

Data Analysis Through Modeling: Thinking and Communicating in Context

Kris Green and Allen Emerson

Fall 2017 Edition¹

¹©2017 Kris H. Green

About this text

Data Analysis Through Modeling is a one-semester data analysis and calculus text that can be used as part of a one-, two- or three semester sequence of mathematics courses usually required of business and management undergraduate majors. We believe the following features distinguish this text from other texts in the curriculum:

- ⇒ A wide array of problems, from small, data-driven problems that focus on a single skill, to more complex, open-ended problems that draw upon multiple skills
- ⇒ Software (e.g. spreadsheets and R) integrated throughout the text as a tool for thinking deeply about problems
- ⇒ Key problems framed in business contexts that require communication skills as well as analytical skills
- ⇒ Content and approaches based on the MAA's Curriculum Foundations Project CRAFTY report for business and management

The increasingly information-driven demands of business in the 21st century require a different emphasis in the quantitative skills and ways of thinking than traditional mathematics courses have provided in the education of managers. This emphasis has to do with becoming comfortable in the world of data and mathematical models, being able to use technology as a tool through which to think, and expressing one's thinking effectively in writing.

The key, we believe, is data analysis through modeling. Data analysis for us means "What can we find out about this data set relevant to our problem?" Models for us are such things as: averages, boxplots, histograms, single- and multivariable regression equations, both linear and nonlinear. These models are proxies for data that are too complex to understand any other way. We think of calculus as a way of analyzing certain kinds of models, which in turn, reveals something about underlying data structures. Our treatment of calculus emphasizes basic concepts, such as rates of change, constrained optimization, and interpretations of area under a graph, and their applications to business problems. We use spreadsheets to develop numerical methods for both differentiation and integration while deemphasizing symbolic manipulation. We use routines like Excel's Solver routine instead of the simplex method to solve linear programming problems. Using Solver has the advantage that we can also solve nonlinear programs.

As we developed this text, we found the introduction of spreadsheet technology for analysis of data not only changed our teaching approach and the content of the course, but it caused us to modify our assignments as well. We found that we simply could not get the quality and depth of understanding we desired from our students by using conventional exercises. We found that students have to explain their thinking and make explicit their assumptions and inferences. In short, we had to supplement our more conventional exercises with memoranda problems with accompanying data files that students respond to in an appropriate business format that are, in turn, read by their supervisor. Further, we find that students learn more by having a chance to revise their work based on instructor/supervisor feedback. All of which should give an indication as to why the book is subtitled "Thinking and Writing in Context."

Although the text has a unit of descriptive statistics and develops regression all the way through multivariable regression with interaction terms, *Data Analysis Through Modeling* is

not a statistics text. Most one-semester introductory statistics courses do not treat regression at the level presented in this text. Moreover, most introductory statistics texts do not give the same emphasis to descriptive statistics that this text does, which is to use these relatively simple concepts for rather deep analysis. *Data Analysis Through Modeling* fits well with an introductory statistics course that primarily deals with probability and hypothesis testing.

How this text fits into the curriculum

We recommend the following tracks for a three-credit-hour, semester-long course using *Data Analysis Through Modeling*:

- For students not having a prior statistics course: Chapters 1-9, 11-12 [11 chapters]. This course would not contain calculus and would be the first in either a two- or three semester sequence: 1) data analysis and statistics or 2) data analysis, statistics, and calculus. In our experience, students then do quite well in the follow-up statistics course after their experience with our approach to data analysis.
- With a statistics prerequisite: Chapters 1-3, 7-9, 11-17 [12 chapters]. This course would contain calculus and constitute the second course in a two-semester sequence containing probability and hypothesis testing, data analysis, and calculus.

The basic concepts of calculus are emphasized and applied to business problems involving marginal analysis, optimization and area under a curve. As recommended by CRAFTY, formal techniques of symbolic manipulation are kept to a minimum, whereas spreadsheets are used extensively not only for finding numerical solutions but, equally important, for the development of the basic concepts of calculus themselves.

The Technology Used in this Text

The material in this text is not designed for passive reading. Rather, you should be reading the material while you have some sort of software package to help you work through the examples. Most modern spreadsheets (e.g. Microsoft Excel) will easily allow you to follow along. Throughout the book, examples will be shown as they typically appear in MS Excel. However, this software has limitations. Eventually, the difficulty of getting the spreadsheet to do what you need outweighs the quality of what you get. This is why we advocate using the right tool for the job. So in the supplemental technology guide, you will also find information for installing and using R, a powerful, open-source (read “free!”) software package for doing statistical analysis, visualization, and modeling with data. Together, these software tools provides a dynamic environment for problem solving.

Excel and other spreadsheets are very useful and visual when it comes to organizing your data and making simple graphs and charts. Thus, it is useful for most of the first two units of the text and the beginning of unit 3. However, even this has caveats, since Excel does not include any built-in tools for generating boxplots of any sort, and the steps needed to make a histogram from un-summarized data are cumbersome. For this reason, we recommend

Unit	Thinking Strategy
Quantifying the World	Students learn the importance of data and how to locate data in real world situations.
Analyzing Data Through Summary Models	Students learn how to use basic charts and graphs to deeply understand a problem situation.
Analyzing Data Through Linear Models	Students learn how to apply proportional reasoning to understand data with one or more independent variables.
Analyzing Data Through Nonlinear Models	Students learn to build models by linearizing non-proportional data and learn how to interpret these in realistic situations.
Analyzing Data Through Calculus Models	Now that students understand how to build models from data, they learn how to use concepts from calculus to understand the problem from which the data and the model were derived.

Table 1: Units and thinking strategies covered in the text.

introducing R in the second unit and using both as needed, until the complex modeling in chapter 9 makes R by far more useful.

In particular, readers will have the opportunity to learn about and use the following tools: pivot tables, sorting, stacking and unstacking data, basic statistical functions, frequency tables, sumproduct, building boxplots and histograms, correlation tables, simple regression, multivariable regression (quantitative and qualitative), scatterplots, trendlines, Goal Seek, SOLVER table and graphing in three dimensions. In addition, students will develop many basic computer literacy abilities, such as copying and pasting and integrating numerical, textual and graphical analyses into a single Word document. But what is most important about the way students learn these tools is that they are all taught in the context of solving business-type problems; this context, we believe, is critical for students learning how to transform these tools from a set of instructions to follow into a method of thinking and analyzing data.

The Structure of the Book

This text is organized into five units, not all of which can be covered in one semester, as mentioned above. The chapters in each unit are all connected through a common “thinking strategy”. The thinking strategies are described in the table 1. The breakdown of topics in each chapter within the units is described later.

Each chapter is designed to be covered in one week of a typical semester course. Since the homework problems (see below) come at the end of a chapter, the homework schedule should, ideally, consist of one assignment per week. Each chapter’s introduction provides a brief overview. It also includes a list of goals and objectives that the student should have after completing the chapter. After the introduction and overview, the main content of each chapter is separated into two major sections, each of which consists of the following:

Discussion. This presents a short overview of the chapter or discusses a short motivational example illustrating the use of the chapter material. The material in this section is conceptual in nature.

Definitions and Formulas. This lists the factual information of the chapter in the form of definitions, formulas, graphs, and methods of computing. It is intended as a reference guide.

Worked Examples. These offer worked examples of using the formulas and techniques of the chapter. This material is more often procedural in nature, but uses concepts to unpack and apply the material to realistic situations from the business world.

Explorations. These involve small scenarios, often supplemented with a data file for you to explore in whatever software you are using. They are open-ended and require discussion and scaffolding. These are basically guided-discovery type activities and are ready-made in-class activities, but can also be completed by students outside of class in order to enhance their understanding of the chapter material.

Homework Problems

Each chapter within a unit is designed to provide the material for a weekly homework assignment at the end of the second section of the chapter. The problems at a chapter's end come in three types: Mechanics and Technique Problems, Application and Reasoning Problems, and Memo Problems (which include Communication and Professionalism skills). Although we consider the memos to be the heart of any course using this book, the number of memos instructors choose to assign on a weekly basis will vary and the two other types of problems work very well to provide a balanced weekly assignment load.

Mechanics and Technique Problems. These problems involve straightforward calculations by hand or, more often, with the computer, and use the basic definitions, formulas, and computer techniques from the chapter.

Application and Reasoning Problems. These problems require students to analyze data or apply the concepts of the chapter to small decision-making scenarios. Many of these require students to explain their thinking in a few short sentences so that the inferences they have drawn from the data and other information are made explicit.

Communication and Professionalism Problems. These questions are designed to focus on writing clearly about the data and analysis you have done. You will also have to think carefully about the audience and their perspective in order to adequately address these questions.

Memo Problems. Each unit includes a larger problem that requires you to tie together the ideas from the chapters in the unit. These are framed as an assignment from a supervisor at the mythical Oracular Consulting. The assignments are written in the style of a management memo, often having a rather open-ended feel, and will most often direct

the analysis staff (which is you) to analyze some data for a client, using the tools of that unit (and possibly previous chapters). You are expected to reply to these memos with a professionally written report. Most memo problems usually permit more than one “correct” response. We have developed detailed grading checklists for assessing each memo which are invaluable should instructors choose to have use these memos. These rubrics do not contain “answers” per se, but rather statements to be checked off by the instructor that note lapses in analysis, missing pieces, incorrect or misapplied mathematical/computer procedures, or point out structural writing difficulties. These statements are divided into three discrete areas: Mechanics and Technique, Applications and Reasoning, and Communication and Professionalism, and each of these three is divided into two levels of competence, Expected and Impressive (see the appendices for an example). In addition to the unit-wide memos, there are some additional memo problems in the appendix that are each more focused on a single chapter of material.

Reading Complex Texts

As you read the text, not only should you be working through the examples by hand and with the software, but you should also be taking notes to help you remember and understand what you have read. Here is one technique for analyzing written material and making sense out of it for you. This will help you in reading material like this, but will also help you to learn how to write material like this for yourself.

It may help to separate each section into paragraphs, then describe each of these paragraphs with two different words or phrases. The first word could refer to the content of the paragraph; for example, it might be about variables or about observational data. The second word should refer to the purpose of the paragraph in the overall document; for example, is it an introduction or a clarification? Some other possible purposes for a paragraph are to define a term, provide an example, illustrate exceptions or clarify conditions, state conclusions, or to summarize the analysis. There are many more possibilities. The point here is not to get the “right answer” but to come to some understanding on your own. You may be asked to discuss this in class.

Entering Student Profile

As a student entering a course using this book, or as someone using this book on their own to gain new skills, techniques, and concepts about quantitative analysis in the business world, you should have some skills in the areas of mathematics, the use of technology, and writing.

Mathematics background: Basic algebra skills are essential, but the text does not require well-honed algebraic skills as a pre-requisite. What is most essential is the abstraction that algebra supports in moving from concrete objects to expressions and functions with parameters and variables. Students should have had a mathematics background up to, but not necessarily including, precalculus.

Technology background: The text does not assume that the students have any knowledge of spreadsheets, though in our experience most have some familiarity with computers and spreadsheets, Excel in particular.

Writing Background: In our experience, students gain the most from this text when it is taught in a writing-intensive format, using a selection of the chapter memo problems (including revisions). Most first-year college writing course requirements will have prepared students sufficiently to write at the level the memos demand.

Exiting Student Profile

By the end of a course based on this text, we expect students to have developed capabilities in the three areas of the homework problems. The first area (mentioned above) is “Mechanics and Techniques,” which includes knowledge of basic mathematical notation and symbol manipulation as well as basic technological (especially spreadsheet) skills for structuring problems for solutions. The second area is “Application and Reasoning,” which covers the ability to contextualize the mathematical ideas, to extract quantitative information from a context, and to make logical inferences from quantitative analyses. The final area is “Communication and Professionalism,” which covers the ability to write coherently about a problem and its proposed solution and to communicate this analysis in a professionally appropriate manner. Specifically, a student earning an average grade in a course based on this text would have the capabilities in each of the three areas shown in the outline below.

Mechanics and Techniques

- Has had experience formulating and interpreting algebraic, graphical and numerical mathematical models
- Has used spreadsheets to apply various mathematical, statistical, and graphical tools to business situations
- Understands enough about data analytic techniques to effectively communicate with statisticians and other types of expert analysts
- Is competent and comfortable with spreadsheets
- Has learned to use technology as a tool with which to think

Application and Reasoning

- Understands how to define a problem situation in terms of data
- Understands the basic design of data collection forms and how to employ them
- Has experience in working in open-ended, ambiguous problem situations
- Understands the interpretive power of graphical displays of data
- Understands the power and limitations of mathematical models
- Has experience in interpreting the parameters and coefficients of mathematical models
- Is capable of drawing contextual inferences from statistical and graphical analysis

Communication and Professionalism

- Knows the importance of writing in the workplace
- Can write competent memos and reports as part and parcel of one's job
- Knows how to integrate and arrange statistical and graphical elements in a word processing document to produce a convincing argument
- Has learned to consider the reader's response to a memo
- Has learned to plan ahead to meet the demands of the course
- Persists when the path is not clear
- Has learned self discipline in accomplishing long and complex tasks

Some Words About Level of Difficulty

Viewed apart from a context of a memo, the mathematics, technology, and writing demands of certain chapters may not seem very difficult when taken separately. But when students analyze a data set, interpret and draw inferences from mathematical formulations within specific problem contexts and then organize the various charts, computer output, and text into a coherent and readily understood memo, they find the work to be anything but easy. Indeed, instructors of this text invariably comment on how they themselves have been challenged by the problems. The open-ended nature of the problems (e.g. see the Chapter 1 memo) contributes to this challenge, as well as the sheer amount of time it takes to complete the whole process. This is one of the reasons that instructors may not wish to assign a memo problem every week, especially when requiring revisions, which students mightily appreciate and benefit from.

Chapter Details

Chapter 1 “Problem Solving” is about learning to ask questions in order to help frame a problem in terms of data that will help in the analysis and solution of the problem. This is followed in chapter 2, “Understanding the Role of Data,” with how to organize your data and understand the types of data you may have to use in your analysis.

Chapter 3, titled “Using Models to Interpret Data” is about building simple summary models to analyze data using the mean, standard deviation and pivot tables. Since a picture (or graph) is said to be worth a thousand words, you will learn how to create and use boxplots to analyze data in chapter 4, “Box Plots” and then supplement these with histograms to display more information about the distribution of data in chapter 5, “Histograms.” In the final chapter of this unit, chapter 6 “Interpreting Spatial Models”, you will learn to estimate statistics from summary data and connect the different spatial models (boxplots and histograms) to build a more complete understanding of a set of data.

The one-variable summaries of unit two are extended in unit three, starting with chapter 7 “Correlation,” which explains how to picture and quantify the relationship between two variables using correlation and trendlines. Chapter 8, “Simple Regression,” enriches this approach by introducing simple linear regression to measure the effect of one variable upon another and to interpret how well our models fit the data. Chapter 9, “Multiple and

Categorical Regression,” extends these regression models into many dimensions and shows how to use qualitative variables in your models. But analysis is not just about building and interpreting models; you have to understand their quality and whether you might be able to simplify them. This is explored in chapter 10, “Is the Model Any Good?” which also introduces interaction terms into the models.

Eventually, though, any modeler will find a situation that cannot be adequately modeled with linear models. Chapter 11 “Graphical Approaches to Nonlinear Data” parallels chapter 7, but looks at how curvilinear models, like power functions, logarithms, and exponentials, can be used to build more accurate models of certain data. You can then learn about how to create such models through regression techniques in chapter 12, “Modeling with Nonlinear Data” where you will also gain experience interpreting these models and their quality. Chapter 13 “Nonlinear Multivariable Models” offers a brief overview of modeling multidimensional nonlinear data, using tools you have developed in previous chapters.

The chapters in the final unit introduce you to one of the most powerful mathematical concepts ever developed: calculus. In “Optimization and Analysis of Models” (Chapter 14) we generalize the idea of slope to the idea of a derivative and use this tool to find maximum and minimum values of polynomial and power models. Chapter 15, “Deeper Exploration of Logs and Exponentials” extends derivatives to the analysis and optimization of logarithmic and exponential models. “Optimization in Several Variables” (Chapter 16) teaches how to use your prior modeling experience to define constraints and use tools like the SOLVER routine to find optimal solutions in more complex situations. Derivatives are only half of what calculus is about, though. Chapter 17 “Area Under the Curve” shows you some of the other half, by finding areas under curves using both the Fundamental Theorem of Calculus and numerical methods. This idea of an integral has numerous applications in business.

Copyright Notice

This edition of *Data Analysis Through Modeling: Thinking and Writing in Context*, including all written material, examples, problems, associated data files, and supplemental material, is the property of Dr. Kris H. Green, copyright 2017.

Dedication and Acknowledgements

First and foremost, this book is dedicated to Dr. Allen Emerson, my co-author and long-time friend, who passed away during the completion of this project. His hard work, tenacious intellect, and willingness try new ideas made this book possible. Our spouses also deserve a great deal of the credit for this work. Cheryl Forbes teaches writing and rhetoric, and her influence on Allen's approach to teaching mathematics was enormous. My wife, Brenda, has had a profound influence on my approach to teaching overall and on helping me understand the business world enough to bring a new approach to mathematics into it. Both of them put up with our tendencies to lose sight of everything but this project, at times spending upwards of twelve straight hours a day trying to understand student learning in the course we wrote this book to support.

We would also like to thank Anne Geraci for her invaluable assistance. She has provided enormous editorial support in reviewing the materials and helping to prepare this updated edition of the textbook. Any errors, typos, or omissions are entirely due to our work and not her excellent reviewing of the material.

I would also like to thank Carol Freeman, the department of Mathematical and Computing Sciences at St. John Fisher College, and the School of Business at Fisher. They have provided us with opportunities to try new approaches to an old course and have supported our ideas, no matter how strange they seemed. The course we designed, and ultimately, the textbook we wrote, would also not have been possible without the assistance of many adjunct faculty members who helped us with suggestions, revisions and ideas: Mike Rotundo, Rebecca Tiffin, and Mary Ann Cape.

In addition, Ginger James provided us with invaluable assistance in the early years of the course, attending class, tutoring students, and offering suggestions while still an undergraduate at St. John Fisher College. We have also benefited from the able tutoring of several undergraduates, and thank all of them for their assistance in supporting the course.

I	Quantifying the World	1
1	Problem Solving By Asking Questions	13
1.1	Defining the Problem	14
1.1.1	Definitions and Formulas	15
1.1.2	Worked Examples	16
1.1.3	Exploration 1A: Assumptions get in the way	20
1.2	Why Data?	22
1.2.1	Definitions and Formulas	24
1.2.2	Worked Examples	25
1.2.3	Exploration 1B: Beef N' Buns Service	29
1.3	Homework	31
2	The Role of Data	35
2.1	Extracting Data from the Problem Situation	37
2.1.1	Definitions and Formulas	40
2.1.2	Worked Examples	42
2.1.3	Exploration 2A: Extracting Data at Beef n' Buns	47
2.2	Organizing data for Future Analysis	48
2.2.1	Definitions and Formulas	49
2.2.2	Worked Examples	51
2.2.3	Exploration 2B: Entering Beef n' Buns Data into a Spreadsheet . . .	55
2.3	Homework	56
II	Analyzing Data Through Summary Models	59
3	Using Models to Interpret Data	65
3.1	The Mean As A Model	67
3.1.1	Definitions and Formulas	69

3.1.2	Worked Examples	71
3.1.3	Exploration 3A: Wait Times at Beef n' Buns	79
3.2	Categorical Data and Means	80
3.2.1	Definitions and Formulas	81
3.2.2	Worked Examples	82
3.2.3	Exploration 3B: Gender Discrimination Analysis with Pivot Tables	88
3.3	Homework	89
4	Box Plots	93
4.1	What Does "Typical" Mean?	94
4.1.1	Definitions and Formulas	94
4.1.2	Worked Examples	95
4.1.3	Exploration 4A: Koduck Salary Increases	98
4.2	Thinking inside the box	99
4.2.1	Definitions and Formulas	100
4.2.2	Worked Examples	101
4.2.3	Exploration 4B: Relationships Among Data, Statistics, and Boxplots	105
4.3	Homework	106
5	Histograms	111
5.1	Getting the Data to Fit a Common Ruler	113
5.1.1	Definitions and Formulas	113
5.1.2	Worked Examples	114
5.1.3	Exploration 5A: Cool Toys for Tots	118
5.2	Profiling Your Data	119
5.2.1	Definitions and Formulas	120
5.2.2	Worked Examples	121
5.2.3	Exploration 5B: Beef n' Buns Service Times	128
5.3	Homework	129
6	Interpreting Spatial Models	135
6.1	Estimating Stats from Frequency Data	136
6.1.1	Definitions and Formulas	137
6.1.2	Worked Examples	138
6.1.3	Exploration 6A: Data Summaries and Sensitivity	144
6.2	Two Perspectives are Better than One	146
6.2.1	Definitions and Formulas	147
6.2.2	Worked Examples	148
6.2.3	Exploration 6B: Stock Investment Decisions	152
6.3	Homework	154

III	Analyzing Data Through Linear Models	159
7	Correlation	165
7.1	Picturing Two Variable Relationships	166
7.1.1	Definitions and Formulas	167
7.1.2	Worked Examples	170
7.1.3	Exploration 7A: Predicting the Price of a Home	175
7.2	Fitting a Line to Data	177
7.2.1	Definitions and Formulas	178
7.2.2	Worked Examples	179
7.2.3	Exploration 7B: Adding Trendlines	182
7.3	Homework	184
8	Simple Regression	187
8.1	Modeling with Proportional Reasoning in Two Dimensions	189
8.1.1	Definitions and Formulas	190
8.1.2	Worked Examples	191
8.1.3	Exploration 8A: Regression Modeling Practice	194
8.2	Using and Comparing the Usefulness of a Proportional Model	195
8.2.1	Definitions and Formulas	195
8.2.2	Worked Examples	198
8.2.3	Exploration 8B: How Outliers Influence Regression	202
8.3	Homework	203
9	Multiple Regression Models	207
9.1	Modeling with Proportional Reasoning in Many Dimensions	209
9.1.1	Definitions and Formulas	210
9.1.2	Worked Examples	213
9.1.3	Exploration 9A: Production Line Data	218
9.2	Modeling with Qualitative Variables	219
9.2.1	Definitions and Formulas	220
9.2.2	Worked Examples	220
9.2.3	Exploration 9B: Maintenance Cost for Trucks	223
9.3	Homework	224
10	Is the Model Any Good	227
10.1	Which coefficients are trustworthy?	229
10.1.1	Definitions and Formulas	230
10.1.2	Worked Examples	230
10.1.3	Exploration 10A: Building a Trustworthy Model at EnPact	234
10.2	More Complexity with Interaction Terms	235
10.2.1	Definitions and Formulas	235
10.2.2	Worked Examples	236
10.2.3	Exploration 10B: Complex Gender Interactions at EnPact	240
10.3	Homework	241

IV	Analyzing Data with Nonlinear Models	243
11	Nonlinear Models Through Graphs	249
11.1	What if the Data is Not Proportional	251
11.1.1	Definitions and Formulas	251
11.1.2	Worked Examples	257
11.1.3	Exploration 11A: Non-proportional data	261
11.2	Transformations of Graphs	263
11.2.1	Definitions and Formulas	264
11.2.2	Worked Examples	266
11.2.3	Exploration 11B: Shifting and Scaling the Basic Models	271
11.3	Homework	274
12	Modeling with Nonlinear Data	279
12.1	Non-proportional Regression Models	281
12.1.1	Definitions and Formulas	282
12.1.2	Worked Examples	283
12.1.3	Exploration 12A: Learning and Production at Presario	289
12.2	Interpreting a Non-proportional Model	290
12.2.1	Definitions and Formulas	291
12.2.2	Worked Examples	293
12.2.3	Exploration 12B: What it means to be linear	297
12.3	Homework	298
13	Multivariate Nonlinear Models	303
13.1	Models with Numerical Interaction Terms	304
13.1.1	Definitions and Formulas	305
13.1.2	Worked Examples	305
13.1.3	Exploration 13A: Revenue and Demand Functions	311
13.2	Interpreting Quadratic Models in Several Variables	313
13.2.1	Definitions and Formulas	314
13.2.2	Worked Examples	316
13.2.3	Exploration 13B: Exploring Quadratic Models	322
13.3	Homework	324
V	Analyzing Data Using Calculus Models	329
14	Optimization	333
14.1	Calculus with Powers and Polynomials	335
14.1.1	Definitions and Formulas	337
14.1.2	Worked Examples	340
14.1.3	Exploration 14A: Finding the Derivative of a General Power Function	344
14.2	Extreme Calculus!	346
14.2.1	Definitions and Formulas	346

14.2.2	Worked Examples	347
14.2.3	Exploration 14B: Simple Regression Formulas	352
14.3	Homework	354
15	Logarithmic and Exponential Models	357
15.1	Logarithms and their derivatives	358
15.1.1	Definitions and Formulas	359
15.1.2	Worked Examples	360
15.1.3	Exploration 15A: Logs and distributions of data	364
15.2	Compound interest and derivatives of exponentials	366
15.2.1	Definitions and Formulas	366
15.2.2	Worked Examples	367
15.2.3	Exploration 15B: Loan Amortization	371
15.3	Homework	373
16	Optimization in Several Variables	375
16.1	Constraints on Optimization	377
16.1.1	Definitions and Formulas	377
16.1.2	Worked Examples	378
16.1.3	Exploration 16A: Setting up Optimization Problems	384
16.2	Using Solver Table	386
16.2.1	Definitions and Formulas	386
16.2.2	Worked Examples	387
16.2.3	Exploration 16B: Sensitivity Analysis	394
16.3	Homework	396
17	Area Under a Curve	401
17.1	Calculating the Area under a Curve	403
17.1.1	Definitions and Formulas	405
17.1.2	Worked Examples	406
17.1.3	Exploration 17A: Numerical Integration	409
17.2	Applications of the Definite Integral	410
17.2.1	Definitions and Formulas	410
17.2.2	Worked Examples	411
17.2.3	Exploration 17B: Consumers' and Producers' Surplus at Market Equilibrium	417
17.3	Homework	418
VI	Appendices	423
A	Memo Problems	425
A.1	Carnivorous Cruise Lines	426
A.2	Carnivorous Cruise Lines, Part 2	428
A.3	Carnivorous Cruise Lines, Part 3	429

A.4 Matching Managers to a Company	430
A.5 Service at Beef n' Buns	431
A.6 Portfolio Analysis	432
A.7 Truck Maintenance Analysis	433
A.8 Commuter Rail Analysis	434
A.9 Gender Discrimination	435
A.10 Truck Maintenance Expenses, Part 2	436
A.11 DataCon Contract	438
A.12 Insurance Costs	440
A.13 Revenue Projections	441
A.14 Profit Analysis	442
A.15 Loan Analysis	443
A.16 Advertising Costs	444
A.17 Pricing Dispute	445
B Sample Rubric for Evaluating Memo 7	447

Part I

Quantifying the World

Thinking of the world as data

In today's world, everyone is collecting data. It's everywhere. Some even say we are inundated with data, so much so that we cannot keep up with the amount of data we can generate and collect. With this in mind, consider the following definition of **data**:

Data: Information extracted from real-world contexts that has been organized for analysis in forms that can be used for making decisions.

Given this definition, who among the following are more likely to think of the world as data in their professional work?

- Mathematicians?
- Scientists?
- Reporters?
- Detectives?
- Business managers?

We can be fairly safe in saying that mathematicians do not see the world as data in their day-to-day work. Only a relatively few mathematicians deal with the real world at all in their professional work. While they may construct mathematical models that others (such as scientists or business managers) may find very useful in making sense of real-world data, mathematicians themselves are often quite unconcerned about the real-world usefulness of their work.

Scientists, on the other hand, use data extensively in their everyday work, but they use it under carefully controlled circumstances. They are interested in data in terms of its experimental reproducibility. They tend to think of the world in terms of patterns of data that occur and reoccur under certain specified conditions. They tend to think of real-world data in terms of how it conforms to predictable scientific laws.

Reporters think of the world as stories. Not rambling stories, but stories told in a certain way so as to communicate a lot of information in a short space. They are often focused on getting to the heart of a story by looking at the **causes**, **symptoms**, and **effects** of what has happened. To sort everything out, they efficiently extract information from the cacophony of life's events by asking questions built on the 5-W's: Who, What, Where, When, and Why. This helps them find and answer to the questions that start with "How".

Depending on the nature of the story, experienced reporters usually try to work the answers to the 5-W's into the first paragraph or so of the story. This enables them to accurately convey the context and gist of the story as soon as possible, so that anything else further down the column is merely an elaboration of what is already known. The point is that without the 5-W's approach news reporting becomes less focused, more meandering, and results in less accurate information transmitted for the amount of print expended. But do reporters see the world as data in the sense we have defined it above? They certainly see the world as story based on information, but as data?—probably not. Even if a reporter were

writing a financial story, for example, and even if that story contained numerical information that was organized in a form that could be analyzed in some way (for example, graphs or charts), that data would not be specific enough or numerous enough to be of much use to a bank or stock brokerage firm for decision making. Indeed, such businesses would probably have their own data analysis staff anyway or would contract out such services. Nevertheless, the 5-W tool for extracting and organizing information from the world is a useful one from which managers can profit. You will get the opportunity to try it out for yourself in this first unit.

Do detectives see the world as data? “Nothing but the facts, ma’am,” says Joe Friday, the laconic police investigator of that ancient TV show *Dragnet*. Clearly, detectives think of the world as data. Not management-type data, but certainly as information that is organized for analysis in forms that can be used for making that one bottom-line decision - whodunnit? Of course, there are a host of decisions that precede this big one. The detective makes these decisions by drawing inferences from evidence, which is another way of saying “by analyzing the data.” So while the data that detectives work with is quite different from the data that managers compile, there is a similarity in what the two do with the data, the way they marshal compelling evidence and draw inferences from that evidence as they argue their case. Everyone knows that detectives cannot make judgments or decisions that will hold up in court without the proper supporting evidence. So it is with business managers. They likewise must present their arguments based on proper supporting evidence. We will be concerned with what constitutes proper evidence and how to present it in almost all of the homework problems.

Because business managers have to constantly make decisions in less-than-certain circumstances, it is to their advantage to think of the world as data, almost to the extent that it becomes second nature to them, a way of seeing. While it is true that certain aspects of business occur with regularity, such as manufacturing processes or financial dealings, it is also true that many important aspects of business, such as sales trends or employee equity issues, are not reducible to known scientific laws. Then too, all aspects of business eventually come down to that one irreducible basic fact, the bottom line. For example, here is a list of bottom-line questions that a manager has to answer on a day-to-day basis that should make clear the case for thinking of the world as data:

- How are we performing?
- Do we have a problem? If so, what is it?
- What can we expect will happen in the future if we continue doing what we’re doing now?
- What will happen in the future if we make some key changes?

Putting the case plainly: Would you place a person in a management position requiring answers to these kinds of bottom-line questions if they could not see the world as data? This raises another question: Does this mean that managers have to be statisticians?

You may have noticed that the list of professions above does not include statisticians. To be sure, they are the real data professionals and data is their bread and butter. But

statisticians are, in a sense, generalists. While they probably do see the world as data in a way that few others do, chances are that they do not see your particular business world as you do. As a business manager, you are in a position of responsibility and you are the one who has to make those bottom-line decisions that often have far-reaching consequences. Nevertheless, you do need to think of the world as data.

Which brings us to this point: is this then a statistics text for business managers? The answer is “no, not really.” While you will gain experience in dealing with those all-important bottom-line questions listed above by using some rather basic techniques, it would indeed take a lot of statistical background to be able to answer them the way a statistician would. But companies do not, as a rule, hire statisticians as their managers. Similarly, this book is not written for prospective statisticians, but rather prospective managers who will have learned enough from the text to not only appreciate the value of data but also to be able to manage its collection and analysis. This means that they will be expected to understand the technical language of professional data analysts, at least enough to effectively communicate with them, and then to make sense of it all for both their employees and their supervisors. This book takes seriously the assumption that you will be involved either as managers or as team members of a group of professional data analysts in projects similar to those presented in the memo homework problems. This is why the book begins with a unit on thinking of the world as data.

Key Thinking Strategy: Thinking of the world as data.

How does one even begin to recognize and then collect the necessary data that will enable us to first define the problem and then to analyze it? Restated: How does one go about isolating what is relevant to the problem and what is not relevant from the undifferentiated flow of activities or actions or states of existence that we confront in a real-life situation?

One of the easiest and most effective ways to think of the world as data is from the reporter’s point of view. The job of a reporter is to tell a story, a story based on facts. The reporter collects facts mostly by asking questions. This an excellent starting point for the business manager as well. In this unit we will use the 5 W’s-plus one extra - as a strategy to help us see the world as data: Who, what, when, where, why and how. Although we will be using the 5W’s+H, or a selected subsets of them, as a thinking strategy throughout the book, they may take on different meanings and emphases in different sections of the book, depending on whether we are doing the initial work of defining the problem, or creating a plan and timeline to carry out the project or using a particular mathematical technique to analyze the data or writing a memo to convey the results of our analysis. The point is that while it is important to be able to roll the 5W’s+H off the tongue, it is also important to be aware that not only will we not always talk about all of them all the time but that even when we do we may not be thinking of them in quite the same way from situation to situation. Then too, the W’s are not necessarily mutually exclusive, meaning that, for example, there may be situations in which it does not make sense to ask What? without asking Where? in the same breath or How? without asking When?

Key Communication Strategy: Be Professional

In this course, all solutions to memo problems should be written so as to “stand alone.” That is, any reader should be able to pick up your solution and read it without knowing anything else about the problem (but might need to know something about the content). In other words, you must think objectively about your response to the memo. Ask yourself, “What if I got this on my desk and knew nothing about the project? Would I understand the memo?” If the answer is no, then your response needs more information about the context and the goal of your writing.

Writing is one of the most (if not THE MOST) important aspects of a career - any career, but especially one in business. Writing is a way of seeing, understanding, explaining, and envisioning the world around you. Managers with good communication skills will never be outsourced. In this course, you will practice preparing two types of responses to memo problems: memos and reports. Since writings like these may be passed around to many people with whom you have never directly interacted, making your communications as solid as possible may be the only first impression you ever get to make on these individuals! Regardless of whether you are writing a short memo or a more involved report, the writing should be professional. One of the ways to do this is to make sure that you DO NOT include a familiar closing, like “Sincerely, John.” Such closings are good for personal letters, but the heading of a memo or report should tell the reader who wrote it. And since you have submitted the memo or report as a part of your job, it should go without saying that you are sincere in what you have written. It’s best to just avoid such familiarities in writing altogether.

Writing business memos

The first type of response you will produce is an informal, routine type of memo (see the sample in Example 1 - problem at Gamma). In a routine memo, you should:

- Use an informative subject line that summarizes the purpose of the memo.
- State the purpose of your memo in a direct manner in your introduction.
- Organize the body of your memo for readability, using visual cues, bulleted lists, and tables.
- Summarize your analysis, findings, or recommendations, as appropriate in your closing paragraph.

Writing business reports

The second type of response you will be preparing is really what would be called a Business Report. In this type of response, you should include the following sections:

1. The Executive Summary (1 - 2 paragraphs)

- (a) Introduction. Tell the reader what the problem is about. Briefly tell the reader what data has been collected. Explain what you want to determine or find out. Make sure this introduction answers the questions: (1) Why are you writing this report? and (2) What is the context or scenario that was presented? Much of this can be restated directly from the assignment you were given.
 - (b) Preview. Then tell the reader what you are planning to do and give him/her an idea about how the analysis of the data will flow. Explain what the structure of the memo is.
 - (c) Conclusions of the Analysis. Briefly describe the conclusions, if possible.
2. Analysis (the bulk of the report)
- (a) Steps to Reach the Conclusion. Explain what you did; explain all the steps and the reasoning that led to each step.
 - (b) Supporting Evidence. Provide all the supporting evidence - graphs, tables, charts, etc. Label each item appropriately and refer to it by name in the text of the response. Organize the evidence and the explanation so the reader can follow the argument.
3. Conclusion (1 - 2 paragraphs)
- (a) Summary of results. Provide a summary of the results of the analysis, collected into a convenient form (like a table).
 - (b) Context. Put your solution into context and explain what the results actually mean for the situation you are analyzing. Interpret the results so the reader does not have to guess.
 - (c) Sensitivity. How accurate are your results? How much are they likely to change as a result of changes in the underlying data? Give the reader a sense of what the limits are to your analysis, so that he/she knows how much to rely on these results.
 - (d) Advice. If the original assignment/memo called for a decision, be sure you provide a clearly stated response to that requirement, and clearly connect your evidence and process to the decision you are advocating.

Elements Of An Effective Communication

Another helpful hint is to use headings to improve both the visual and logical organization. These may take many different visual formats (bold type, underlining, larger font, etc.) but should match with the structure provided in the introduction of the memo, so that an interested reader can easily locate the information he or she needs. This helps with your overall goal in writing a good memo: Make your thought process transparent for both your reader and yourself. After all, you may need to come back to this project in six months or a year; a good memo now will save you a lot of time later. Here's a handy checklist of other things to think about:

- ✓ Writing is competent (grammar, spelling and sentence structure are correct and clear)
- ✓ Directed at the right person (your audience)
- ✓ Fully addresses the problem
- ✓ Margins are clean
- ✓ Organized and focused so it can quickly be read and understood
- ✓ Sufficient analytic detail to justify all conclusions
- ✓ No charts or tables split across page breaks and none of them “fall” off the page
- ✓ Charts and tables are labeled and referred to by these labels in the text
- ✓ Evidence (charts and tables) is embedded in the report, somewhat near the discussion about it (integration of text and graphics)
- ✓ Charts and tables are all legible
- ✓ Document sections are in order and not fragmented
- ✓ Introduction provides an overview of the memo and its structure
- ✓ Introduction reminds the reader (briefly) of what the problem context is and what you have been asked to do
- ✓ Include a header that explains who wrote this, when, and who should be reading it
- ✓ Avoid familiar closing salutations

Elements of an Exceptional Report

- ✓ Additional tables used to organize all conclusions and make comparisons easier
- ✓ Features like bold text, shading, and headings used to highlight important information
- ✓ Conclusion summarizes the analysis
- ✓ Report includes an analysis of accuracy and possible errors

The 5W+H Strategy

- ✓ Who writes? To whom?
- ✓ What do they write about?
- ✓ When should you write?
- ✓ Where should you write?
- ✓ Why should you write?

Email protocols for professional communication

You are probably used to sending and reading emails. You probably even have multiple email accounts to manage. Since all systems are slightly different, we want to talk about something more important to the general concept of sending emails, whether you are sending them to your course instructors, your family, your boss, or your friends.

Personal Information. It is critical that several things appear in any email message. These should include *Your name*. It may seem redundant; after all, the recipient has your email address. But consider this. Many usernames are not clearly connected to your name. When someone receives an email from abC1434, they don’t usually have time to think about everyone they know that might have that as a username. They really don’t know who sent the message. If you aren’t willing to sign your actual name to the message,

you should really think about whether it is a message you should be sending. You should also include *your job title and address* so that the reader understands your role in whatever processes are discussed in the email and for whom you are working (i.e. the client, the consulting firm, the legal department, etc.) Finally, it's good to provide *alternative contact information* like a phone number, so that the reader can, if they are interested or need to, follow up on the email to get more information. You could include all this in a **signature file** that is automatically included in emails you write. If your e-mail system allows signature files, it probably will let you create several different ones. You could have a signature file for professional communications and a separate one for personal communications, with different information in each.

Subject Line. Be sure that you put something meaningful in the subject line of your messages. With the thousands of junk email messages that most people receive each week, the subject line will often be the deciding factor in whether your email gets read or not. If your email is to a course instructor, for instance, your instructor will appreciate it if the subject line contains the course number (like MATH 130) as the first part of the subject line. In the business world, many people will automatically delete messages with inappropriate subject lines or without a subject altogether, not to mention that many spam filters will automatically remove messages without subject lines before the intended recipient even gets an opportunity to examine them.

Be Professional. Make sure that you do a little proofreading in your emails. They may be short, but don't be sloppy. Grammar, punctuation, and spelling are all important in emails. Very often, this may be your only chance to make a first impression, so make it good.

Memo Problem: Initial Planning for StateEx Contract

Each unit of this text has a memo problem associated with it. These problems will draw upon all (or most) of the skills and ideas you explore in the chapters of the unit. The memo below is your first, and will let you practice working with the content of chapters 1 and 2 in the context of a trucking company having trouble with its loading and unloading processes. Read it carefully, then go through the chapters of the unit. When you have finished working through the unit, come back to the memo and prepare the required report in order to show off what you have learned.

To: Analysis Staff
From: Project Management Director
Date: June 28, 2017
Re: Shipping and unloading process at StateEx

The warehouse manager for StateEx has contacted us to investigate their company's shipping and loading process in order to identify any inefficiency. The operation involves trucks picking up cargo from the airport, the shipyards and the train yards and transporting it to a central warehouse for distribution locally. The cargo is packed into various sized boxes that are weighed at the pickup location before being loaded onto the trucks. The manager is convinced that unloading the trucks at the warehouse is taking too long, slowing down the processing of the cargo and its distribution.

We need to come up with a preliminary proposal in order to win the contract with StateEx. I will review your proposal, give you some feedback, and then pass it on to the marketing team, who will cost it out. I will write the final cover letter and submit the final proposal to the manager of the warehouse. The marketing team will need your proposal to address all the standard requirements of an RFP so that they can accurately cost out this job for the contract bid.

The final proposal should include a written report with data collection forms and spreadsheets incorporated into it.

Standard requirements for an RFP

An RFP is a standard **Request For Proposal**. Clients who have a problem often shop around for consultants to help by sending out RFPs. These include a short description of the problem and the company, and ask for interested companies to put together a proposal. After receiving several proposals, the client reviews them and hires ONE of the companies to work with. So an effective proposal is essential to earning a new client.

Proposals are tricky, though. We haven't been hired and paid at this stage. And we haven't actually got any data upon which to base an analysis or recommendations for action. So we have to show them we understand the situation, have a plan for getting at what's really happening, and that we can develop a reasonable solution, regardless of what direction

the data shows us. As a consultant at Oracular, you'll be writing lots of these. In all your proposals, be sure to pay attention to the guidelines below.

1. Explain what the perceived problem we are to investigate is all about - if we can't clearly restate their problem, or push deeper to show that we understand their may be more underneath it than they realize, we won't get the contract.
2. Describe some possible reasons for the problem - the client often mentions some of their own ideas, but we need to be sure we come up with a few reasonable ones on our own to impress them with our knowledge and experience.
3. Make sure you clearly explain the short and long term consequences of the problem, including some that the client may not have thought about.
4. Clearly describe your proposal for gathering data to identify the problem. This should include sample data collection forms and a rough timeline for the whole data collection and analysis process.
5. Use your possible reasons and possible solutions as a way of ensuring that your data collection gets what you might need; that is, use these as a reality check on your thinking process.
6. Include some sample spreadsheets showing how the data will be coded and organized, with explanations for the codes used, units for the variables in the data, and at least 15 observations of each variable. This will "show off" the range of values and demonstrate that you are really thinking about everything to be sure you haven't left any possibilities out.
7. Attach all spreadsheets in a single workbook with relevant names for the sheets, so that the marketing team doesn't overlook them. Below the sample data for each variable you should also describe the type of variable (e.g. categorical, discrete continuous, etc.).
8. Identify any possible difficulties, problems, or expenses (there will indeed be some) that might be encountered in collecting and analyzing the data. Don't include dollar figures because our marketing team will handle this.

CHAPTER 1

Problem Solving By Asking Questions¹

Sometimes the data that is needed to solve a problem has already been gathered and is sitting in a data bank just waiting to be analyzed. Sometimes it is not. If this is the case, one of the first steps in solving the problem is to gather the necessary data. But exactly what data does one need? Clearly, that depends on the problem and that, in turn, assumes we know what the problem really is. This chapter is about letting go of our preconceived notions of what the problem is and then developing ways of getting at the data that will not only define the problem situation but also point to a solution.

- Section 1.1 looks at the nature of problems and how to ask questions to help find solutions.
- Section 1.2 investigates the role of data in helping one to both ask good questions and develop good solutions.

<i>As a result of this chapter, students will learn</i>	<i>As a result of this chapter, students will be able to</i>
---	--

- | | |
|---|--|
| ✓ The importance of not making assumptions about a problem | ✓ Better understand how to read complex interrelated texts |
| ✓ To understand a problem within its context | ✓ Develop a plan for gathering data |
| ✓ The importance of data to identify the true root cause of a problem | ✓ Identify multiple potential root causes for a particular perceived problem |
| ✓ What is involved in gathering data | ✓ Develop a rough timeline for a project |
| | ✓ Write a memo in response to a problem |

¹©2017 Kris H. Green and W. Allen Emerson

1.1 Defining the Problem

One of the first things that an aspiring manager or consultant has to learn about solving a problem is that he or she is not being paid to provide unsubstantiated beliefs about what might or might not be a good solution. Rather, the successful manager or consultant is paid to propose solutions based upon pertinent data and a well-reasoned analysis of that data. The manager's feelings or guesswork or intuitions can be helpful in exploring a problem situation but they cannot by themselves be the basis for making sound and reliable decisions. Again, it is having a clear idea of how to define the problem and having a plan for collecting the relevant data that constitute the professional approach.

The first step to solving a problem is to define the problem. This is not as obvious or as simple as it sounds; there are numerous case studies showing how businesses have wasted large quantities of money trying to solve the wrong problem. Listed below are the other steps in the general problem solving process. Keep in mind that the process is usually not sequential. You will usually find yourself jumping steps and repeating steps in an attempt to refine your solution. Throughout this book, we will be focusing on the first stages - defining the problem, collecting data, and analyzing data - rather than the last stages.

1. Problem formulation stage

- (a) Define the problem
- (b) Identify possible causes and their effects
- (c) Determine data to be collected

2. Data collection stage

- (a) Determine what the variables are and how they will be coded
- (b) Construct data collection forms
- (c) Construct the database for analysis

3. Solution development stage

- (a) Interrogate the data
- (b) Determine a root cause for the problem
- (c) Develop possible solutions
- (d) Use the data to select the best solution

4. Refinement stage

- (a) Test the solution with sample data
- (b) Modify the solution based on the tests

5. Implementation stage

- (a) Present your findings and your plan

- (b) Put your solution into practice
- (c) Collect data as to the effectiveness of the solution
- (d) Modify the solution as needed, based on data

One of the reasons that it is vitally important to define the problem you are studying is because real-world problems are often multifaceted. Their causes may be well hidden, and what you observe - the **perceived problem** - may mask the real problem's causes. Part of your job in studying a problem is to think of possible causes for the perceived problem, then determine ways to investigate the situation by collecting data that can sort through these causes. Making this even more difficult is the fact that a single cause can have multiple effects, each of which may generate more effects of its own, some of which may overlap. Identifying this **chain of cause and effect** is really what understanding the problem is all about.

1.1.1 Definitions and Formulas

Data Information extracted from real-world contexts that has been organized for analysis in forms that can be used to inform decision-making.

Consultant/Business Manager A person who is paid to propose solutions that are based upon the collection and analysis of pertinent data.

Perceived problem What the supervisor or employee or customer or client thinks is happening, which may or may not be the actual problem. It might be a symptom, a secondary effect, or one of the causes. It might even be a red herring, and not be relevant at all.

Problem situation The circumstances in which a problem takes place and that give rise to the problem

Cause The cause of a problem is often very unclear. The cause is what is really keeping your situation from being ideal. Very often, you will need to brainstorm possible causes and then collect data in order to rule out one or more of them.

Effect, Impact, or Symptom This is the real problem, the result of the cause of the problem. It is what you really want to correct. It may be something obvious like lost revenue, or something less clear like low morale. There are usually more than one effect emanating from a single cause.

Short-term Effect This is an effect that is immediately resulting from the situation, such as unhappy customers. Eventually, this may lead to long-term effects.

Long-term Effect This is an effect that may not immediately occur, but if the causes are left unaddressed, will be inevitable in the future, such as a loss of customers as a result of word of mouth spreading about poor customer service.

Chain of cause and effect Very often, a single root cause will "ripple" through the situation, leading to an intermediate effect, which itself becomes the cause of another problem, which has an effect, which causes another problem, and so on. Identification of the real problem and its cause then becomes more difficult because you are forced to backtrack from the obvious problem all the way to the root cause in order to most effectively solve the problem. For example, you may be experiencing the symptom of abdominal pain. In order for the doctor to help you, she must determine why you have the pain (the cause): it could be something you ate, an ulcer, a broken rib, a bruise, or something even more serious. Each possible cause has a very different solution. However, if the cause of the pain is, say, an ulcer, what is causing the ulcer? Stress? Spicy food? Poison?

Law of Unintended Consequences Any change you introduce into a system - even one intended to solve a problem you have identified - will have a ripple effect on the rest of the system, and may make matters worse in the long run.

1.1.2 Worked Examples

The examples below are related to a memo from the CEO of Gamma Technologies, a firm that makes electronic sensors and filters for medical imaging equipment. The company is fairly large, has been around for many years, and has a varied and diverse workforce.

To:	All department managers
From:	CEO, Gamma Technologies
Date:	May 1, 2008
Re:	Working environment at Gamma

I have received a number of complaints that the working environment at Gamma is unfriendly to older workers. As a result, it is believed that older workers are leaving the company in such numbers that they are drastically underrepresented in the company. What should we do about this?

Three of the CEO's managers (Xerxes, Yanni, and Zena) wrote a response to this memo. Excerpts from these responses are explored in the examples here and are critiqued to help you understand the problem solving process more fully. This will help you learn to ask useful questions and avoid predicating your problem solving on unsubstantiated beliefs and assumptions without looking for data to help.

Example 1.1. Manager Xerxes

How Xerxes responded (an excerpt):

"...Age discrimination is clearly a problem in today's workforce and it will become even more so in the immediate future as baby-boomers begin retiring later in life than previous generations of workers, either by choice or because of the increasing difficulty of accumulating a sufficient nest egg. Attitudes toward and perceptions

of aging workers must be addressed head on. I recommend that a required-attendance series of sensitivity training classes be inaugurated immediately...”

Notice that Xerxes’ response is entirely built on the belief that age discrimination is a problem in today’s workforce. While this may be true, no evidence has been presented to support this belief, and Xerxes has not proposed any data collection that would help to evaluate it. Xerxes follows this up with a big assumption, that younger workers at Gamma have an attitude problem, namely unfriendliness, toward older workers. Besides the claim in the CEO’s memo, how do we know this is true? Again, there is no evidence presented or data collection proposed. This leads to the next assumption in the chain of reasoning, that the unfriendliness at Gamma (if it exists) to older workers is due to age discrimination. But there are many possible reasons for people to be unfriendly toward one another, and not all are discriminatory. Moreover, if there is indeed an unfriendly environment at Gamma, it is possible that it is due to the unfriendly attitude of older workers to younger workers, not the other way around.

This manager concludes with a proposal intended to fix the problem: implementing sensitivity training. Clearly, Xerxes believes that sensitivity training effectively curbs discrimination. However, there is no evidence presented that such is the case. Even if Xerxes had presented data supporting the effectiveness of sensitivity training based on studies conducted at other companies, X would have to demonstrate that Gamma fits the profiles of these other companies. Finally, this solution will be costly, in terms of money to hire the trainers and lost productivity while employees attend the training. Without knowing the facts, jumping to this solution may cost money that would have been better spent or may not even solve the problem. Even worse, this solution could generate resentment among the employees toward the management: the managers have made assumptions about their attitudes and behaviors and forced them into sessions that may be perceived as a waste of time. Ultimately, this could cost the company more than money, as employees look elsewhere for jobs.

Example 1.2. Manager Yanni

Perhaps Yanni’s response will lead the company forward more productively.

“...Underrepresentation, whether regards to gender, race, or age, is a serious matter and puts Gamma at risk of a major class-action discrimination law suit. I recommend that management immediately establish 1) a secure hotline to handle complaints and 2) a review board that will investigate such complaints and 3) a set of procedures that establish clearly what actions will be taken upon the review board’s conclusions...”

Certainly Yanni’s statement about the legal ramifications of discrimination is correct. But, as with Xerxes, Yanni has skipped much of the problem solving process and simply gone straight to proposing a solution. This solution is based on the assumption that there is an under representation of older workers at Gamma. But we have no evidence (data) as to what the representation of older workers actually is at Gamma or how it compares to that of

similar companies. We do not know if it has changed over time. If it has changed, we do not know why; we only have suspicions, and these are based entirely on a few lines in the CEO's memo. Further, this solution assumes that Gamma is at imminent risk of a class-action law suit. Even if there is under representation of older workers at Gamma, how do we know it is of sufficient proportion to precipitate a class-action law suit? Much more data needs to be collected before implementing a solution like this, since it would involve a great deal of time, effort, and expense that might not even solve the problem.

Example 1.3. Manager Zena

An excerpt from Zena's response is a little more promising:

“..It is difficult to say with certainty how much of the underrepresentation of older workers is due to an unfriendly environment and how much to other factors, such as wanting a career change or having accumulated sufficient financial resources for early retirement..”

This is still founded upon the assumption that there is an under representation of older workers at Gamma. But it does point to some positive directions for exploration, since the question is raised as to why the under representation exists, and whether it is related to age alone. This demonstrates a degree of analytic sophistication not found in the other manager's responses. This response is starting to move toward good problem solving since it explores some possible alternative explanations and might eventually lead to the collection of data rather than immediately proposing a solution.

Example 1.4. Common issues with the responses

In fact, all of the managers took as a given that what the CEO says is a problem is indeed a problem. A good manager or consultant understands that there is a perceived problem (what the boss or employee or customer or client perceives to be the problem) but also understands that the person being consulted should not propose solutions to bogus problems. The client's view of the problem will almost certainly be stated in terms of one or more assumptions and beliefs. This is to be expected, since if the client/customer knew what the situation really was, he or she would not need a consultant. It is the consultant's job to use the client's perceptions as a first approximation, a way of framing the problem, but to not buy into these perceptions unless analysis of the data supports them.

Here's one place to help you understand the process better: Look closely at the words used in the memo from the CEO. You should notice that several of these words are vague or ill-defined. Each of these offers an opportunity to ask deeper questions or to collect data:

- Up front, the CEO cites “a number of complaints.” How many complaints? From whom? How were these complaints gathered? Has the number increased over time? How many such complaints should trigger an investigation? How many do other similar size company's experience?
- The word “unfriendly” could be interpreted in many different ways. It might be anything from not smiling when passing each other in the hall to not being included in after-work social activities.

- The entire memo expresses concerns about the “older workers” at Gamma, but who counts as older?
- Right in the middle of it all the CEO gives you a big clue that this might be a snipe hunt with the phrase “it is believed...” What follows that phrase should immediately be questioned, and data should be gathered to substantiate or refute the beliefs.
- Finally, be wary of adverb phrases like “drastically under-represented.” They evoke emotional responses that could raise your concern for the issue and bypass your more analytical instincts to question the claims and investigate deeper.

While it is true that the manager/consultant is not paid for his or her unsubstantiated beliefs about what might or might not be a good solution (even to a genuine problem), one’s beliefs or intuitions can be useful tools for figuring out what data should be gathered. It may turn out that some of the assumptions managers Xerxes, Yanni, and Zena made might, in fact, be true and could be supported by the gathering and analysis of appropriate data.

1.1.3 Exploration 1A: Assumptions get in the way

The beliefs and assumptions underlying the managers' responses to the CEO's directive at Gamma are all plausible, but they are not grounded in an analysis of data. We can use assumptions like these, however, to help us determine what kinds of data we need to gather in order to explore as many dimensions of a situation as we can without assuming we know the answers. The managers' "solutions" were likewise plausible but were proposed at the final stage of the problem-solving process instead of at the problem formulation stage where they could be most useful. In similar situations, we can use our imagined solutions as a way of testing whether we have thought of all the data we need to gather in order to adequately support them.

Briefly describe how you would gather the data needed to test the assumptions/beliefs of Xerxes, Yanni, and Zena. Then complete the second table, explaining what data you could use to clarify the language used or to substantiate the ideas in the vague phrases from the CEO's memo. Once you have completed these tasks, explain why you think that the data you have described will assure that you have gotten beneath the perceived problem to the real problem (they may be one and the same, of course). NOTE: If a particular belief or assumption does not seem to be particularly helpful for collecting the kind of data you need, explain briefly why not and move on to the next one.

Manager	Belief or Assumption	Data Needed or Explanation
Xerxes	Age discrimination is a problem	
	Young workers are unfriendly toward older workers	
	Unfriendliness is due to age discrimination	
	Sensitivity training curbs age discrimination	
Yanni	Older workers are under-represented at Gamma	
	Gamma is at risk of a lawsuit	
Zena	Under-representation may be due to more than just unfriendliness	

Word or Phrase	Data Needed
“a number”	
“unfriendly”	
“older workers”	
“believed that...”	
“drastically under-represented”	

1.2 Why Data?

More often than not, the most important step in solving a real-life problem is finding out exactly what the problem is. Obvious - and sometimes not so obvious. There are times when you will be called on to work with supervisors or clients who come to you for a solution but who don't really have a clear idea of what the problem is, and it will be your job to get to the bottom of the situation by gathering the data that define the problem and that provide all that will be needed to solve the problem once it is identified. Defining the problem and coming up with a plan for gathering the appropriate data to think about the problem go hand in hand.

But how, for any given problem situation, do we recognize what data will enable us to first define the problem accurately and then later to analyze it? We need to learn how to isolate what is relevant to the problem from what is not relevant to it. We will see in the Examples and Exploration how to use the 5W+H thinking strategy to recognize and isolate relevant information from a problem situation. In most cases, the strategy must be applied twice: once to the **problem context**, in order to understand what is going on and how one might resolve it, and once to the **communicative context**, in order to understand the purpose for solving the problem and how the results are to be shared.

In the problem context, we ask questions to help understand the perception of the problem, the causes of the problem, and the consequences of the problem. In other words, they help us develop an accurate picture of chain of cause and effect involved in the problem. If the perceived problem is "poor customer service" at your fast food restaurant, one needs to know a great deal more before trying to solve the problem. Some possible questions that you might ask are listed here:

- Who is complaining about the customer service? Is it a particular type of customer, customers placing a particular order, or something else?
- When are the complaints occurring? Are they around the clock, only at certain times of day/night, or are they connected to a particular staffing arrangement?
- Where are the complaints centered? Are they at the counter, drive-through, or both?
- What does the phrase "customer complaints" even mean? Are they in regards to waiting too long for food, lack of friendliness, lack of cleanliness, or some other aspect of customer service?
- Why have these complaints just been brought up? Has something changed about the customer service?

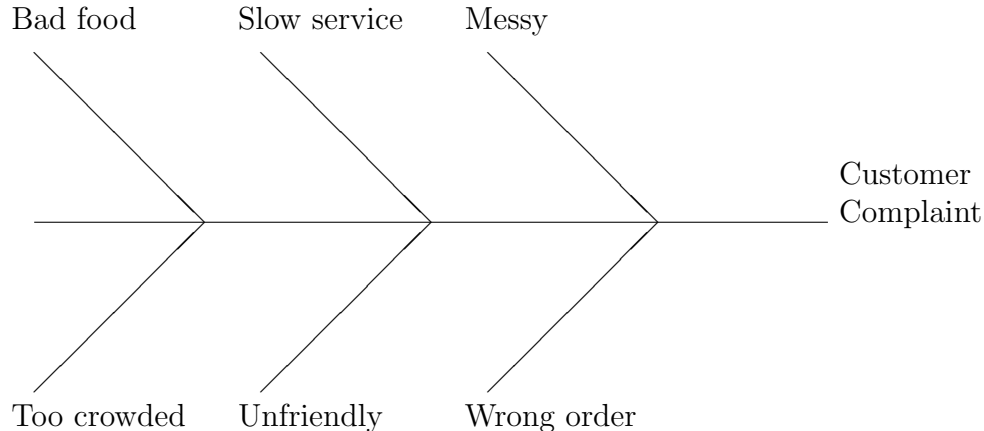
By first asking these questions, and then systematically collecting data to answer them, one can develop a better picture of the problem context. One can then attempt to resolve the problem as it actually exists. Notice that the questions above are organized into categories that come from the standard 5W+H strategy: Who, What, When, Where, Why, and How. Figuring out how to answer these questions will require data. Without relevant data in the restaurant example, a manager might be tempted to push her employees to work faster, trying to minimize service times, when the real issue is that no one is keeping the dining

area clean. Solving the wrong problem is often costly in time and money, and it still leaves the original problem unsolved.

Capturing the needed information in analyzable forms, however, is not always easy. Here are four ways to collect data that you might consider when defining the problem:

- Observations
- Survey Questionnaires
- Interviews
- Archives

We will use observations and survey questionnaires in later sections. You might consider, however, using any or all of these methods of data collection when doing the memo problem at the end of this chapter. However, there is a bigger issue that you, as the manager/problem solver will need to engage even before you begin the data collection: Making sure you are planning data that will identify the actual cause of the problem. This means getting at the **root cause** rather than at a mere symptom or a side problem. In the case of customer complaints described above, we might start this process by thinking about the restaurant as if we were a customer, and then trying to figure out what things might make us complain. After a little brainstorming, we might fill in a **fishbone diagram** like the one below.



Notice how each possible cause points to a different solution and to different data that need to be collected in order to zero in on that particular cause. This is why it is absolutely vital that you get used to engaging in root cause analysis. Such an analysis also forces you to dig deeper. You cannot just be satisfied with an answer like “customers are complaining because the service is slow.” You must dig: Why is the service slow? Is it slow everywhere or just at the counter (or drive-through)? Is it slow at all times? Slow compared to what? Have our service times changed recently? Are these changes caused by changes in staffing or changes in our business? No matter how deep you dig, there will always be more questions you can ask. Eventually, you will develop your own professional judgment regarding how deep to dig.

1.2.1 Definitions and Formulas

5W+H Strategy A method for making sure that you have considered as many different aspects of the problem situation as possible by asking six essential questions: Who? What? When? Where? Why? and How?

Problem Context This is the situation giving rise to the problem. It includes everything about the situation. For example, if there are complaints about a restaurant's service times, then the food preparation process, the layout of the store, the menu, the customers, and everything else involved in placing and picking up an order can be considered part of the problem context.

Communicative Context This is an additional aspect of solving a problem that is often ignored. The problem context describes the situation; the communicative context helps one to understand the purpose and goals for solving the problem. Typically, this is because your boss has contacted you and given you deadlines and goals for the project.

Loaded Words Very often, a problem or challenge will present itself that contains words that can be interpreted in many different ways, such as "Find the best way to do something." Words like "best" are ill-defined and should be interrogated in order to determine what they mean in this particular context: Best for whom? Best in which variables? How do we compare things to determine which is better?

Observation Either a person or some sort of mechanical process (or combination) records the occurrence of some pertinent event, usually recorded in a format specially prepared for this purpose, e.g. keeping tallies on a lined paper form or noting times in one-hour blocks or an electric eye keeping tallies and times of the traffic flow through a gate.

Survey Questionnaire A form (paper or electronic) filled out by customers, usually requires some sort of short answer or circle (check mark) of possible responses, e.g. Check: Male/Female; On a Likert scale of 1 (most liked) to 5 (least liked), circle one of the following, etc.

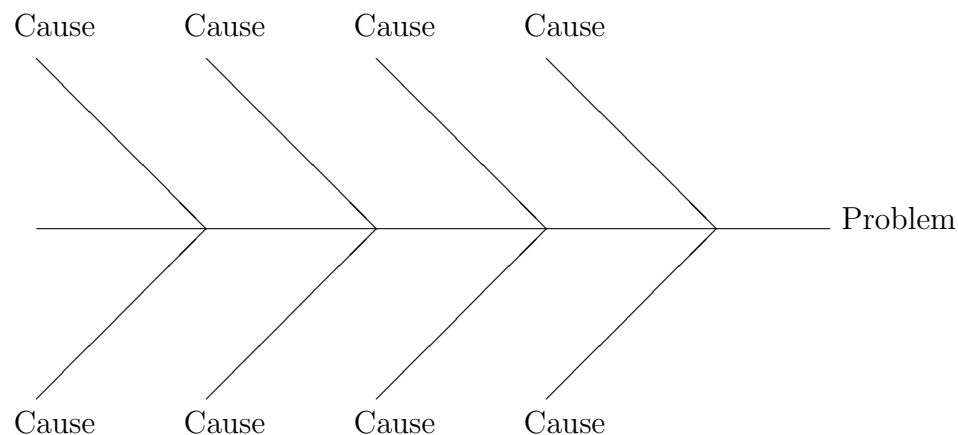
Interview Either structured (the interviewer asks all interviewees exactly the same questions) or semi-structured (the interviewer asks each interviewee the same basic questions but "goes with the flow," according to how the interviewee responds). Structured interviews lend themselves most readily to quantitative analysis and often are somewhat like a questionnaire, except they are usually longer and have certain advantages, such as the interviewer making sure that all the questions are answered and understood.

Archival Data Data that is compiled from already-existing sources, e.g. company data banks, government reports, or trade/industry tables that are available in print form, on CDs or that can be downloaded from the web.

Timeline A schedule of the events or tasks needed to complete a project along with the length of time each task will require and, usually, the personnel needed for each task.

RFP A request for a proposal. A business solicits proposals from other companies to undertake a project. The business will evaluate all submitted proposals on a competitive basis with regard to how well the proposals address the task or problem at hand and at what cost. The business will then award a contract to the company submitting the best proposal.

Fishbone Diagram A diagram that helps isolate possible causes and consequences of a perceived problem. The diagram looks like the skeleton of a fish, as seen in the figure below. Notice that the diagram has space for you to include several different causes, and it can be extended as far as needed to add more. It is also typical that a particular type of business may have generic categories of causes that you will want to explore. For example, service industries may want to think about causes coming from Policies, Procedures, People, or Plants/Technologies; manufacturing might focus on causes derived from Machines, Methods, Materials, Measurements, Mother Nature (Environment), and Manpower (People).



1.2.2 Worked Examples

These examples start with a memo from a regional manager of a fast-food chain to you, the manager of a local restaurant in the chain. This is followed by some notes on how you might respond to the memo, and a sample response memo to critique. The subsequent exploration begins with the response from the regional manager to your memo and gives you the opportunity to explore how you might address the regional manager's concerns.

Example 1.5. A Problem at Beef n' Buns

To: Local Manager, Beef n' Buns
 From: Chad R. Chez, Regional Manager, Beef n' Buns
 Date: May 8, 2008
 Re: Customer Service

I seem to be getting a lot of complaints from across the region that our Beef n' Buns service is slow. I want each of you to send me a detailed plan and a rough timeline for addressing this problem. I will review them and get back to you.

My Notes Toward a Reply There seems to be a lot going on with this memo, so I'm going to deal with it on two levels:

1. How should I deal with my boss in the context of the memo?
2. How can I come up with what he wants?

The Memo Context. This will help ensure that I do everything the boss wants in the way that he has asked for it to be done. While it doesn't resolve the problem itself, it will probably help me understand the boss better, and it will certainly help me keep my job.

WHO? My boss, the Regional Manager

WHAT? Wants me to devise a plan and a timeline for addressing the perceived problem. I'll send my response as a well-thought out memo

WHEN? ASAP (if I know what's good for me!), then I'll wait for his comments to see when I will actually have to do put the plan into action

WHERE? Sent to him

WHY? Perceived Problem: Complaints about slow service. Whether I feel these complaints are justified or not, I am responsible for addressing his concern

HOW? Send him a memo with two things: A plan and a timeline. I think I'll also send along some idea of what the project will cost, just to let him know that such things don't come free.

The Problem Situation. This is where I find out about the causes of the problem and hopefully find some ways to fix it.

WHO? My customers

WHAT? Perceived problem: slow service.

WHERE? At the drive-up window? At the walk-in service counter? In the kitchen? In the dining area?

WHEN? Does the time of day matter? Is it tied to a particular set of staff members?

WHY? To see if my restaurant has a problem with service times. Are lousy-service complaints justified at my business? Is it related to something other than service times (like cleanliness, friendliness or something else?)

HOW? How can I and my staff go about gathering data to find out about service times?

PLAN: I don't know exactly what the problem really is. I need to find out if it's a matter of unacceptably long service times at the drive-up and/or the walk-in counter. I'll need to collect data on both. I (or my staff) will have to observe, time, and record the service-time data and do some analysis of the data.

OBSERVATIONS: Where should I position my observers? When should they be there? How will they actually do it? Over what period of time should we collect data?

TIMELINE: I don't need a definite starting date for my project at this point, but I do need some kind of estimate of how long each of the tasks in my plan will take. Trying to set up a timeline really points out the missing pieces of my plan, and that is helpful. For example, who will carry out all these tasks? Seeing the overall project laid out like this also gives me an idea of the extra personnel cost and personnel scheduling problems I will encounter in carrying out the study—I will want to at least mention these things in my memo. Part of the timeline should include the time required to analyze the data. All in all, it is important to show the Regional Manager that I have thought about some of the ins and outs and that I have a realistic picture of the whole.

Example 1.6. A Proposed Plan and Timeline for Beef n' Buns

To: Chad R. Chez, Regional Manager, Beef n' Buns
 From: Local Manager, Beef n' Buns
 Date: May 8, 2008
 Re: Customer Service

This is in response to your request for a plan and timeline for determining if customers are receiving poor service at my location.

Plan My staff and I will collect service wait times at the two venues, the drive-up window and the walk-in counter. One of us will record the time each order has taken from the moment it is placed to when the completed order is delivered to the customer. We will gather the wait times during a continuous one-hour interval for the periods we are busiest, that is, breakfast, lunch, and at dinner. Here is my timeline for the project:

Task	Time	Personnel
1. Create detailed plan for data collection	1 week	1 me
2. Actual data collection	2 weeks	2 people for 3 hrs/day
3. Analysis of data	2 weeks	1 me + consultant
4. Writing of report	1 week	1 me

Although there may be some additional expenses, most of the cost of the project will come from filling the slots vacated by the observers during the period of data collection and the hours the consultant puts in. I have identified a reputable statistical consultant at our local university who charges \$50/hr and would be interested in the project.

Cost Estimate:

2 people x 3 hrs/day x 14 days @ \$10.00/hr=	\$840
1 Statistical consultant x 4 hrs x \$50/hr	= \$200
Miscellaneous expense (forms, etc)	= \$100
<hr/> Total	<hr/> =\$1140

I await your reply as to when to begin and how I should take care of the accounting.

1.2.3 Exploration 1B: Beef N' Buns Service

Consider the following response from your boss:

To: Local Manager, Beef n' Buns
From: Chad R. Chez, Regional Manager, Beef n' Buns
Date: May 11, 2008
Re: Customer Service

You are definitely headed in the right direction. Here are some things I would like you to think about for your revision.

I'm thinking that the type of order might have something to do with the wait time. We might find that some items or combination of items take longer than others. What might make one order take a significantly longer time than another? Just the size? Is there some way of comparing all these different combinations? If so, we could pinpoint the problem items and figure out ways to cut down their prep or processing times. Also, should the drive-up data be kept separate from the walk-in data or is it sufficient just to identify which is which in the same data base?

Collecting data during certain times of the day seems reasonable. Have you thought about the day of the week as a variable as well?

Have you considered that the complaints might mean something other than too-long wait times? What about customer relations? You know—friendliness, courtesy? Is there a way of getting at this possibility?

I think your timeline ought to include something about the time it is going to take to design the data collection forms and also maybe a training period for your observers. Shouldn't there be some trial runs to catch any bugs and then some time to make any necessary modifications to the collection forms? Also, what about the time it will take to enter the data into spreadsheets? Considering the amount of data you will be collecting, entry time might be significant enough to figure in the timeline. Much of the data can be captured and stored by the computer as the orders are placed and so your observers need not write down everything at the moment.

I am pleased that you thought to include the services of a statistical consultant in this project because so many of my local managers did not. I suspect that either they didn't even think of a consultant or were afraid that they didn't know enough to deal with one. As a matter of fact, I think that you might consider bringing in the consultant at other times in the process, instead of just for the data analysis.

Anyway, give me a revision of your plan and timeline ASAP. I will let you know how we will deal with the cost and the accounting of the project later on.

Make some notes as to how you would modify your plan based on the regional manager's memo and then create an expanded timeline to include his suggestions.

Notes on how to deal with the differences in orders:

Notes on collecting data on customer relations:

Revised timeline:

1.3 Homework

Mechanics and Techniques Problems

1.1. Two sets of train tracks run parallel to each other, except for a short distance where they meet and become one set of tracks over a narrow bridge. One morning, a train speeds onto the bridge. Another train coming from the opposite direction also speeds onto the bridge. Neither train can stop on the short bridge, yet there is no collision. How is this possible?

1.2. Identify the chain of cause-and-effect in the following scenario: Something...

Application and Reasoning Problems

1.3. Look again at the Gamma Technologies scenario, example 1. The CEO contends that there is a problem with age-discrimination at the company. Describe what either Manager X or Manager Y seems to believe is the the chain of cause and effect.

1.4. Let's say you are a writer for a major national newspaper. You are asked to write an article to go with the one of the following headlines. Choose one and describe your hypothesis and how you would collect data to support that hypothesis.

- Big City? Less Safe!
- Are Major League Baseball Salaries too high?
- Chicago Cubs play better at night
- Pentagon skimps on Health Care for Vets

1.5. Your boss asks you to respond to the memo below from Jenny Eggs...

To: Oracular Consulting
From: Jenny Eggs, Owner of Over-Easy Diner
Date: Today
Re: Unkind words

As you may be aware, my restaurant, Over-Easy Diner, has been serving breakfast and lunch to the citizens of this fine town for the last 50 years.

Recently I have overheard a number of comments from the servers indicating that the customers are complaining to them about the comfort of the chairs in the dining area. Last week an anonymous editorial appeared in our local paper branding us “The Worst Seat in Town”.

Can you help me solve this problem?

- What is the perceived problem at Over-Easy Diner?
- How do we know there is a problem?
- Describe 3-4 possible causes of this perceived problem.
- For each cause you identified above, what would you need to do to correct the situation?
- What are some short and long term effects of this perceived problem?

Communication and Professionalism Problems

1.6. Identify the 5W+H for the news article in figure 1.1:

DO THE BILLS PLAY BETTER IN THE COLD?

ORCHARD PARK (AP) - As the playoffs begin, Bills fans pile on the jackets and mittens to brave the bitter winds of Orchard Park and cheer on their team. Loyal fans fill up their tailgate coolers and portable barbeque cookers with food and drink to ward off the cold.

The question is, do the Bills actually play BETTER in cold weather? Perhaps the fact that Bills players have more cold-weather practice with the mittens and extra layers leads to a better outcome when the snow flies on Sunday.

We looked at the quantitative data for the 2005 and 2006 football seasons to see how the cold weather affects the players.

In order to determine the effect of the cold weather, we looked at the number of fumbles made by the team in each game. On average, the Bills made 1.73 fumbles per game during warm weather (over 40 degrees), whereas they made very slightly less (1.4) fumbles during cold weather. This is probably not a statistically significant difference.

We found that the correlation

	Average Fumbles
Cold (< 40°)	1.40
Warm (> 40°)	1.73
Total	1.63

between temperature and number of fumbles was very weak ($r=.046$), suggesting that the cold weather is not generally associated with more fumbles.

The predictive linear model **FUMBLES = .007(TEMP) + 1.321** could be used to predict the number of fumbles based on the temperature, however the coefficient of determination, R^2 , equal to .0023 further indicates that this model is not going to be very accurate. Statisticians prefer to see a coefficient of determination closer to 1.0, as it determines the percentage of variation that can be explained by the model.

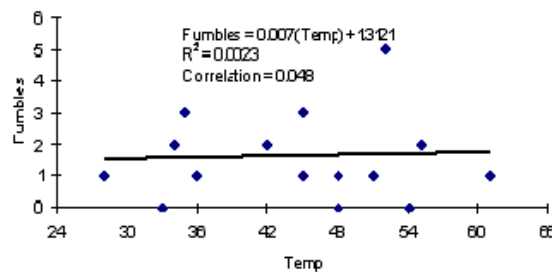


Figure 1.1: News article for Mechanics and Techniques problem 6.

CHAPTER 2

Understanding the Role of Data¹

Quantifying the world is often a bit more involved than simply determining how much there is of variable A or how many there are of variable B. The complication: “it depends.” There may be other variables C or D that need to be taken into consideration. For example, suppose you are the CEO of a large company and you want data on the salaries of your employees in order to ensure fairness and equity, provide incentives, control costs, and yet keep your company competitive. A simple approach: How much does employee 23 earn? employee 24? Etc. This is certainly useful data to have at hand—you know how much of variable A and how many of variable B. But that is not enough. As CEO, it would be much more useful for you to know, in addition, the employee’s department, years of experience at the company, job grade, educational level, age, and gender. What you really want to know is how much of A and how many of B broken down by categories C, D, E, F, G, and H. Quantifying the world, then, does not necessarily mean thinking of the world in terms of numbers only, but also in terms of categories. We will learn how to distinguish and classify various kinds of variable data in the first section of the chapter. In the second section, we will practice coding these differing data and entering the data into a spreadsheet.

- Section 2.1 takes a deep look at the different forms data can take.
- Section 2.2 explores how to organize your data to support effective analysis and problem solving - something that is much harder than it sounds.

¹©2017 Kris H. Green and W. Allen Emerson

As a result of this chapter, students will learn

- ✓ The differences between numerical and categorical data
- ✓ The importance of attending to units and categories
- ✓ How to extract data from a problem situation
- ✓ The purpose of identifiers in a data set

As a result of this chapter, students will be able to

- ✓ Design data collection forms
- ✓ Code numerical and categorical data from a data collection form
- ✓ Set up a spreadsheet for analysis
- ✓ Correctly organize data for analysis with software
- ✓ Properly define the required variable names
- ✓ Properly document information about the coding of the data

2.1 Extracting Data from the Problem Situation

In the previous chapter we learned how to define a problem. We recognized that a real-world problem is often embedded in an interconnected web of events taking place in time and space usually involving people, objects, or machines. To gather meaningful data about a problem we must think of how the data is related to its surroundings. For example, in order to gather the kind of data that we can use to identify and then correct excessive wait times at Beef n' Buns, we need to consider when a “wait time” begins and when it ends and then connect these wait times to the types of orders being filled during these wait times because not all orders are created equal with regard to wait times.

In order to gather the kinds of data that we can use to identify and then correct excessive wait times, we need to understand why not all orders are created equal with regard to wait times. And one of the first things that we recognize as we try to understand this connection is that there seems to be an inherent difference between wait-time data and type-of-order data. In this section we move ahead by learning how to recognize different types of data in a problem situation and how to record them on data collection forms. This is the process of extracting data from the problem situation.

Before we can complete the data extraction process by recording the data on data collection forms, we need to know exactly what type of data we are recording in order to know either “how many of what” to mark down or what category to check, depending on whether the data is **numerical/quantitative** or **categorical/qualitative**. (To keep it interesting, qualitative data is often referred to as **factor** data.) We will try to stay consistent in this text and refer to non-numerical data as categorical data.

Types of Data

As we mentioned above, not all data has to do with numbers. Data that does have to do with numbers, that is, counting or measuring something, is called numerical data and that which has to do with classification or categorizing something is called categorical data. Examples of numerical data are salaries, sales, heights, weights, number of customers, number of children. Examples of categorical data are gender (male, female), job classifications (e.g. office staff, management, vice president), day of week, marriage status. Sometimes it is obvious what type of data we are dealing with in a particular problem situation; other times we have to make a conscious decision as to whether we want to record our data numerically or categorically. In the latter case, we have to ask ourselves if it would be more beneficial for our analysis to retain the numerical differences between the individual things we are observing or whether it would be better to group them into categories. Each has its advantages.

Almost any type of numerical data can be converted into categorical data by some sort of classification scheme. For example, individual numerical heights could be lumped into short, medium, tall, and very tall categories by some sort of scheme, such as, all heights below 60 inches will be placed in the “short” category, all heights between 60 inches and 68 inches will be placed in the “medium” category, etc. Categorical data, however, cannot be converted to numerical data, however. So if you recorded heights of people as tall, medium, or short, you could no later go back and determine the average height of a person, since you do not have the numbers themselves. One complication is that factor data are often **coded** with

numbers, making them appear to be numerical. For example, the gender categorical data. It would not make sense to find the add-up-and-divide average of the categories “female” and “male” even if we decided to record a female in our data as “0” and a male as “1.” It would make no sense to talk about $(0+1)/2$ or .5 as gender. In general, we can distinguish numerical and categorical data by this rule of thumb: if you can do meaningful arithmetic with the data itself, it is numerical; if not, it is categorical. If all you can do numerically is some variation on counting how many of them there are, it is probably categorical.

When coding data, note that numbers can be used as codes for categorical data. So above we used 0 for female, 1 for male. We could record someone’s opinion on a scale from 1=strongly agrees to 5=strongly disagrees. Without prior knowledge or provided information, it is often difficult to distinguish between numerical data and categorical data that has been coded with numbers: E.g. Age: 59, 52, 58, 12, 43, 23. This data could either be numerical or categorical, depending on the purpose and design of the study. That is, if it were to be considered numerical, 59 would have a different impact on the sum of all the ages, for instance, than would 52, whereas if age were considered to be categorical data, then both 59 and 52 might be lumped into the “middle-aged” category, whereas 70 and 80 might be counted in the “senior” category.

Each type of data, numerical and categorical, has two subtypes. Numerical data can be either **discrete** or **continuous** and categorical data can be either **ordinal** or **nominal**. In short, continuous numerical data can take on values that fall anywhere within a continuous range of numbers, whereas discrete numerical data can only take on particular number values and nothing in between them (non-continuous); with ordinal categorical data, the categories are related by some sort of “more than” or “later than” or “better than” structure, whereas nominal categorical data (name-only categorical data) does not have any kind of inherent ordering structure (see Definitions and Formulas for examples). There are cases, however, in which some of these distinctions break down, but the point of trying to make them in the first place is that they give us more than just a way of focusing on and thinking about data as we attempt to extract it from a problem situation. They also give us the vocabulary to talk about it, especially when we are deciding how to record it.

The Units for Recording Numerical Data

Numerical data is recorded in units. In some cases, there is more than one choice for the units. For example, bottled soft drink could be measured in metric units or conventional English units. A bottle with volume 500 ML is 16.9 Fl oz., which could be measured as .5 L or as .53 qt. The business manager must be constantly aware of units. For example, if you hurriedly ran your eyes over an invoice and saw an order of 10,000 bottles of soft drink, each recorded on the invoice as having a volume of .5, you might assume that the order was for 10,000 half-quart bottles. This would have a volume of 5,000 quarts. But if the unit is actually a liter, then you would be making a large error. $10,000 \times 0.5 \text{ L} = 5,000 \text{ L}$. And since each half liter is equivalent to 0.53 quarts, this 5,000 L volume of soft drink is equivalent to $5,000 \times 0.53 / 0.5 = 5,300$ quarts. Why would such a mistake matter? After all, you actually have more soft drink than you thought, so you can sell it for more money. But if you were expecting to pay shipping on 5,000 quarts and it was really 5,300 quarts, you could be facing a much steeper delivery charge.

The issue of units, however, is more fundamental than committing oversight errors. The choice of units can change the nature of the data we are extracting from a problem context. The different units in the bottled soft drink example all measure the amount of liquid as volume. We could have measured the amount of soft drink in units measuring the mass of liquid (grams or kilograms) or its weight (in pounds). Each unit, mL or grams, measures a quantity of water, but the units of data, whether measured in volume or in weight, determine the ease with which we can use incorporate the data into other problem contexts. For example, if the soft drink is being transported, there may be a weight limit, but the units are in mL (volume). In this particular case, we could, with time and effort, make the necessary conversion from volume to weight to see if our shipment is under the weight limit. The point is that we have to give some thought as to how our data might be used in the future when we go about extracting it from its context.

Categories for Recording Non-Numerical Data

Units are usually associated only with numerical data. Non-numerical data is recorded in categories that have to be explicitly defined unless they are obvious. Gender is an example of non-numerical data whose categories are obvious when recorded as Male or Female or even when recorded as M and F. Gender data is not obvious, however, when recorded in the categories 0 or 1. In this case, we should make a note (for example, by adding a “comment” to the cell in EXCEL) that explicitly states that, for example, 0 is being used to represent Male and 1 is being used to represent Female (the numbers could, of course, be reversed for male and female).

Raw Data, Summary Data, and Computed Fields

A very important idea in data collection is the difference between the raw data, a data summary, and a computed field. **Raw data** is the data as directly collected: one set of values for each variable per observation. In newspaper articles and other readings, it is not common to display the raw data, however, as it may contain thousands (or even millions) of observations. Consider collecting data on employees at a company. The raw data might appear as in the first four columns of table 2.1. This raw data table would have one line per employee, so it could be hundreds or thousands of lines long. Such a table would be difficult to read and interpret. This raw data file would typically be large and have many entries, but it is necessary in order to do any data analysis that you have this file of raw data. Another clue that you are looking at raw data is that there should be an identifier for each set of observations (in the table below, this is the employee ID.)

Instead of presenting raw data, reports and articles that are based on data usually present the data in **summary form**. A possible summary of the data from table 2.1 is shown in table 2.1. Rather than containing one observation of each variable for each employee, the entries in a table of summary data collect together results on many of the employees in the data and do not correspond to any single employee. Table 2.1 reports the number of male or female employees and the average salaries of male and female employees in the data. In a summary, notice that we cannot tell anything about individual employees; we have information about the aggregate set of employees, instead.

Employee ID	Annual Salary \$1,000's	Gender	Height Inches	Gender Code (0=Male, 1=Female)	Height Range	Monthly Salary \$
90020	31.5	Male	68	0	Medium	2,625
90034	40.3	Female	64	1	Medium	3,358
92300	65.1	Male	72	0	Very Tall	5,425

Table 2.1: Table of raw data (salary, gender, and height) and computed data (gender code, height range, monthly salary) on employees.

Gender	Count	Average Height (inches)
Male	452	69.4
Female	309	65.6

Table 2.2: Summary of the employee data shown in table 2.1.

This example above also illustrates the idea of a **computed variable** or a **derived variable**. Table 2.1 has three computed variables: Gender Code presents the information from the raw data field “Gender” as a dummy variable which is either a 0 or 1. The new variable “height range” provides a descriptor based on the height variable, recording employees as either tall, short, or something similar. The “monthly salary” variable is calculated from the annual salary by dividing it by 12 and multiplying it by \$1,000. In these cases, someone started with the raw data on the employees in terms of their heights, genders, and salaries, then created new variables that compares the raw data (Gender as male or female; actual height in inches) to a set of values and assigns a new number or name based on the employee’s information. This might also involve a computation from existing data: Once we have the annual salary, we can compute the monthly salary easily, we just divide by 12. And while the variable contains no new information compared to the original raw data, it does show the information in a different way. This might be useful if, for example, we are trying to put together a project proposal that would involve some of these employees being assigned to the project for different amounts of time than a full year; having the monthly salary would allow us to cost out the project more accurately.

2.1.1 Definitions and Formulas

Numerical data Data that can be arithmetically combined in meaningful ways, that is, added, subtracted, multiplied, divided, or averaged. E.g. number of children, age, number of years of experience, salary, sales, acreage

Discrete numerical data This type of numerical data takes on whole number values and usually represents a count of some kind. “In-between” values do not, therefore, do not make sense. E.g. number of children, age, number of years of experience. Note: This is numerical data because adding, for example, numbers of children, ages, or years makes

sense. It is discrete because we usually round off age or years of experience to a whole number of years for data collection in business

Continuous Numerical Data Apart from rounding, this type of numerical data could theoretically take on any number of in-between values because it is not counting discrete things; rather it measures things whose magnitudes fall on a continuous scale. E.g. salary, sales, weight, acreage. Note: This is numerical data because "averaging" salaries, sales, or weights makes sense. Weight and acreage are probably the only data that clearly fall on a continuous scale, depending of course on the accuracy of the scale (tenths, hundredths, thousandths, etc). Salary and sales are considered continuous for all practical purposes, because, theoretically, they could be broken down into hundredths of a dollar (cents), which are not whole numbers.

Categorical data Data that is used to classify, type, or categorize groups of individual things. E.g. Preference rankings (1, highest preferred, 5, least), Gender (male, female); State (NY, WI, TN); Marriage status (M, U, D). Such data may be recorded (or **coded**) using any kind of symbol: numbers, words, or letters.

Ordinal categorical data In addition to classifying or categorizing, this type of data also has an inherent order that provides additional information. E.g. The numbers 1 through 5 in an opinion poll where 1 is the most preferred and 5 the least preferred. Note: This is categorical data because adding "most preferred" to "least preferred" does not make sense. Also, the integers 1-5 are not used to "count" data and hence do not constitute discrete numerical data

Nominal categorical data This type of categorical data contains no inherent order but merely classifies or categorizes information. E.g. Gender (male, female); State: NY, WI, TN; Marriage status (M, U, D)

Quantitative data Categorical data is often referred to as qualitative.

Factor data Categorical data is often referred to as factor data, with specified levels of the factor determined by your coding.

Quantitative data Numerical data is often referred to as quantitative.

Code book Categorical data must be recorded by first determining a list of codes or levels that will be used, such as M/F for gender, tall/medium/short for height, or local/regional/national for areas. The Code Book is an index of all these codes and their meanings that is shared by everyone involved in a project, to ensure consistency. A code book is often presented in table form. Here is a template for your own code books:

Variable	Type	Units/Categories	Notes
Var 1	Numerical or Categorical	Feet, dollars, Male/Female, etc.	Anything extra
Var 2	Numerical or Categorical	Feet, dollars, Male/Female, etc.	Anything extra
Var 3	Numerical or Categorical	Feet, dollars, Male/Female, etc.	Anything extra

2.1.2 Worked Examples

The worked examples below should help you decide what type of data you are extracting from a problem situation as well as the units or categories in which it should be recorded.

Example 2.1. Salary Data: Type and Units

Consider organizing data about the salaries of employees at a company. We might be interested in each employee's salary as well as his or her position with the company and experience. Our analysis, and thus our findings, will clearly depend on what data we collect, but just as importantly, the analysis will depend on how we record, or **code**, that data. Even with just a few simple variables in our data, we have many options to consider. In the first table, we record the data much as you might initially expect.

Variable	Type	Units/Categories	Notes
Employee	IDentifier	No units	Employee ID Number
Salary	Numerical continuous	Dollars (e.g. \$34856)	Annual Gross Salary
Dept	Categorical nominal	S = Sales P = Purchasing A = Accounting R = Research	Department in which employee works
YrsExp	Numerical Discrete	Years	Years of working experience (not necessarily all with this company).

There is nothing wrong with this fairly straightforward approach to recording the data. However, the salary data requires a good deal more information than probably needed, and the years experience will vary widely across the company. So one might consider simplifying these, recording the salary in thousands of dollars and treating experience as a categorical variable. Thus, an actual salary of \$34,856 would be coded as 34.9 (thousand) after rounding it off.

Note that we change how we can analyze the data we have collected pertaining to the years-of-experience above by changing the data type, that is, the way we record the data. Recording this as a number allows us to pinpoint the typical experience of an employee by computing the arithmetic mean of YrsExp because it is numerical data. We cannot do this when the data is coded categorically. Imagine if we had coded YrsExp as either “New”, “Experienced”, or “Seasoned” to indicate whether an employee had less than 3 years of experience (New), more than 15 years (Seasoned) or somewhere in between (Experienced). We can no longer compute the average experience; all we can do is count the number of employees in each category. On the other hand, the categorical coding offers us a broader picture of the company's workforce experience by counting the number of employees falling in the New, Experienced, and Seasoned categories. Such a summary of the data would be more difficult if the data were recorded in actual years of experience. For maximum flexibility, one might even consider having two variables for years of experience: In one, the experience is recorded in as in the first table, using the actual years; in the second version of the years of experience variable, it is recorded categorically to allow for easier data summaries to be

produced. In fact, one could record the actual age and also include a second variable which is computed from the first to be a description of the age.

Variable	Type	Units/Categories	Notes
Salary	Numerical continuous	Thousands of Dollars (e.g. 34.9)	Annual Gross Salary
Dept	Categorical nominal	1 = Sales 2 = Purchasing 3 = Accounting 4 = Research	Department in which employee works
YrsExp	Categorical Ordinal	New: < 3 years Junior: 3 to <10 years Middle: 10 to < 20 years Senior: 20 or more years	Years of working experience (not necessarily all with this company).

Example 2.2. Designing an observational data collection form

Consider the following request from Jenny Eggs, regarding her restaurant:

To: Oracular Consulting
 From: Jenny Eggs, Owner of Over-Easy Diner
 Date: Today
 Re: Seating complaints

As you may be aware, my restaurant, Over-Easy Diner, has been serving breakfast and lunch to the citizens of this fine town for the last 50 years. Recently I have overheard a number of comments from the servers indicating that the customers are complaining to them about the comfort of the chairs in the dining area. Last week an anonymous editorial appeared in our local paper branding us “The Worst Seat in Town”. In order to better understand the potential causes of customer discomfort, I would like for you to collect some data for me. I am particularly interested in the following:

- What are the actual seating patterns (number of people in each seating area)?
- Where did the customer sit?
- When were the customers in the restaurant?
- What are the customers’ opinions of the restaurant layout?
- What are the customers’ opinions of chair comfort?

Over Easy serves breakfast and lunch. There are three distinct seating areas, the Nook, the Cranny, and the Hole, where diners seat themselves. The manager wants to redesign

the cafeteria and would like to collect data on the seating occupancy patterns in the three dining areas every day over a two-week period beginning on Monday, June 9. Our goal is to first design an observational data collection form, including an explanation of the units and categories.

Step 1. Decide what data is to be collected

Variable	Type	Units/Categories	Notes
Date	Numerical discrete	MM/DD/YYYY	Date observations were recorded
Day of Week	Categorical	M: Mon, F: Fri T: Tues, S: Sat W: Wed, N: Sun H: Thurs	
Time	Numerical continuous	HH:MM AM/PM	
Nook	Numerical discrete	Customers	How many customers are seated in “Nook”?
Cranny	Numerical discrete	Customers	How many customers are seated in “Cranny”?
Hole	Numerical discrete	Customers	How many customers are seated in the “Hole”?

Step 2. Design an data collection form for the OBSERVATIONAL data.

A simple data collection form for seating patterns might look like the sheet above, with columns for each of the variables, and rows for each set of observations. In this case, we have an **observational form**; someone will have to look around the restaurant at particular days and times and record the data. Such observational data, no matter how they are gotten, are essential for understanding what is actually happening in a problem situation.

BLANK DATA COLLECTION FORM FOR OVER EASY

Date (MM/DD)	Day (MTWHFSN)	Time (HH:MM AM/PM)	Nook	Cranny	Hole

COMPLETED DATA COLLECTION FORM FOR OVER EASY

Date (MM/DD)	Day (MTWHFSN)	Time (HH:MM AM/PM)	Nook	Cranny	Hole
06/12	M	09:30 AM	23	24	16
06/15	H	01:00 PM	28	15	34
etc.					

Example 2.3. Designing a survey questionnaire form

The memo suggests that the cafeteria manager also wants to collect some customer preference data before remodeling the cafeteria. We need to design a **questionnaire** for this purpose. The manager will offer free juice, coffee, or side orders to induce customers to fill out the forms, one per customer. Information about the variables and ways of measuring them appears in the table below.

Variable Name	Type	Units/Categories	Notes
Survey Code	Identifier	Six-digits	Pre-printed on forms
FirstVisit	Categorical	Y=Yes, N = No	Is this your first visit?
Room	Categorical	P = Plenty E = Enough N = Need more space	Is there enough room between the tables?
ChairSize	Numerical discrete	1 to 4 (1=great, 4=terrible)	Rank the comfort of the chairs.
ChairCushion	Numerical discrete	1 to 4 (1=great, 4=terrible)	Rank the cushioning of the chairs.
ChairFit	Numerical discrete	1 to 4 (1=great, 4=terrible)	Rank the fit to the body of the chairs.
Keep	Categorical	Y=Yes (keep) N = No (combine)	Should we keep the separate areas?

A possible survey form might look like the one below. Notice that this data is all opinion data. This is why we need multiple methods of data collection to triangulate the data; this gives us information and helps us corroborate data from each of the different methods of collection.

Over Easy Customer Satisfaction Survey				
Please circle your answers:				
1. Is this your first visit to Over Easy? Yes No				
2. Is there enough room between the tables? Plenty Adequate Need more space				
3. Please rank the comfort of the chairs on a scale of 1 to 4 (1 is great; 4 is terrible)				
(a) Size:	1 Great	2	3	4 Terrible
(b) Cushioning:	1 Great	2	3	4 Terrible
(c) Fit to Body:	1 Great	2	3	4 Terrible
4. Should we keep the Nook, Cranny, and Hole areas, or should we make one large area? Yes, keep them No, make one large area Doesn't matter				
5. Any additional comments about your experience at Over Easy?				

Note: Questions 1, 2, and 4 collect categorical nominal data. Question 3 collects categorical ordinal data

2.2 Organizing data for Future Analysis

We are now at the place where we have learned something about extracting data from a problem situation and recording it on data collection forms. Recording “live” data that we have extracted from a problems situation, however, may not be the only way to gather the data we need to solve problems. Some or all of the data could have been collected by someone else and stored in computer data banks or archived in some other medium. By whatever means we have gathered our data, we will eventually need to input that data into a computer program so that we can use that program to analyze the data. The most common kind of program that is used in business to analyze data is the spreadsheet, and the most commonly used spreadsheet is Microsoft Excel. This section will teach you how to code and organize your data so you can process it with whatever data analysis tool you are most familiar.

Data should be organized in rows and columns. The intersection of a row and column is called a cell. Each column contains the data associated with a **variable**, e.g. salary, or age or gender or opinion. An **observation** is a complete row of data and contains all the information about a particular individual or a particular case of what we are studying. You may also see observations referred to as **records**.

EmpID	AnnualSalary (thousands of dollars)	Gender	Height (inches)	Dept	YrsExp (years)
90020	31.5	Male	68	Sales	5.4
90034	40.3	Female	64	Research	0.5
92300	65.1	Male	72	Admin	15.1
92305	40.1	Male	69	Sales	6.1
92307	32.6	Female	68	Admin	7.8
92455	51.9	Male	70	Sales	3.1
94500	28.9	Male	65	Research	3.2
94700	44.0	Female	62	Sales	9.1
94545	49.9	Male	71	Admin	8.3

There are a few rules that must be followed when entering data in a spreadsheet. Following these rules will help make the data **useable**, which is the primary requirement. The organization of the spreadsheet should also be done for **readability**, but not at the expense of the useability. Once the analysis is complete, one can worry about making the data or the output of the analysis look nice for presentation, but that should be the last concern. The main considerations about spreadsheet organization are these:

1. Every column of data must have a variable name at the top of its column. This is the purpose of the column headings “EmpID”, “AnnualSalary”, “Gender”, “Height”, “Dept”, and “YrsExp” in the table above. Note that when entering variable names into spreadsheets, you may need to be careful to avoid blanks spaces or non-alphanumeric characters. If so, you can use an underscore character (e.g. Annual_Salary) or run the words together as we have above.
2. Every observation should have a unique **identifier**, usually at the beginning of its row. The column “EmpID” serves this purpose in the data above, clarifying to which

employee a particular row of data refers. Such identifiers could be as simple as a sequential number that is different for each record in the data, or the name of the person represented in that record, or an address if the records are data on buildings.

3. A data cell can contain only one kind of information; that is, two variables cannot share the same cell. We will see examples of this later.
4. If the data is numerical, the units should appear in the column heading or a comment, not as part of the data entered into the cell. The information in parentheses for each variable name defines the units in the table above, e.g., years, inches, thousands of dollars, etc. It is vital that you use the same units throughout a variable. In other words, if you are recording wait times for orders, you cannot record the time in minutes for some records and seconds for others; if you do, then your analysis will be critically flawed.

2.2.1 Definitions and Formulas

Identifier Usually the leftmost column in your data, it should contain a name or other piece of information for the purpose of identifying each set of observations separately. Identifiers should be unique; that is, no two observations should have the same identifier. Examples include: names of employees, social security numbers, and home addresses. An identifier gives us a way of quickly and accurately locating all the information about a particular observation from among all the observations in the data set, something that we quite frequently have to be able to do in our analysis. Sometimes an identifier is nothing more than what its name implies, a way of identifying a particular observation, which is certainly important. In other situations, however, identifiers might be coded in a way so that they do indeed contain information that can be used for data analysis beyond their identification purpose. The point is that the analyst must be on guard when it comes to identifiers. A column of identifiers may look like data, and may even have a heading that looks like a variable name, but because they are no more than identifiers they should not be included along with the actual data when performing analysis. To do so might give rise to some very peculiar - and erroneous - results. Identifiers can be extremely helpful in the analysis phase for identifying data that may have been entered incorrectly or data that may represent outliers.

Row (Observation or Record) Each row of your data should contain the observations of the different variables that are all associated with one identifier. If data is collected on people including name, age, education level, and salary, then a complete set of information is called a record or observation of the variables. Usually the term **record** is used in databases, and the term **observation** is used in statistical settings. When the data is organized into a spreadsheet, the records usually appear as rows.

Column (Variable or Field) Each column of in your data should contain a set of observations of a single variable. In database terms, variables are called **fields**.

Coding This is the process by which the information is converted the raw form in which it was collected into data entries for analysis. For example, when collecting information on the gender of employees, the data could be coded in several ways:

- You could enter the words “Male” or “Female”
- You could enter “M” or “F”
- You could enter “0” for male, “1” for female
- You could enter “0” for female, “1” for male

The choice you make determines the way the data is coded. It is a good idea to include a comment for each variable that explains how it has been coded and what each code means.

Computed Variable A data item that is not collected directly from the problem situation, but computed based on the collected data. For example, we might collect an employee’s BirthDate, then compute his/her age as of a certain date.

Derived Variable Another name for a computed variable

Raw Data Field This is a data field that remains as it was recorded originally.

Cross-sectional data Cross sectional data is data in which the variables are all observed at some “frozen instant in time”. Each of the observations is independent of the other observations (has no effect on it). Such data is usually used to capture information about a population by cutting through the entire population and recording information on all the variables for each individual in the population.

Time Series Data If the same variables are observed at different times, then the data is time series data. Analysis of time series data is more difficult than the analysis of cross sectional data since usually the values of the variables at one time have an effect on the values of the variables at the next time they are observed. For example, if a stock closed up one day, this has an effect on the likelihood of the stock closing up the next day. This means that the observations are not independent of each other.

Population Populations are collections of individual items (people, houses, companies, countries, cars) that are being investigated. For cross-sectional data on populations, each observation in the data is for a different member of the population. For example, in collecting data on incomes for families, you could define a population to be “all families in cities with less than 100,000 people” or “all families with two children in the United States”.

Sample When collecting data, it is rare indeed to collect information from every member of a population. Usually this is impractical because of time or expense, so some portion, usually randomly chosen according to some carefully defined criteria, is sampled. Each member of the sample produces an observation of the variables in the data. However, it is possible that the sample you have collected is not representative of the entire population. It is critical that you make certain that the sample and population are as

similar as possible. When you calculate any statistical information based on a sample, you are using this information to infer the characteristics of the population. This will usually modify the statistical calculations. (For an example, see chapter 3 on the standard deviation.)

Experimental unit This is the level at which observations are made. If you are sampling individual people, then your experimental unit is people and each observation in your data will represent data from a single person.

Binning This is a way of creating codes to construct a categorical variable from a numerical variable. each category represents a consistent-size range of values. For example, salary (in thousands of dollars) could be binned as 0-25, 25.1 - 50, 50.1-75, 75.1-100, etc.

2.2.2 Worked Examples

Example 2.4. An example of poor data entry

The following spreadsheet shows an incorrect attempt to enter the data from the Data Collection Form for Seating Patterns and the Remodeling Questionnaire form that were developed previously (example 2). In order to save space, only a few observations are shown (A2-J4). The description of the categories (a piece of the code book) for each of the variables is found in cells A10-I16; normally these would appear as comments in spreadsheet in the column headings.

	A	B	C	D	E	F	G	H	I	J	K
1	DATE	DAY	TIME	NOOK	CRANNY	HOLE	TABLE SPACE	CHAIR COMFORT	AREA	COMMENTS	
2	30-Jun	W	11:30 AM	21	25	35	P	S-1, C-3, F-2	Y	Everything's great	
3	1-Jul	H	12:00 PM	35	18	32	A, N	S-2, C-2, F-2	Y	It's OK	
4	1-Jul	H	12:30 PM	30	20	15		S-1, C-2, F-1	N	Needs improvement	
5	2-Jul	F	10:00 AM	15	20	10	P	S-1, C-3, F-2	Y		
6	2-Jul	F	12:00 PM	30	10	5	N, P	S-2, C-2, F-2	DM		
7	2-Jul	F	12:30 PM	25	23	31		S-1, C-2, F-1	N	I love this place!	
8											
9											
10	Key:	DAY					TABLE SPACE	CHAIR COMFORT	AREA		
11		H: Thursday					P: Plenty of space	S: chair size	Y: Yes, keep areas		
12		S: Saturday					A: Adequate	C: chair cushioning	N: No, don't keep areas		
13		D: Sunday					N: Needs more space	F: chair fit	DM: Doesn't matter		
14											
15								1: Great			
16								4: Terrible			
17											

Figure 2.1: Example of poor data organization.

There are several major errors in the way the data has been entered into the spreadsheet in figure 2.1.

1. An observation in this spreadsheet incorrectly consists of two types of observations run together, one from the Data Collection Form for Seating Patterns and the other from

the Remodeling Questionnaire. An observation from the seating pattern form consists of counting people in the three areas at a particular time of day. An observation from the questionnaire consists of one person's opinions.

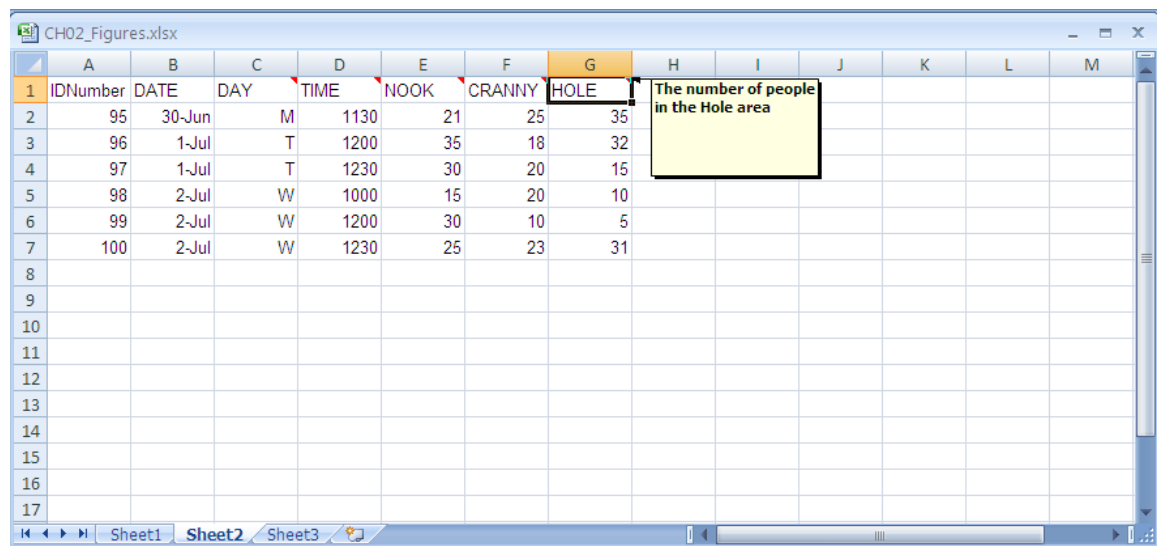
2. The observations have no meaningful identifiers. Notice that multiple records have the same date or same time, so those are not acceptable identifiers.
3. The variable names "TABLE SPACE" and "CHAIR COMFORT" each contain a space, which might cause data analysis problems, depending on the software used.
4. Cells G3, H2-H4, I3 all contain two or more entries.
5. The column under "COMMENTS" does not contain coded data.

NOTE: The next two examples and spreadsheets show a better way of entering the data from the above data collection forms. (All of these are in data file C02 *Over Easy*.) Two spreadsheets (or at least two workbooks in a single spreadsheet) are necessary because the observations cannot be combined into one spreadsheet, as we saw in example 4. Explanations for the coding of the data from each form are provided (these are not the only correct answers; there are different options for each). Four sample observations are shown in each spreadsheet.

Example 2.5. Coding the Data from the Collection Form for Seating Patterns

Our data collection form contains the following variables. Each variable is described, along with its unit and categories. The data type for each variable is described in parentheses after the description.

Variable Name	Type	Units/Categories/Notes
IDNumber	Identifier - Numeric	A one- to three-digit number identifying consecutive observations starting with 1, the first observation taken on June 30 at 11:30 A.M.
Date	Numeric - discrete	
Day	Categorical - ordinal	M: Monday, T: Tuesday, W: Wednesday, H: Thursday, F: Friday, S: Saturday, D: Sunday
Time	Numeric - discrete	The time of day will be converted to military time, where: 5:00 A.M. is 500, 5:30 A.M. is 530, 12:00 P.M. is 1200, 1:00 P.M. is 1300, 1:30 P.M. is 1330. Military time eliminates the necessity of using the A.M./P.M. designators. Although the conventional A.M./P.M. way of recording time is probably more user friendly for the people who had to collect the data, they disrupt the natural order of time necessary for analysis.
Nook	Numeric-discrete	The number of people in the Nook area
Cranny	Numeric-discrete	The number of people in the Cranny area
Hole	Numeric-discrete	The number of people in the Hole area



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	IDNumber	DATE	DAY	TIME	NOOK	CRANNY	HOLE						
2	95	30-Jun	M	1130	21	25	35						
3	96	1-Jul	T	1200	35	18	32						
4	97	1-Jul	T	1230	30	20	15						
5	98	2-Jul	W	1000	15	20	10						
6	99	2-Jul	W	1200	30	10	5						
7	100	2-Jul	W	1230	25	23	31						
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													

Figure 2.2: Observational data for Over Easy entered into a well-organized spreadsheet.

Notice that the name variable “HOLE” has been selected in the spreadsheet (figure 2.2) and a pop-up comment has been displayed describing how this variable has been coded. You can also see the other comment triangles in the upper right part of the other cells in the row. These contain the descriptions of how each variable has been coded.

Example 2.6. Coding the Data from the Remodeling Questionnaire

In the spreadsheet in figure 2.3, the codes for the data are written at the bottom of the data on the spreadsheet itself for convenience. Most often, descriptions of codes are either inserted as comments in the variable name cells (as we saw above) or written separately from the spreadsheet in the report of the analysis.

	A	B	C	D	E	F	G	H	I	J	K
1	ID	TABLE_SPACE	CHAIR_SIZE	CHAIR_CUSHIONING	CHAIR_FIT	AREAS					
2	236	P	1	3	2	Y					
3	237	A	2	1	2	DM					
4	238	N	1	2	1	N					
5	239	N	4	3	4	DM					
6											
7											
8				CODES FOR DATA							
9		P: Plenty		1: Great		Y: Keep them					
10		A: Adequate		to		N: Make one area					
11		N: Need more		4: Terrible		DM: Doesn't matter					
12											
13											
14											
15											
16											
17											

Figure 2.3: Survey data for Over Easy entered into a well-organized spreadsheet.

2.2.3 Exploration 2B: Entering Beef n' Buns Data into a Spreadsheet

In exploration 2.1.3, you designed a data collection form and a questionnaire form for Beef n' Buns. Set up a spreadsheet to capture data from these forms. Create fake data consistent with your data collection forms and enter these data into two separate data forms, one will be for the observational data and one for the survey data. Create comments for each variable name on the spreadsheet itself (as in example 6), stating its units or categories. Rename each of the sheets in the workbook with an appropriate title. Include at least ten observations of each variable. Use the space below to plan and record your thoughts.

Planning form for the observational data

	A	B	C	D	E	F
1						
2						
3						
4						
5						
6						

Planning form for the survey data

	A	B	C	D	E	F
1						
2						
3						
4						
5						
6						

2.3 Homework

Mechanics and Techniques Problems

2.1. Open the spreadsheet **C02 Homes**. This file contains data on over 270 homes that sold in the greater Rochester, NY areas during a three-month period in the year 2000. Identify each variable in the data. Classify each variable as either numerical or categorical. For numerical variables, give a rough idea of the range of the variable. For categorical variables, list each of the categories and how they are coded.

Variable Name	Type	Range/Units/Categories	Notes

2.2. Problem situation: Demand for analysts at Delphinium Consulting, Inc. is growing. Delphinium often loses its best consultants to its competitors in the industry, although consultants who stay with Delphinium for at least three years tend to stay with the company much longer.

Problem: The CEO of Delphinium is concerned about the retention of her analysts and has identified data she would like to collect below. Your job is to specify reasonable units or codes for each of these variables.

Variable	Description	Units/Codes
StartingSalary	Salary upon hiring at Delphinium	
OutOfOffice	Percentage of time consultant spends out of the office working with clients	
LocalGrad	Whether or not the employee graduated from a local university/college or not	
Major	Undergraduate major	
Tenure	Time employee has spent with the company	

2.3. In problem 2, change the numerical variables StartingSalary, OutOfOffice, and Tenure into categorical variables. For example, to change a numerical variable like TaxPercentage into a categorical variable we might define three categories:

Low less than 10%
 Middle between 10% and 20% inclusive
 High greater than 20%

2.4. Create a spreadsheet using the variables you defined in problems 2 and 4 above. Create test (fake) data for 5 observations that demonstrate the range of values for each of your variables.

Application and Reasoning Problems

2.5. Under what circumstances would it be preferable to code salary as a numerical variable? When would it be more advantageous to code it as a categorical variable?

2.6. Why might it always be preferable to record your data numerically, if possible?

Part II

Analyzing Data Through Summary Models

Key Thinking Strategy: Asking Questions

In the last unit, we focused on helping you to understand that the world is made up of information that can be organized into data. These data can then be used to help make real-world decisions. In this unit, we focus on how this data can be used to get the answers to these real-world questions. Think of this as interrogating the data in order to find out what story it tells. If we do not interrogate the data, we may find that we have collected thousands of pieces of information, have it organized nicely into a spreadsheet or a database, but have no idea as to what it all means.

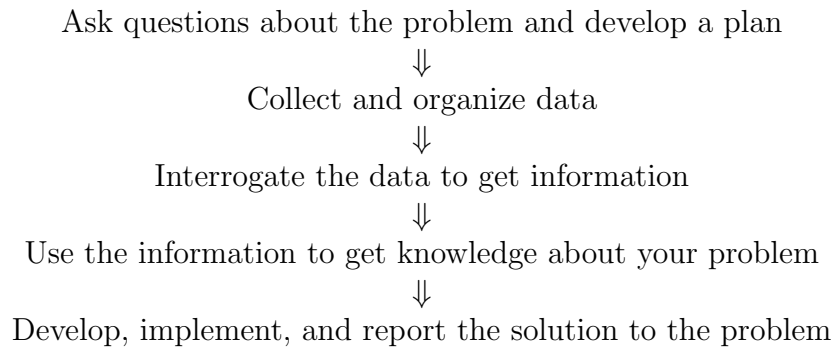
In practical terms, to interrogate the data involves asking questions and finding ways to manipulate the data to get answers. We will model the types of questions you should ask, but remember, every data set comes from a different context, so every data set will need to be asked slightly different questions. In other words, we are hoping that you will start to develop a facility with the kinds of questions that need to be asked, rather than simply going down a list of questions that someone else developed. Some people may not think that this is part of mathematics; in some sense, we agree with them. Asking questions of data is more the purview of a scientist or a detective than a mathematician. However, as a data analyst, you are a detective. Keep in mind also that there are no right or wrong questions to ask. There are questions that will take you further than others, but once you get in the habit of asking questions, you will see which questions are productive and which questions are not.

After you have asked the questions comes the part most people would consider to be mathematical: getting the answers. Usually, we will need to compute some quantity or quantities. We may carry out these computations by hand, by calculator, or using software, such as a spreadsheet. This is the part where we have right and wrong answers. For example, if we decide that the mean (a type of average) is the tool we want to use to answer a question about the data, then there is one and only one way to compute the mean; if we do not compute it correctly, our answers from that point forward will be incorrect because they are built on a mistake. We hope to help you avoid these mistakes, but you should always go back and double-check your work.

Another useful way to check your work also comes to us from police work: corroborating evidence. Finding the suspect's finger prints at the crime scene is helpful, but does not prove that the suspect was in that location when the crime happened. Finding a witness who saw the suspect enter the location at the same time the crime was committed strengthens the case. Finding that the suspect had a motive for committing the crime pretty much seals the deal. Once the police can establish motive, means, and opportunity, they consider that they have enough evidence to arrest the suspect for the crime. One calculation that supports an answer is okay, but several different quantities or representations derived from the data that lead to the same conclusion (maybe a calculation of the mean, a boxplot of the data, and a histogram of the distribution of the data) make for a much stronger case.

This idea of corroborating evidence is a little different from the way a scientist would approach the problem. A scientist typically collects data from an experiment. This experiment can be recreated and rerun several times. When all the data from each of the runs of the experiment are compared, the scientists can be satisfied they are on the right path if the results are all nearly the same. For business management data, it may be impossible to recreate the data collection method in order to get new data to compare with the old data.

Conditions will change too much between attempts to gather data, or it may be too costly. Thus, rather than looking for multiple sets of data that agree in order to settle on one set of numbers (the scientific approach), you should look for different kinds of evidence that shed light on the same question (the detective approach). This technique is called triangulation and is very common in fields of study such as management, education, and psychology.



It is critical that you also understand that this process is not a linear process. Just because we have answers to the questions we have asked does not mean that we stop. Usually, the answers will lead us to more questions. Sometimes, these answers will lead us back to the beginning and require us to design a method for collecting additional data on the situation.

Key Communication Strategy: Summarizing Data

We are inundated by data these days, whether we realize it or not. Data is being gathered, used, and thrown back at us from every direction. Just about every industry and business is being tasked to find “data-driven” or “evidence-based” reasons for doing what they do. But as we have seen in the previous unit, data can take a multitude of forms and be quite complex. So how is it that we can make any sense of all these numbers and descriptors in order to make informed decisions?

The first part of effectively using data - that is moving it from *data* to *information* - is to summarize it accurately and communicate that well. One of the goals of a good data analysis is to focus on the *descriptive statistics* to provide an accurate picture of the data using a summary. These descriptions usually involve some quantitative measures, like the mean (or another average) and spread (usually standard deviation). They may involve graphical displays of the data, like histograms, boxplots, and scatterplots. They may involve pivot tables that show relationships among the different categories within your data.

Obviously, the first part of communicating your data is to describe the data itself - what the variables are, the units of measurement or categories, and how it was all collected. The next step is to help your reader develop a model of the particular data you gathered in his or her mind so that they can better understand it. This means that you have to provide a thorough picture that includes the context of the data. As you will see, there are a huge variety of options for how one might summarize data to build these descriptive models; selecting effective tools is an important part of your job.

Also notice that we have emphasized the descriptive nature of this communication strategy. In later chapters, we’ll turn to models that help us develop more understanding of the

relationships in the data among the variables, and use those to draw inferences about the underlying context from which the data was gathered. Right now, though, these descriptive tools should lead you to asking more and more questions about the data and the situation. Answering these questions may require either more data gathering or more model building.

Memo Problem: Summary of StateEx Data

To: Analysis Staff
From: Project Management Director
Date: July 14, 2017
Re: Shipping and unloading process at StateEx

As you know, we won the contract with StateEx a while ago and have been collecting data to help the manager investigate the unloading times at his warehouse. What you may not know is that the contract has expanded and we are looking at how the unloading process works not only at the primary warehouse, which has a loading dock, but also at the endpoints of the delivery process. As you can see from the data collected (see attachment) we have aggregated some of the information from many different data collection forms into a single spreadsheet for analysis, with one row of data on each delivery. All of the variables are described in the data file.

At present, we need to give the manager of StateEx's warehouse some preliminary analysis of the data on unloading times with some recommendations about the system as it currently stands. We'll be doing deeper analysis on this later and modeling the actual relationships among the variables in the future, but for now, we need to look at the various data collected and give the manager a rough idea of what the deliveries are like. This should include numerical and graphical analyses of the data to describe what a typical delivery is like. You should also describe a typical daytime, evening, and night delivery and make recommendations on how to staff the deliveries during each of the three shifts StateEx currently runs. To do this, you may need to standardize the data with either z-scores, percentages, or both in order to ensure that you are comparing similar things across each shift. It may also help to break the data apart into the different types of trucks that are used in the deliveries.

Most importantly, though, the manager needs some information on the unloading times so that he can start to make better decisions for staffing. These data should be broken down in different ways in order to see if there are any particular types of deliveries that seem to be taking too long to unload.

Attachments: Data File StateEx_Deliveries

Using Models to Interpret Data¹

What is this chapter about? It's about taking data, possibly thousands of numbers, and finding a few measures (values) that help you make sense of the data and represent it effectively. The main tools you will use are the mean, the standard deviation, and Pivot Tables. Pivot tables are a powerful and dynamic Excel tool; there are similar tools in other spreadsheets and other software, such as the table tools in R like `ttfamily` **aggregate**.

The **mean** turns out to be the simplest and most commonly used model of data. The **standard deviation** can be thought of as a measure for how closely this model fits the data (or equivalently, how appropriate the mean is in modeling the data). Thus, we have the two basic pieces of a model: the model itself (the mean) and a measure of how well the model fits (standard deviation). Another way to think about this process is that we are taking a huge amount of information (the original data) and compressing it, reducing it to fewer pieces of information that give us a sense of the entire data set. Of course, we lose some of the information in the process, but we gain efficiency and a way of communicating and making decisions that would be extremely difficult using only the data itself. In this sense, the mean is the simplest possible model we can produce: we take all of the numerical data, no matter how numerous, and reduce it to one number for each of the numerical variables in the data set. In order to evaluate the quality of this model for each variable, we then compute the standard deviation of that variable.

- Section 3.1 shows you how to use the mean as a model for the data, and how the standard deviation is a measure of how well this model represents the data.
- Section 3.2 of the chapter shows you how to reduce data that has several variables, some of which are categorical, to several means using **Pivot Tables**.

¹©2017 Kris H. Green and W. Allen Emerson

As a result of this chapter, students will learn

- ✓ What a mean is and how it can be used to model the average or typical data point
- ✓ How to use the standard deviation as a tool for determining how well the mean represents the data
- ✓ What pivot tables are and how they are useful

As a result of this chapter, students will be able to

- ✓ Compute means and standard deviations by hand and with spreadsheet tools
- ✓ Make a Pivot Table that cross-sections your data in order to help you analyze it

3.1 The Mean As A Model

Consider what we have so far: a lot of information in the form of spreadsheets filled with data that we arranged into variables and observations. But what do we do with all this? Unless you're really special, you probably can't learn a lot from looking at a list of one thousand numbers. You probably know even less from looking at a thousand observations for each of four different variables. Sets of data in business and science are usually larger than this, so we need to think of a more efficient analysis tool. The tool we will use is to build a model of the data. A **model** is a number or formula that represents a set of data - it is not the data itself, but is meant to capture certain important features of the data that would otherwise not be recognizable in a long list of numbers.

Using models help us to understand or simplify a situation. They can also help us make predictions about future events. For example, weather models help us analyze current weather and predict potential future weather patterns. Architectural models help us visualize the design of a building before we commit it to bricks-and-mortar. In this section we will deal with what is possibly the simplest and most widely used model, called the **mean** of a set of data. Other commonly used models are given by graphs and equations, which we will develop in future chapters, eventually having models that include all sorts of features, like categorical variables.

Rather than look at the entire set of data, we want to look at the data one variable at a time in order to find out what that one variable tells us about the situation about which we collected data. To make things even easier, we want to reduce the data down to one number that represents the typical data point for that variable. In general, a number used to represent an entire variable is called a **parameter**, such as the average age of people in the United States. If we estimate that parameter based on some sample of the population, we refer to the estimate as a **statistic**. If that statistic is meant to represent the typical data point, we call it an **average**. Note that there are many ways to compute an average; the most common are the mean and median.

Selecting the appropriate measure depends on several factors, such as one intends by the term "typical" in the current setting. For example, suppose we want to talk about the average salary of employees at a company. Since the mean of a set of numbers is greatly affected by outliers in the data, the pay of a few top-tier executives like the CEO will inflate the mean salary. Thus, the management might choose to argue that the mean salary of the company is very high, and try to justify lower raises with this fact. Meanwhile, the workers might focus more on the median, which is not as influenced by outliers, and use this to show that many employees earn far less than the salary that the management claims is typical.

Let's look at another example. Shown below are the fat and protein counts for 10 of the most popular sandwiches sold at Beef n' Buns.

Item	TotalFat	Protein
Super Burger	39	29
Super Burger w/ cheese	47	34
Double Super Burger	57	48
Double Super Burger w/ Cheese	65	53
Hamburger	14	18
Cheeseburger	18	20
Double Hamburger	26	31
Double Cheeseburger	34	35
Double Cheeseburger w/ Bacon	37	38
Veggie Burger	10	14

We can reduce all this data down to the following simplified model, from which we might conclude that the “typical” sandwich has 34.7 grams of fat and 32 grams of protein.

Statistic	Total Fat	Protein
Mean (g)	34.7	32.0

The question we should ask ourselves is “*how well does the mean represent a given set of data?*” Looking at the data above, we see that although the typical sandwich has 34.7 grams of fat, there are some that have much higher values than that and some that have much less.

The first step in getting an overall measure for how the data values differ from the mean is to develop a standardized ruler to measure how close the observations are to one another. For example, in a crowd of people, your arm-length is a good measuring stick for “closeness”: If someone is less than one arm-length away from you, you would consider them “close”. However, this distance is not appropriate when driving down the freeway. A more appropriate measuring stick for this situation would be the length of a car. The Federal Aviation Administration has yet another definition of close: aircraft are not allowed within 1000 feet of each other without declaring a “near miss.”

These situations all describe ways of measuring closeness that refer to real physical distances. Seldom, however, do managers deal with these kinds of distances. More commonly, they collect data measured in dollars or years. Can we find a way to measure distance that will make sense for almost any situation that managers encounter?

As you’ve probably guessed, we can. To do so, however, we need to decide where to start measuring from. Most of the time we start measuring at zero, but this may not help very much when looking at sales figures in millions of dollars, especially if none of the figures is near zero. Rather than pick a single fixed place from which to always measure zero, it makes more sense to use a **measure of central tendency**, namely the mean for the variable.

Once we have selected the mean as the reference point we can then look at the **deviation** of each observation from the mean: Is each observation above the mean or below the mean? By how much? Thus, we will always be measuring the spread of our data from a central reference point that pertains to that particular set of data.

The measuring tool that we will use to measure the spread of our data is called the **standard deviation**. This number is different for each set of data, but it is calculated through the same formula each time.

	TotalFat (g)	Protein (g)
Mean	34.700	32.000
Standard deviation	18.209	12.561

Looking again at the Beef n' Buns data, we now have a model for the sandwiches at the restaurant: The typical sandwich has 34.7 grams of fat, and the majority of sandwiches have a fat content ranging from 26.5 grams (subtract 18.2 from 34.7) to 52.9 (add 18.2 to 34.7) grams.

You have probably encountered the standard deviation before. If you did, you may have thought that the formula was a little complicated and hard to understand. We are going to take a close look at the formula for standard deviation, because if you understand this formula you will understand a lot about statistics. Although the formula looks difficult, you will quickly learn that every piece of the formula makes sense and has a reason for being there. It wasn't developed by some genius who made the formula up from thin air. The formula was developed as the simplest possible way to find an appropriate measuring stick for any set of data. In fact, the formula for standard deviation is essentially the best way to measure the average deviation of the data from the mean.

3.1.1 Definitions and Formulas

Model A model is a number or formula that represents a set of data; it is not the data itself, but is meant to capture certain important features of the data that would otherwise not be recognizable. Models can be descriptive (used to describe a particular situation or set of data), predictive (used to help understand the likely future outcomes of a situation), or interpretive (designed to help one understand how the current situation came about or where the data came from), and can take the form of numbers, graphs, pictures, equations or descriptions.

Empirical Model An empirical model is based only on data and is used to predict, not explain, a system. An empirical model usually consists of a function that captures the trend of the data

Parameter Any number meant to capture a feature of an entire population. For example, normally distributed data can be characterized by two parameters: the mean and standard deviation. Knowing these two parameter values, together with the knowledge that the population is normally distributed, allows one to create a very accurate model of the population.

Statistic An estimate of a parameter based on a sample of a population. For example, the mean of a set of sampled data is a statistic that we hope accurately reflects the parameter mean of the underlying population from which the data was taken.

Measure of central tendency A statistic that is intended to provide a measure of what a "typical" data point is for a single variable. The most common measure of central tendency is the arithmetic average, or mean. Others include the median, mode and geometric average, all of which will be explored in the next chapter.

Mean An average computed by adding all the observations of a variable together and then dividing by the number of observations. This is more properly called the **arithmetic mean**. This is the most commonly used average, and it is the most robust average (it will change the least under repeated sampling of the population). In symbols, the mean of the data $x_1, x_2, x_3, \dots, x_n$ is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Sigma, Σ This symbol provides a compact way to represent adding a large number of items together if they follow a pattern. For example, the formula $\sum_{i=1}^5 (i + 2)$ means that we are adding together five objects that look like $i + 2$, that is, each object is a number, i , plus 2. So, the first term in the sum starts at the smallest value of i (in this case, 1) and increments up for each term. So, the nice compact formula really represents a much larger addition problem:

$$\sum_{i=1}^5 (i + 2) = (1 + 2) + (2 + 2) + (3 + 2) + (4 + 2) + (5 + 2) = 3 + 4 + 5 + 6 + 7 = 25$$

The sigma notation (the symbol is the uppercase Greek letter S, for "sum") provides a much cleaner way to write the formula. After, all, if we had to add from $i = 1$ to $i = 10,000$, writing each term out by hand would be tedious and rather pointless.

Deviation The deviation of a data point is its signed distance from the mean. To calculate this for data point x_i simply subtract the mean from the data point: $x_i - \bar{x}$. This deviation will be positive if the observation is larger than the mean and negative if the deviation is smaller than the mean.

Total Variation (SSD) This is the sum of the squares of all the deviations of all the observations in the data. In symbols, this is

$$\sum_{i=1}^n (x_i - \bar{x})^2$$

The total variation is always positive (since you are adding a bunch of squares of numbers) or zero (if each observation is equal to the mean).

Sample Standard Deviation This is a sort of average deviation for all the observations in the data. The sample standard deviation for a set of data labeled x is denoted by the symbol S_x . To compute this, we take the total variation in the data (see above), divide by the number of degrees of freedom (usually $n - 1$) and then convert back into the right units by taking the square root:

$$S_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Degrees of Freedom (DOF) This is related to several key ideas: the number of observations in your data and whether the data is from a population or from a sample. If the data is from a population, then the number of degrees of freedom is the same as the number of observations. However, if you are taking data from a sample and calculating quantities (such as the mean) that describe the population, then you lose a degree of freedom for each calculation you are inferring about the population. For example, to compute the standard deviation of a sample, you must calculate the (inferred) mean of the population. This costs you one degree of freedom, taking you from n to $n - 1$.

Outlier An outlier is a data point that atypical and may have undue influence on the statistics one computes. There are quite a few methods for identifying whether a data point should be considered an outlier. For instance, one may define any values that are more than three standard deviations from the mean to be outliers.

3.1.2 Worked Examples

Example 3.1. Computing a mean

For this example, we want to compute the mean of a set of test scores:

55, 60, 67, 70, 78, 81, 84, 88, 90, 95, 99

The mean of the data is given by adding the observations and dividing by the number of observations, $n = 11$. Thus, the mean of this data² is

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{11} = \frac{55 + 60 + 67 + 70 + 78 + 81 + 84 + 88 + 90 + 95 + 99}{11} \approx 78.82$$

Thus, we would say that *a typical student received a score of about 79 on this test.*

Example 3.2. Deviations and average deviations

Computing the deviation of a data point from the mean is simple. We just subtract the mean from the data point. The result is a signed number (it could be positive or negative) that tells us how far the data point is from the mean. So in the data for burgers at Beef n' Buns, we can compute the deviation of each burger's fat content from the mean fat content of 34.7 g.

²Notice that the last symbol before the result is a squiggly equal sign (\approx). This indicates that the answer has been rounded off.

Item	TotalFat	Deviation
Super Burger	39	4.3
Super Burger w/ cheese	47	12.3
Double Super Burger	57	22.3
Double Super Burger w/ Cheese	65	30.3
Hamburger	14	-20.7
Cheeseburger	18	-16.7
Double Hamburger	26	-8.7
Double Cheeseburger	34	-0.7
Double Cheeseburger w/ Bacon	37	2.3
Veggie Burger	10	-24.7

From this, it is clear that the Veggie Burger has much less fat, as it has a deviation of -24.7 grams. This indicates that it is very “far” away from the mean of 34.7 grams. Likewise, the Double Super Burger with cheese has a positive deviation of 30.3 grams, indicating that it is also very “fa” away from the mean (on the other side). Notice also that none of the fat contents are actually right at the mean of 34.7 grams; the Double Cheeseburger is close, but a little low. Given this, if we randomly chose a burger to eat, what would we expect its fat content to be? This is really just another way to ask what the average deviation of the fat content is.

Well, this is just an average of the deviations, right? So we can add up the deviations and divide by the number of burgers. Unfortunately, we find that the sum of the deviations is zero, giving an average deviation of zero. But how can this be? Not only do none of the burgers have exactly the mean fat content, but many of them are quite far away from the mean. (Just in case you think we have rigged this example, try it with the protein content of the burgers. Then try it with any list of numbers you want to use - your favorite football team’s points per game, for example.)

In fact, the sum of the deviations from the mean is zero, for any set of data points. We can use some algebra to decide whether this conjecture is really true.

$$\begin{aligned}\text{Old Data} &= x_i \\ \text{Old Mean} &= \bar{x} \\ \text{New Data} &= x_i - \bar{x}\end{aligned}$$

Now simply add all these data points up and compute the mean of the new (shifted) data:

$$\text{New Mean} = \frac{\sum_{i=1}^n (x_i - \bar{x})}{n} = \frac{\sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x}}{n} = \frac{\sum_{i=1}^n x_i}{n} - \frac{\sum_{i=1}^n \bar{x}}{n} = \bar{x} - \frac{n\bar{x}}{n} = 0$$

Notice that this calculation works in general. We did not need to have a specific set of data, or a specific mean, or a specific number of data points. By using algebra we can show that the mean of the deviations of any set of data is zero (and thus, the sum of the deviations is zero). This example shows both the power and beauty of mathematics. Rather than work hundreds of examples and rather than calculate each sum of deviations separately, we now have a powerful understanding of what is beneath the actual data.

Why does this happen? Another way to calculate the mean is to “take from the tall and give to the small”. The amount you take from the tall (large data values) is equal to the deviation for that stack. The amount that the small needs is the deviation for that stack, which is a negative number. So the mean is gotten by making all the deviations zero. Thus, the reason for the sum of the deviations equaling zero is related to the fact that some deviations are positive and others are negative. Adding the positive and negative numbers cancels out the deviations completely.

What does this result mean, in practical terms? It means that since the sum of the deviations is always zero, we cannot use the deviations themselves to compute an “average distance from the mean”. We must construct a new tool to measure the typical distance of an observation from the mean of the data.

Example 3.3. The Standard Deviation Formula: What it all means

The last example showed that the sum of the deviations is always zero because there is always the same total amount of positive deviation (above the mean) as there is negative deviation (below the mean). What we need is a way to turn all the deviations positive; after all, we are really interested in the average distance of the observations from the mean, not in which direction the observation falls. What ways can we take positive and negative numbers and make them all positive? If you’re like most people, you can think of at least two ways:

- Drop the negative sign from the negative deviations (this is the same as taking the **absolute value** of the deviations before you add them together)
- Square all the deviations before you add them together

For technical reasons, mathematicians prefer the second method, squaring all the deviations. If we then add up the squared deviations, we get the **total variation** of the data:

$$\text{Total Variation} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

This number by itself is not very useful. First of all, no matter what the units of the original data were, the total variation is never in those units; it’s always measured in the square of those units. Thus, if the original data were measured in dollars, the total variation would be in square-dollars, whatever those are. Second, the total variation is the sum of a bunch of squared numbers. If a number bigger than 1 is squared, the result is much larger than the original number. Thus, the total variation is often a huge number. Third, this is a total amount, not an average amount. This leads us to the next step: divide by the number of degrees of freedom to compute an average variation:

$$\text{Average Deviation} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}.$$

This number still suffers from the problems of being in the wrong units and being huge. But this is relatively easy to fix. We got the numbers larger and into the wrong units by squaring them. What’s the opposite of squaring a number? Taking the square root! (The squaring function and the square root function are **inverse functions**.) This one simple

solution will make the numbers smaller and put them into the proper units. We are then left with the sample standard deviation³ :

$$S_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}.$$

So you can see that although the formula for the standard deviation looks complicated, every piece of it is in the formula for a specific reason. For the data listed in example 2, we can determine the standard deviation. (Remember, the mean is 34.7.)

Item	TotalFat x_i	Deviation $x_i - \bar{x}$	Squared Deviation $(x_i - \bar{x})^2$
Super Burger	39	4.3	18.49
Super Burger w/ cheese	47	12.3	151.29
Double Super Burger	57	22.3	497.29
Double Super Burger w/ Cheese	65	30.3	918.09
Hamburger	14	-20.7	428.49
Cheeseburger	18	-16.7	278.89
Double Hamburger	26	-8.7	75.69
Double Cheeseburger	34	-0.7	0.49
Double Cheeseburger w/ Bacon	37	2.3	5.29
Veggie Burger	10	-24.7	610.09

$$\text{Total Variation} = \sum_{i=1}^{10} (x_i - 34.7)^2 = 2984.1.$$

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{10} (x_i - 34.7)^2}{10 - 1}} = \sqrt{\frac{2984.1}{9}} \approx \sqrt{331.57} \approx 18.21$$

This says that, on average, most data points (approximately 68% of the data points) are within 18.21 units above and 18.21 units below the mean. This would give a range of $\bar{x} - S_x = 34.7 - 18.21 = 16.49$ up to $\bar{x} + S_x = 52.91$ for most of the data. By counting the data points, we see that 6 out of the 10 data points (60%) fall inside this range. Going out to two standard deviations above and below the mean should give us 95% of the data. The lower end of that range would be $\bar{x} - 2(S_x) = 34.7 - 2(18.21) = -1.72$ and the upper end would be $\bar{x} + 2(S_x) = 71.12$. We see that the data has all 10 points (100%) within this range.

Example 3.4. Empirical Rule for normally distributed data

³One can also compute the *population standard deviation*. In this case, one must be working with data from the entire population, rather than a sample, so that the mean is not estimated but is in fact the actual mean of the population. Then you keep all n degrees of freedom. Since n is bigger than $n - 1$, the population standard deviation is always less than the sample standard deviation, indicating that we have more certainty and less variability in the population statistics.

In general, if the data is **normally distributed** we expect the following **empirical rule**⁴ to be true:

- Approximately 68% of the data points should be within one standard deviation above and below the mean. That is, 68% of the points are less than one standard deviation from the mean.
- Approximately 95% of the data points should be within two standard deviations of the mean.
- Approximately 99.7% of the data points should be within three standard deviations of the mean.
- The data should be symmetrically distributed, with about equal numbers of data points above and below the mean.

Checking to see if your data is distributed in a way that is consistent with a normal distribution can be done a number of ways. The simplest is just to compute the mean and standard deviation, and then count the percentage of data that falls in each band around the mean. You also need to check whether the data is symmetric, with about the same amount of data on each side of the mean. Rarely will the numbers exactly match the **empirical rule** but most software can give you a way to check this automatically. For now, it's enough to know that methods exist. Note that we have carefully referred to this situation as “consistent with a normal distribution” since we can never be sure it *is* a normal distribution unless we actually get the data from the entire population. It is also the basis for a popular management system known as “Six Sigma” or 6σ . This refers to the use of the symbol σ (a lowercase Greek “s”) for the standard deviation. One of the goals of the Six Sigma method is to minimize the amount of production (or whatever you are involved in that can be measured) that falls outside of three standard deviations above and below the mean. At most 0.3% of all data should fall that far away.

Example 3.5. Identifying outliers using the mean and standard deviation

The data below shows the total monthly sales for each branch of Cool Toys for Tots in two different regions of the country, the north-east region and the north-central region. (See file C03 Tots.) Which of these two regions is performing better?

⁴These rules of thumb can be derived using techniques from calculus, but it does not hurt to think of them simply as rules based on observations of lots of examples.

Sales NE	Sales NC
\$95,643.20	\$668,694.31
\$80,000.00	\$515,539.13
\$543,779.27	\$313,879.39
\$499,883.07	\$345,156.13
\$173,461.46	\$245,182.96
\$581,738.16	\$273,000.00
\$189,368.10	\$135,000.00
\$485,344.87	\$222,973.44
\$122,256.49	\$161,632.85
\$370,026.87	\$373,742.75
\$140,251.25	\$171,235.07
\$314,737.79	\$215,000.00
\$134,896.35	\$276,659.53
\$438,995.30	\$302,689.11
\$211,211.90	\$244,067.77
\$818,405.93	\$193,000.00
	\$141,903.82
	\$393,047.98
	\$507,595.76

Right off the bat, we notice that we being asked a question with a loaded word in it: better. One way we could define “better” in this context would be to compare the mean sales in each region. We find that the northeast region has mean sales of \$325,000, and the north-central region has mean sales of \$300,000.

Based on this information, we might conclude that the northeast region is doing better. But we must consider whether the mean is a good way to model this data. As a clue, when looking at the northeast region, we notice some of the lowest performing stores in the sample! And notice that there is one store in the north-east region with sales of \$818,405.93. This is much higher than the sales for the other stores in either region. This single high value is pulling the mean for the north-east region up, even though the stores in the north-central region are typically doing better, as evidenced by the fact that many of them (almost half) are well above the north-central region’s mean. In the northeast, however, the stores performing below the mean are typically far below the mean.

This sensitivity to high or low scores is one of the drawbacks of the mean. This is why the Olympics (and many other sports bodies) drop the high and low scores for a competitor before computing the mean. In later chapters, you’ll learn such **outliers** and gain a powerful graphical and numerical tools for determining which data points are likely to have too much influence on the mean.

We already know that the mean of the NE region sales is \$325,000 and the mean of the NC region is \$300,000. What about the standard deviations for each region?

	Sales NE	Sales NC
Mean	\$325,000.00	\$300,000.00
Standard Deviation	\$217,096.62	\$141,771.13

Now we have some useful information. The NE region has a much larger standard deviation than the NC region. In the NC region, though, this smaller standard deviation indicates a much narrower spread of the data. This means that the stores will perform more similarly to each other, indicating more consistency and more stable, dependable sales results in the long run. The stores in the NE region are, on average, more spread out than those in the NC region. We are likely to have very high and very low sales in this region. Notice the last store on the list for the NE region. It had sales of \$818,405.93, which is very high compared to the other stores. This store is a potential **outlier** (a data point so atypical that we should not consider it.) What happens if we remove it from the data?

	Sales NE (without outlier)	Sales NE (with outlier)
Mean	\$292,106.27	\$325,000.00
Standard Deviation	\$178,742.58	\$217,096.62

Notice the dramatic change in the results! Although the NE region (minus the outlier) is still slightly more spread out than the NC region, it is nowhere near as spread out as it was. Also, the mean for the NE region without the outlier is now slightly below the mean of the NC region. This shows you how important a single observation of the data can be. If you have outliers in the data, it is a good idea to report the statistics with and without the outliers. In this case, the \$292,106.27 value is more representative of the bulk of the stores in the NE region than the \$325,000 value, which was exaggerated by the outlier. Also, without the outlier, the stores in the NE region have more similar sales results, indicated by the smaller standard deviation.

Example 3.6. Writing meaningful paragraphs about your analysis

In this chapter, you have learned how to compose a simple model for a set of data using the mean (a measure of the center) and standard deviation (a measure of the distance to the mean). Communicating this information to your reader, who very well may not know what these terms mean, is an important skill. As an example, let's look again at the Fat and Protein content of the burgers at Beef nBuns:

Statistic	Total Fat (g)	Protein (g)
Mean	34.700	32.000
Standard deviation	18.209	12.561

A superficial paragraph communicating these descriptive statistics might be as follows:

“The average Fat for the burgers is 34.7 with a standard deviation of 18.2. The average for Protein is 32.0 with a standard deviation of 12.56.”

Remembering that your reader may not know what these statistic terms mean, this excerpt leaves a number of questions unanswered:

- Is this the average of all burgers in the whole world, or some smaller subset (context)?
- What is the unit for 34.7? Grams? Pounds? Kilograms? (units)
- What does standard deviation mean? (statistical terminology)

Lets write a more meaningful paragraph:

“The typical burger sold by Beef n Buns (context) contains 34.7 grams (units) of fat, with an average deviation of 18.2 grams. In other words, we expect that the majority of the burgers to contain between 26.5 and 52.9 grams of fat (+/- one standard deviation). On the other hand, the typical burger will contain 32.0 grams of protein, with a typical deviation of only 12.561 grams. We would expect the majority of the burgers to contain between 19.4 and 44.6 grams of protein.”

3.1.3 Exploration 3A: Wait Times at Beef n' Buns

For this exploration we will look at the waiting times collected by our observational data forms at Beef n' Buns. Assume that the following data (C03 Waittime) represents the number of seconds the customer waited to receive his/her food at Beef n' Buns. Also included is a column of data showing a sample of service times from a competitor.

1. By hand, calculate the average wait time for a typical customer.
2. How many customers waited less than the average? How many waited more?
3. By hand, compute the deviations from the mean in a new column to the right of WaitTime.
4. What was the largest deviation from the mean?
5. What is the Total Deviation (or sum of the deviations) for this set of data?
6. Manually compute the Squared Deviation in a new column to the right.
7. Manually compute the Sum of the Squared Deviations (SSD) for this set of data.
8. Use your previous calculations to compute the Standard Deviation as follows:

$$S_x = \sqrt{\frac{\text{Sum of Squared Deviations}}{n - 1}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

9. Now use built-in spreadsheet functions to compute the standard deviation.
10. Use your software to compute the mean and standard deviation for your competitor.
11. For which of the two fast food restaurants is the mean a better model for customer wait times? Why?
12. Write a meaningful paragraph comparing the waiting time at these two restaurants.

WaitTime (sec.)	Competitor (sec.)
90	210
152	0
113	118
239	0
54	185
47	0
72	16
276	43
114	17
74	165
61	23
88	9
84	134
60	26
55	22
100	26
53	36
80	273
92	83
65	186
56	109
57	140
72	48
59	56
103	132
40	183
52	153
50	30
21	72
120	104

3.2 Categorical Data and Means

The ideas in the previous section - modeling a set of data by using the mean and treating the standard deviation as the “goodness of fit” for the model - will not work if our data is categorical, because we cannot add, subtract, multiply or divide with categorical data recorded as “Male” and “Female”. Yet, we often have both categorical and numerical data and wish to know a variety of things about the underlying situation. Consider shipping records for a company that contain data on each shipment: the total weight, during what shift (morning, afternoon, night) it was loaded, truck size (van, semi, other) and so forth. We might want to profile the data to determine what a typical day-shift load looks like in order to help us plan employee schedules. How might we go about this?

One way would involve going through all the data by hand, adding up the information about just the day shift-related shipments, and then creating a report. However, typical data sets can number in the thousands of observations per variable and such a hand-compiled approach is not practicable. Another way might be to use the software to sort the data by shift and look at it that way. In a package like Excel, we could use the autofilter feature to display only the daytime shift information. But, if we also want to know what a typical afternoon and night shift load are like, we would have to repeat all this again for each shift. An additional drawback is that if the data change (a new month is added, for example) we have to start over.

Fortunately, most modern statistical and data analysis software is designed to easily allow users to cross-section their data in a variety of ways. In Excel, the main tool for this is called a Pivot Table; in R, the `aggregate` command can be used to create similar tables. These can be used to easily produce a report like the following one which has taken a spreadsheet of raw data and created a summary table for how the count of customers varies over two variables (in this case, the day of the week and the location in the restaurant). To construct this **pivot table** we have used *Day* as the Row Variable, *Location* as the Column Variable, *CustomerCount* as the Data Variable, and we have summarized these data by Average (which in Excel is the mean.)

Average of CustomerCount	Location			
Day	Cranny	Hole	Nook	Grand Total
D	26.796	19.115	28.622	24.844
F	19.734	8.770	22.135	16.879
H	16.809	8.592	22.322	15.908
M	16.763	12.056	18.849	15.889
S	19.845	15.697	22.289	19.277
T	16.974	12.260	18.414	15.883
W	13.306	8.743	21.852	14.634
Grand Total	18.604	12.176	22.069	17.616

Table 3.1: Pivot table showing average seating in the areas of Over Easy, by day of week.

This table took about 30 seconds to produce from a set of data with almost 200 observations of each variable. Even for larger sets of data, that’s about all it takes. And look

at what we can easily learn from this table. One thing that really stands out in the above summary table is that more people seem to sit in the Nook area than other other two area, as we see that the Grand Total average for Nook is 22.069 - larger than the other two locations. This could be for a variety of reasons. For example, this area might have the best window view or be closest to the entrance (or farthest from the entrance). On the other hand, the Hole area doesn't ever seem to have many people in it, with a Grand Total average over all the Days of only 12.176. This might be due to lighting (it might be too dark) or it could be located right next to the bathrooms and have lots of traffic or unwanted noise.

At the same time, we have to ask if these results are significant. Maybe this data is skewed somehow and more typical data from a longer time period (several months) wouldn't display such a large difference between the Hole and the Nook areas? To test this, one normally uses statistical tests called chi-square tests or z-tests. Our approach in this book will be to look for overall patterns rather than test for statistical significance.

In the case of Over Easy, there is another important consideration before we read too much into the pivot table: Are all three areas of equal size? If the Hole area only has seating for 20 people, then it is often close to full. At the same time, if the Nook area has seating for 50 people, it is actually not being used much at all. This is one reason why it is often vital to introduce a computed variable into the data. In this case, we should probably run statistics not on the actual counts of people in each area. Instead, we could compute the percent capacity of each area (the number of people divided by maximum capacity of the area) and then run our statistics.

3.2.1 Definitions and Formulas

Pivot Table An Excel tool for allowing you to quickly cross-section and summarize your data. Values can be displayed as means, counts, sums, and standard deviations, percentages (of rows, columns, or totals). The data can be cross-sectioned in up to three variables simultaneously (row, column, and page) and variable ranges can be grouped for easier display.

Cross-sectioning data Term used to refer to taking the data and breaking it down into categories and reporting on each category separately.

Data Mining Process of combing through data, using a variety of tools, such as pivot tables, to find information about the underlying source of the data. For example, companies mine the data on past customer purchases in order to develop targeted advertising and marketing campaigns that address the tendencies of certain types of customers.

Field In a pivot table (or a database) the variables are called fields.

Column Field This is the variable that will be used to label each column in a pivot table and to cross section the data. It is most often a categorical variable from the data. To use numerical variables here, you will want to make it into categories, for example, by creating a new version of actual Age data by breaking it into categories like 0-10, 11-20, etc.

Row Field This is the variable that will be used to label each row in a pivot table for cross-sectioning. It follows the same rules as Column Fields.

Data Field This is the variable that will be calculated in the cells of the pivot table. If it is a categorical variable, you may want to summarize by counts or percentages; if it is numerical, you will probably want to use an average or standard deviation. Other summary measures can also be applied, depending on the analysis desired.

Record In a pivot table (or a database) the observations are called records.

Raw data The raw data is all of the data collected, organized for later analysis, with one observation per row. Contrast this with **summary data** or a **computed variable**.

Goodness of fit This is a measure to determine how well a given model reflects the underlying data from which it was derived. No analysis is complete without considering and reporting on this, since an analysis based on a model that does not fit the data well is not much use. For example, the standard deviation is a way of measuring the goodness of fit for the mean; if the standard deviation is small (compared to the mean) then most data fall close to the mean and it is a reasonable model for the data.

3.2.2 Worked Examples

The following examples deal with data collected at Over Easy using the observational data forms from the last chapter. The data can be found in the file C03 OverEasy2.

Example 3.7. Reading a one-variable pivot table

Given the data on counts of people at OverEasy for the period shown in the data (contained in the variable *CustomerCount*), it might be nice to see how many people typically sit in each area of the restaurant (Nook, Cranny and Hole). We could go through the data and count these up ourselves, or sort the data and look at it, but in a few seconds, our software can produce a pivot table doing this for us. Moreover, it is then easy to change the way the data is displayed and cross-sectioned. For example, the table below shows the average number of people in each area of the restaurant. From this, we see that one area of the restaurant, the Hole section, has many fewer people in it, on average.

CustomerCount		
Location	Mean	StdDev
Cranny	18.604	12.075
Hole	12.176	9.042
Nook	22.069	12.549
Grand Total	17.616	12.046

However, we do not know how significant these differences are. It would help to know how variable the number of people in each section is. By changing the display to show the standard deviation of the counts of people in each section, we get the table below. Notice

that this shows that the overall spread of the data is about 12 people, but that the Hole area is much lower at about 9 people. So not only is the average number of people in that area low, but the spread is much less, indicating that we have consistently fewer people in that area of the restaurant.

Realistically, though, we should only be looking at the total count of people in each area if the areas are equal in size. Assuming they are, the above analysis holds; if they are not equal, and if, for example, the Hole area is smaller than the other two, then we would of course expect lower numbers in that area. In the case where the numbers have different maximum values, we need to find a number that is more stable across all the areas of the restaurant. For example, from the observational data, we could create a new variable called "PercentCapacity" that is the actual number of people divided by the maximum capacity of that area in the restaurant. These numbers would all range from 0% to 100%, making them comparable. For now, though, we assume that all the areas are the same size, which leads us to conclude that for some reason, people don't seem to want to use the Hole area of the restaurant.

To find out why this is we would need to collect more data. Perhaps people do not like the decor in that area. Or maybe it's farther from the door, so that people fill up from the front of the restaurant on back. Maybe there's no view out of the windows in that area, or it is too close to the kitchen or bathrooms, making it noisy.

Example 3.8. Pivot tables in two variables

Let us first revisit the example pivot table from the discussion above (table 3.1.) It shows the average number of people in each section of the restaurant, broken down by two variables: day of the week and location. This lets us explore more patterns in the data and see more clearly what is happening.

Notice that the table shows how much lower the counts are in the Hole area - on every day of the week. And for several days (W = Wednesday, H = Thursday, and F = Friday) the counts in the Hole are about 50% of the counts in either of the other areas. We can also see in the last column that our busiest days tend to be S = Saturday and D = Sunday. Using this information, we could more easily plan how many servers to position in each area each day of the week. But these numbers will clearly change. How much can we expect them to change? The table below displays the standard deviation of the counts, broken down by day and location. Overall, we see a lot of variation in the counts.

StdDev of CutomerCount	Location			
Day	Cranny	Hole	Nook	Grand Total
D	16.331	12.153	16.138	15.538
F	11.864	6.031	12.474	12.023
H	10.138	5.804	12.245	11.275
M	10.171	7.860	10.179	9.874
S	11.748	9.791	12.050	11.554
T	10.027	7.956	9.434	9.540
W	7.982	6.175	11.603	10.402
Grand Total	12.075	9.042	12.549	12.046

Another way to view the data in a pivot table is to look at the data not as counts (the default, usually useful only if the data variable is categorical) or as averages or standard deviations, but as a percentage. Most commonly, we would represent the data either as a percentage of the row or the column variable. For example, the table below displays the average counts, but as a percentage of the row variable (day of the week) showing us what percentage of our customers are in each of the three areas of the restaurant. Such a view of the data lets us quickly see - on a common scale - which areas are most and least popular.

These percentages might be useful for many things that the raw numbers would not directly show. For example, by looking at who is where through a percentage, we can make reasonable staffing decisions: if 25% of the customers are in the Hole on Thursday, then 25% of our staff should be in that area. In addition, this also puts all the areas on an equal footing.

Sum of CustomerCount	Location			
Day	Cranny	Hole	Nook	Grand Total
D	35.95%	25.65%	38.40%	100.00%
F	38.97%	17.32%	43.71%	100.00%
H	35.22%	18.00%	46.77%	100.00%
M	35.17%	25.29%	39.54%	100.00%
S	34.32%	27.14%	38.54%	100.00%
T	35.62%	25.73%	38.65%	100.00%
W	30.31%	19.92%	49.78%	100.00%
Grand Total	35.20%	23.04%	41.76%	100.00%

Example 3.9. Large pivot tables and grouping

Our data on the counts at Over Easy also includes the time of day. It might be interesting to look at the data using this as one of our variables, but there are a lot of times during the day. A quick two-variable pivot table looking at time and location gives us the rather long table shown below. The length is due to the large number of values for the variable “Time”. If our variable had been a continuous numerical variable instead, we could have an even bigger table, with one row for each different value of the variable.

Average of CustomerCount	Location			
Time	Cranny	Hole	Nook	Grand Total
500	3.91	2.60	5.95	4.15
530	16.69	10.84	20.01	15.85
600	31.03	20.59	36.51	29.38
630	32.83	22.33	36.94	30.70
700	28.43	18.73	32.63	26.60
730	25.19	16.88	29.48	23.85
800	16.17	10.72	19.43	15.44
830	12.51	8.36	15.90	12.26
900	6.67	4.27	8.99	6.64
930	5.42	3.69	7.63	5.58
1000	6.56	4.52	8.85	6.64
1030	8.71	5.67	11.19	8.52
1100	16.36	10.70	19.87	15.64
1130	20.63	13.17	24.20	19.33
1200	29.29	18.77	33.52	27.19
1230	31.54	19.85	36.02	29.13
1300	31.06	19.27	35.37	28.57
1330	26.52	17.85	30.84	25.07
1400	3.96	2.55	6.00	4.17
Grand Total	18.60	12.18	22.07	17.62

Large tables are generally undesirable. They are harder to read and harder to interpret. Usually, though, when you use a numerical variable as either the column or row variable in a pivot table, there is a way to group the values of the variable to make it easier to read the table. In this case, we could easily group all the times between 500 and 1000 into a “Breakfast” group all the other times into a “Lunch” group.

Average of CutomerCount		Location			
Time2	Time	Cranny	Hole	Nook	Grand Total
Group1	500	3.91	2.60	5.95	4.15
	530	16.69	10.84	20.01	15.85
	600	31.03	20.59	36.51	29.38
	630	32.83	22.33	36.94	30.70
	700	28.43	18.73	32.63	26.60
	730	25.19	16.88	29.48	23.85
	800	16.17	10.72	19.43	15.44
	830	12.51	8.36	15.90	12.26
	900	6.67	4.27	8.99	6.64
	930	5.42	3.69	7.63	5.58
Group2	1000	6.56	4.52	8.85	6.64
	1030	8.71	5.67	11.19	8.52
	1100	16.36	10.70	19.87	15.64
	1130	20.63	13.17	24.20	19.33
	1200	29.29	18.77	33.52	27.19
	1230	31.54	19.85	36.02	29.13
	1300	31.06	19.27	35.37	28.57
	1330	26.52	17.85	30.84	25.07
	1400	3.96	2.55	6.00	4.17
Grand Total		18.60	12.18	22.07	17.62

At first glance, this has not really helped. The table is in fact larger by one column. But by hiding the details of a grouped variable, we can get a much smaller table, letting us quickly compare the morning (Group 1) and noon (Group 2) rushes.

Average of CutomerCount		Location			
Time2	Time	Cranny	Hole	Nook	Grand Total
Group1		17.88	11.90	21.35	17.04
Group2		19.40	12.48	22.87	18.25
Grand Total		18.60	12.18	22.07	17.62

As a final way of looking at this data, consider using the count variable itself as a row variable. The counts in each section range from 0 to 65, leaving us with 66 rows in such a table. But by grouping them (which can easily be done with a numerical variable like “Count”) we can quickly see how many 30 minute blocks of time (observations) in each area of the restaurant fell into each grouping of the number of patrons at our restaurant.

Count of CustomerCount	Location			
Count	Cranny	Hole	Nook	Grand Total
0-9	675	1030	517	2222
10-19	529	672	464	1665
20-29	498	311	516	1325
30-39	303	102	421	826
40-49	97	12	178	287
50-59	25	1	30	56
60-69	1		2	3
Grand Total	2128	2128	2128	6384

Essentially, this type of pivot table is a frequency table of the variable “Count”. These are useful for creating histograms and other visual representations of one-variable data, which are discussed in chapter 5.

3.2.3 Exploration 3B: Gender Discrimination Analysis with Pivot Tables

To help understand how pivot tables work and can help you analyze data to explore a problem context, we will consider a small private company called EnPact that produces environmental impact statements. (Basically, when a company wants to build in an area or manufacture a product, impact statements help predict the expected impact of this work on the local ecology.) Recently, the company has been sued by a group of female employees on the grounds that males have an unfair advantage in the salary process. By exploring this problem using pivot tables, you will learn a fundamental truth about data mining: the deeper you explore, the more you are forced to reconsider each and every piece of evidence you have.

The company salary data (and employee profiles) are in the file **C03 EnPact**. Open this file. Using simple pivot tables to help you answer the following questions. For each set of questions, write a meaningful paragraph as your response.

1. How many male employees are there? How many female employees? What percentage of the employees is male? Female?
2. What is the average male salary? What is the average female salary?
3. Based on your answers to #1 and #2, write a sentence or two discussing the company's lawsuit.
4. Cross-section the data by both Gender and Education Level. Look at the average salaries of the employees and discuss the company's lawsuit.
5. Cross-section the data by both Gender and Job Level. What does the lawsuit look like now?
6. For a final look at just how complex this issue is, cross-section on three variables simultaneously. Set the pivot table up with Gender as the row variable, Education Level as the column variable, Job Level as the "page variable" (at the top of the pivot table) and average salary as the data.
7. Using the three-variable pivot table, pull down the "Job Level" menu and look at each job level separately in the pivot table. Are there any particular job levels where the male and female salaries, after accounting for education, are roughly the same? Are there any where the salaries are quite different?
8. Select one of the job levels that shows a large difference in salaries by gender. Go back to the original data. Can you account for these differences by looking at the numerical variables (Years of experience and Years Prior)?

3.3 Homework

Mechanics and Techniques Problems

3.1. Download the data file **C03 Salaries**. This data represents salaries for employees at a small company.

1. Describe the typical salary for an employee at this company, using the mean as a model. Write a meaningful paragraph about this result.
2. Add in two new columns of computed data: The first column should contain the salaries of each employee after a flat \$1000 raise. The second column should contain the salaries after a 5% raise.
3. What are the mean, median, and standard deviation of these three different salaries? (Be sure to copy and paste these statistics from your software).
4. What happened to the mean after the \$1000 increase? Why?
5. What happened to the mean after the 5% increase? Why?
6. What happened to the Standard Deviation after the \$1000 increase? Why?
7. What happened to the Standard Deviation after the 5% increase? Why?

3.2. Data file **C03 Incomes** contains a list of 1000 family incomes from each of four fictitious countries. All of the families are the same size (two parents and one child) and the families form a representative sample of such families in their country.

1. Place the four countries in order of increasing average income. Explain what this ordering tells you about these countries. Describe each country as either "Poor", "Average", or "Wealthy".
2. Place the four countries in order of increasing standard deviation. Explain what this ordering tells you about these countries. Describe each country as either "Shared Wealth" or "Disparate". Make sure you consider not just the standard deviation alone, but also how it compares with the mean income of the country.
3. Which country would you want to live in? Explain your reasoning.

3.3. These problems will involve the data set **C03 AllTron**. The data shows information on the employees at AllTron, an electronics design company. Use pivot tables to answer the following questions.

1. How many of the employees in this sample are men (gender=0 for male, 1 for female)?

2. What percentage of the employees is female?
3. What is the average salary of the male employees?
4. What is the average salary of the male employees who have exactly four years of post-secondary education?
5. What is the total number of years of experience at AllTron for the male employees?
For the female employees?

3.4. Using data file **C02 Homes**, answer each question below by preparing a pivot table. Put your answers into a Word document and be sure to include your answer in full sentences as well as the pivot table or other computations that support your answer.

1. How many of each STYLE of home do we have?
2. How many of each STYLE of home, broken down by the number of bedrooms?
3. What is the average VALUE of each STYLE of home, broken down by BED?
4. What is the Standard Deviation of VALUE, broken down by STYLE (row) and BED (column)?
5. What is the average SIZE of the houses, broken down by LOCATION (row), and BED (column)?

3.5. Using data file **C03 BeefNBuns**, answer each question below by preparing a pivot table. Be sure to include your answer in full sentences as well as the pivot table or other computations that support your answer.

1. Calculate the mean and standard deviation for the variable WaitTime.
2. Now create a pivot table and calculate the AVERAGE of WaitTime.
3. Now break that statistic down with Venue for the Row Variable (C for Counter; D for Drive-thru.)
4. Can you break that down by Complexity (as the Column Variable)?
5. How about Venue (as the Row) and Size (as the Column Variable)?
6. How about Time (Time of Day) by Size?

Application and Reasoning Problems

3.6. Consider the data and work you did in problem 1.

1. Will any fixed amount of increase result in the same change to the mean? What about a salary decrease? Explain your reasoning.
2. Will any fixed amount of increase result in the same change to the standard deviation? What about a salary decrease? Explain your reasoning.
3. Will any percentage increase result in the same change to the mean? What about a salary decrease? Explain your reasoning.
4. Will any percentage increase result in the same change to the standard deviation? What about a salary decrease? Explain your reasoning.
5. Do these data seem to follow the **empirical rule** for normal data? Explain, citing specifics.

3.7. Based on your work (and anything else you think is important to investigate) from problem 3, is it more important to have more experience or more education before working at Alltron? Support your claim.

3.8. Consider the quarterly sales figures at a national chain of pet supply stores. The stores are divided into four geographic regions (NE = Northeast, NW = Northwest, SE = Southeast, SW = Southwest).

Region	NE	NW	SE	SW
Mean Sales (thousands of dollars)	409	384	265	241
Standard Deviation in Sales (thousands of dollars)	112	77	73	120

1. Which region is performing better? Justify your answer.
2. Which region is performing the worst? Justify your answer.
3. Which typical sales figure is most trustworthy? Explain.
4. In which region do you expect to find the store with the highest sales? Explain.
5. In which region do you expect to find the store with the lowest sales? Explain.

CHAPTER 4

Box-and-Whisker Plots¹

What is this chapter about? It's about taking data - possibly thousands of numbers - and finding a few measures (values) that help you make sense of the data and represent it effectively. You are probably already familiar with many of these tools, but may not have used them in the way that we describe here.

- Section 4.1 of the chapter shows you how to reduce the data to a single number representing the central tendency of the data.
- Section 4.2 of the chapter shows you how to reduce the data to several numbers and then represent these numbers in a graph.

As a result of this chapter, students will learn *As a result of this chapter, students will be able to*

- | | |
|--|---|
| ✓ What a statistic is and what it is used for | ✓ Compute various summary statistics by hand and with software |
| ✓ What an average is and what the common ways of determining an average are | ✓ Make a boxplot by hand or with software |
| ✓ What quartiles are and what they tell you about data | ✓ Incorporate graphs from software into a Word document effectively to support your work |
| ✓ What an outlier is | ✓ Correctly use data stored in spreadsheet cells in computations |
| ✓ What a boxplot is, how to read the information in a boxplot, and how to interpret boxplots | ✓ Explain what happens to various statistics if the data is increased by a constant amount or by a fixed percentage |
| ✓ How to compare data sets in order to answer real-world problems | |

¹©2017 Kris H. Green and W. Allen Emerson

4.1 What Does “Typical” Mean?

So far, we’ve got a lot of information: spreadsheets filled with data that we arranged into variables and observations. But what do we do with all this? Unless you’re really special, you probably can’t learn a lot from looking at a list of one thousand numbers. You probably know even less from looking at a thousand observations for each of four different variables. Sets of data in business and science are usually larger than this, so we need to think of something fast.

The key is to take it slowly. Rather than look at the entire set of data, we want to look at the data one variable at a time in order to find out what that one variable tells us about the situation about which we collected data. To make things even easier, we want to reduce the data down to one number that represents the “typical” (another loaded word!) data point for that variable. In general, a number used to represent an entire variable is called a statistic. If that statistic is meant to represent the typical data point, we call it an average.

Watch out, though, the word “average” doesn’t really mean what you probably think it does. It has a much more general meaning than “add up the data and divide by the number of data points.” That’s only one method of computing an average, known properly as the **arithmetic mean**. There are many others. In this chapter, we’re interested in the three most common averages: the mean, the median, and the mode.

Another way to think of an average comes from the phrase central tendency. This refers to the middle of the data. You’ll always have some data above the average and some below it. The average is a way of talking about the middle of the data. The three described here (mean, median, mode) are the most commonly used ways to compute the middle. Each has a different meaning and has different applications. All are correct ways to compute the middle; it’s just that sometimes one is more appropriate than the others. When you go about computing an average you may need to check all three statistics (mean, median, and mode) of these in order to determine which of these will be the most appropriate measure of the typical data point.

If you’ve understood the ideas above, you might be amused by the statement below, which was issued by Joan Barb Briggs, the president of Generic University, in a moment of administrative desperation:

By the end of the next academic year, I want all of our instructors to have above average course evaluations.

4.1.1 Definitions and Formulas

Average A statistic that is intended to provide a measure of what a “typical” data point is for a single variable

Median An average computed by first ordering the observations from smallest to largest and then finding the number that splits the observations in half. Observe that this number may or may not be a data point, depending on whether there are an even or odd number of observations. 50% of the observations are less than or equal to the median and 50% are greater than or equal to the median. If there are an even number of points, the median is in between the two center numbers (see example 2)

Mode An average computed by determining which observation(s) is repeated most often (or most frequently). The mode is not necessarily unique, nor is it guaranteed to even exist. This is really only useful for discrete numerical data with a few possible values or for categorical data

Trimmed mean This is a mean in which you first remove a certain percentage of the data from the high and low ends in order to eliminate outliers

4.1.2 Worked Examples

Example 4.1. Computing Mean and Median with an Odd Number of Data Points

For this example, we want to compute the mean, median and mode of a set of test scores:

55, 60, 67, 70, 78, 81, 84, 88, 90, 95, 99

The mode is the most frequently occurring observation. Since none of the test scores are repeated, there is no mode. We computed the mean of this data in example 1 and found it to be about 78.82. Computing the median of the data requires us to put the data in order (this has been done already) and identify the data point in the middle of the ordered list. There are 11 points, so we want the 6th data point (that leaves five numbers less than that observation and five greater than that observation). This makes the median 81, which is slightly higher than the mean, indicating that many students did “above average” on the test. We call a distribution like this “skewed to the left”, since the mean is smaller than (to the left of) the median.

55	60	67	70	78	81	84	88	90	95	99
Lowest five					↑	Highest five				
observations					Median	observations				

Example 4.2. Computing Mean and Median with an Even Number of Data Points

Suppose that we have the same test scores as above, but a student who was absent finally comes to take the test. So now we have twelve test scores:

55, 60, 67, 70, 70, 78, 81, 84, 88, 90, 95, 99

We now have 70 repeated twice, making it the most frequently occurring test score, so the mode of this set of data is 70. To compute the median, we note that with 12 data points, we need to find a score between the 6th and 7th data points. This would be

$$\frac{78 + 81}{2} = 79.5.$$

55	60	67	70	70	78	81	84	88	90	95	99
Lowest five					Middle two		Highest five				
observations					observations		observations				

The mean can be computed using the same technique as above. A faster approach would be to realize that the first 11 scores (from example 1) have a mean of about 78.82. These will contribute a total of $11 \times 78.82 = 867$ to the sum of all the data. Then we add in the new data point, 70, for a total of 937, and divide by the total number of points, 12, to get the mean of the new data as approximately 78.08.

Example 4.3. Comparing Sales Performances

The data below shows the total monthly sales for each branch of Cool Toys for Tots in two different regions of the country, the north-east region and the north-central region. (See file 'C04 Tots.') Which of these two regions is performing better?

Sales NE	Sales NC
\$95,643.20	\$668,694.31
\$80,000.00	\$515,539.13
\$543,779.27	\$313,879.39
\$499,883.07	\$345,156.13
\$173,461.46	\$245,182.96
\$581,738.16	\$273,000.00
\$189,368.10	\$135,000.00
\$485,344.87	\$222,973.44
\$122,256.49	\$161,632.85
\$370,026.87	\$373,742.75
\$140,251.25	\$171,235.07
\$314,737.79	\$215,000.00
\$134,896.35	\$276,659.53
\$438,995.30	\$302,689.11
\$211,211.90	\$244,067.77
\$818,405.93	\$193,000.00
	\$141,903.82
	\$393,047.98
	\$507,595.76

One way to answer this question is to compare the mean and median sales in each region. We find that the northeast region has mean sales of \$325,000 and median sales of \$262,974.85. The north-central region has mean sales of \$300,000 and median sales of \$273,000.

Based on this information, we might have a hard time deciding which region is performing better. Notice that the mean sales favor the north-east region, indicating higher sales across the region, but the median sales favor the north-central region. In fact, there are more stores in the north-central region and half of them had sales of greater than \$273,000. This means that the top half of the stores in the north-central region are doing better in general than the top half of the stores in the north-east region.

Also notice that there is one store in the north-east region with sales of \$818,405.93. This is much higher than the sales for the other stores in either region. This single high value is pulling the mean for the north-east region up, even though the stores in the north-central region are typically doing better.

This sensitivity to high or low scores is one of the drawbacks of the mean. This is why the Olympics (and many other sports bodies) drop the high and low scores for a competitor before computing the mean. In Chapter 3B, you’ll learn what data points like this are called and gain a powerful graphic tool for determining which data points are likely to have too much influence on the mean.

4.1.3 Exploration 4A: Koduck Salary Increases

Koduck, a local company that makes pictures of water fowl, has 10 employees and needs to give raises to each of them. The company wants to know if it would be better financially (for the company) to give everyone a 3% raise or to add \$1000 to each employee's yearly salary. The yearly salaries of each employee are \$24,300; \$25,000; \$45,000; \$40,000; \$36,700; \$70,000; \$19,000; \$44,000; \$15,000; \$43,000.

1. Write down which method (3% raise or flat \$1000 increase) you think would be better.
2. For whom is your method better, the management, all the employees, or only certain employees?
3. Why did you select this option?
4. What would help you to make a more informed decision?
5. Now, try this in your software. Enter the salary data in one column, and then create formulas to compute the salaries after each of the two methods for the raise. Then, compute the mean and median of each data set using formulas for mean and median.
6. Compare the mean and median before and after each raise. What happened?
7. Explain why you think this happened.

As it turns out, there is a mathematical explanation for why each change happened the way it did. Using algebra, we can calculate what will happen to the mean and median of any set of data after a fixed amount is added to each data value or after a fixed percentage increase.

4.2 Thinking inside the box

Very often, we find that the measures of central tendency - mean, median and mode - are not enough to describe the data we are exploring. These numbers give us some idea of what a typical data point looks like, but they cannot answer questions like:

- How much of the data is less than the average? How much is more than the average?
- What is the largest value in the data? What is the smallest value?
- Where is "most" of the data? Is it close to the average?
- Which measure of central tendency best describes this data?

To answer these questions, we will need to have more tools available. This means that we need more information. If you think about it, we start with a collection of data. This might include thousands of observations of each variable. No human mind can process that much data in order to draw conclusions to make decisions. Therefore, we tried the easiest thing possible: reduce all the data down to a single statistic that represents the central tendency of the data. Now we can see some of the limitations of this approach. Any time we reduce thousands of pieces of data to a single number we have lost information about the data. Consider the following statement:

The mean number of children in a U.S. family is 2.2.

Certainly, this does not mean that every family in the U.S. is made up of 2.2 children. In fact, even to claim that the typical family has 2.2 children is a little strange since the number of children in a family is a discrete numerical quantity. Based on this statement only, which of the following statements most closely seems to describe family structures in the U.S.?

- Most families in the U.S. have two children. A few families have zero, or one child. A few more families have more than two children.
- There are more families with two or fewer children than there are families with more than two children.
- The number of families with two or fewer children is the same as the number of families with three or more children.

In fact, without more information, only the third statement can be ruled out. This one is based on the definition and computations used to compute the mean. (See if you can figure out why the third statement is definitely false.) We cannot decide which of the two remaining statements is more accurate without additional information. One common set of statistics used to get more information about a set of data are called quartiles. The idea behind quartiles is to take the data, put it in order from smallest to largest, and then break it into four quarters, each with the same number of data points in it. We then keep track of the data points at the places where the data is broken up, and we call these statistics quartiles. This gives us some idea of how the data is distributed. Graphically, we can represent the quartiles and other information about the spread of the data in a boxplot, which is a type of graph that contains about seven pieces of information to describe the data.

4.2.1 Definitions and Formulas

Minimum The smallest observation of a variable

Maximum The largest observation of a variable

Range The difference between the largest and smallest observations: $\text{Range} = \text{Maximum} - \text{Minimum}$

Quartiles These divide the data into four equal-sized groups of observations, based on an ordered list of data from smallest to largest

First quartile (Q1) The first quartile is the numerical value that exactly 25% of the observations are less than or equal to.

Third quartile (Q3) The third quartile is the numerical value that exactly 75% of the observations are less than or equal to.

Interquartile Range (IQR) The distance between the first and third quartiles: $\text{IQR} = Q3 - Q1$. Exactly 50% of the data falls inside the IQR.

Outliers These are data points that are not large enough or small enough to "fit in" with the other data. A **mild outlier** is an observation that is more than 1.5 IQR above Q3 or more than 1.5 IQR below Q1. An **extreme outlier** is an observation that is more than 3 IQR above Q3 or more than 3 IQR below Q1

Boxplot This is a graph of all the basic summary measures of a single variable. It combines all of the above information, the mean, and the median. It is sometimes called a box-and-whisker plot. A sample plot is shown below.

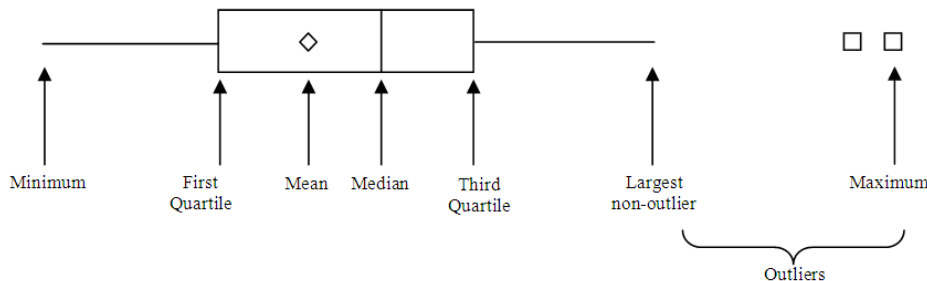


Figure 4.1: Sample boxplot (without scale) showing the major features. Note that any outliers appear past the end of the whiskers.

Side-by-side boxplot Putting several boxplots of a particular variable beside each other on the same scale, where each boxplot represents the data associated with some level of a factor variable, such as salaries of male employees shown side-by-side with those of female employees

4.2.2 Worked Examples

Example 4.4. Making a boxplot when there are an odd number of data points

Consider the list of test scores below:

55, 60, 67, 70, 78, 81, 84, 88, 90, 95, 99

We already determined that the mean of this data is 78.82 and the median is 81. We now divide the list into four equal parts to determine the quartiles. Start by dividing the data into two equal parts, as with finding the median. Then divide each of these into two equal parts. For this data, each quartile should include three data points, since there are 11 total. Notice that the middle data point, the median, is in both the upper half and the lower half of the data when we divide it up.

Lowest 50%					Median	Upper 50%				
55	60	67	70	78	81	84	88	90	95	99
Lowest 25%			Lowest 25%			Lowest 25%			Lowest 25%	

We now have almost everything that we need to make the boxplot. We just need to check whether there are any outliers in this data. An outlier is more than 1.5 IQR from Q1 or Q3. The interquartile range (IQR) for this data is $IQR = Q3 - Q1 = 89 - 68.5 = 20.5$. Thus, outliers must be more than $1.5 \times 20.5 = 30.75$ from the quartiles. Outliers on the low end would be less than $(Q1 - 30.75) = (68.5 - 30.75) = 37.75$. Outliers on the high end would be greater than $(Q3 + 30.75) = (89 + 30.75) = 119.75$. Since there are no data points outside this range, there are no outliers in this data.

To make the boxplot, we simply draw an axis scaled from 55 to 99 (for ease of reading, let's go from 50 to 100 in steps of 5). We then draw the box part of the graph, extending from Q1 to Q3. We put a vertical line in the box at the median. We add a star or a diamond for the mean, and then we extend the “whiskers” of the box from the edges out to the minimum and maximum, since there are no outliers. The final result is shown in figure 4.2

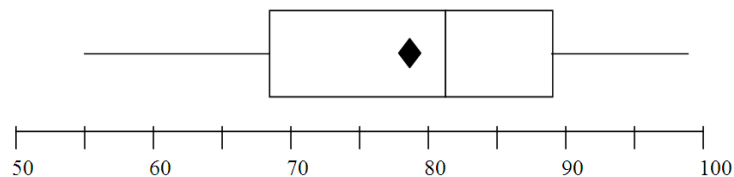


Figure 4.2: Boxplot of test scores

Example 4.5. Reading an Interpreting a Boxplot

Consider the boxplot shown in figure 4.3. It represents the distribution of test scores in a class of 120 students. What can we learn about the class performance from this graph?

To analyze the graph, we will consider a series of questions:

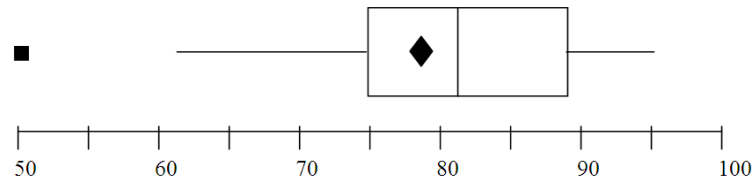


Figure 4.3: Boxplot of test scores for 120 students. What does it say about class performance?

1. What is the minimum test score? What is the maximum? What is the range? From the data, the lowest score is a 50 (which is an outlier) and the highest score is a 95. The range is $95 - 50 = 45$.
2. What are the quartiles of the data? The first quartile is 75. The median (second quartile) is about 82 and the third quartile is about 89. The IQR is $89 - 75 = 14$. This number is about one-third of the range, indicating a fairly tight spread of data (lots of similar test scores).
3. How many students scored between Q1 and Q3? We know that 50% of the data is always between Q1 and Q3. This means that 50% of the observations (in this case student test scores) fall between 75 and 89. Since 50% of 120 (the total number of observations) is $0.50 \times 120 = 60$, we know that 60 students scored between 75 and 89 on the test. However, this is a little misleading. It is possible that there are multiple students with the same tests score. If these duplications happen to be at the quartiles, then a few more students would be in the 75 to 89 range.
4. Assuming that a score of 90 is sufficient to earn an "A", how many students got an "A" on the test? This is harder to answer. Notice that the third quartile is 89. This means that 25% of the class ($0.25 \times 120 = 30$ students) got an 89 or higher. So we only really know that at most 30 students earned an "A". It is possible that most of the scores in the third quartile are right at 89 and only a few of them are between 89 and 95, which would lead to a smaller number of A's on the test.
5. Did most students do well or poorly on this test? We see that the median is slightly higher than the mean. This shows that more than 50% of the class earned a score above the mean. We do not know exactly what percentage scored above the mean, only that it is between 50% and 75% (since the mean is between Q1 and the median). This indicates that the data is negatively skewed, so that more of it is piled up above the mean than below it. Overall, then, it seems that the class did a little better than average on this test.
6. What other questions could we ask about the data?

Example 4.6. Side-by-Side Boxplots

Consider the sales data given above in example 3. (Data file C04 Tots.) Let's use boxplots to compare the two sales regions and select the region that has the better performance. If

you enter the data above into your software and create side-by-side boxplots you should get a graph similar to the one in figure 4.4.

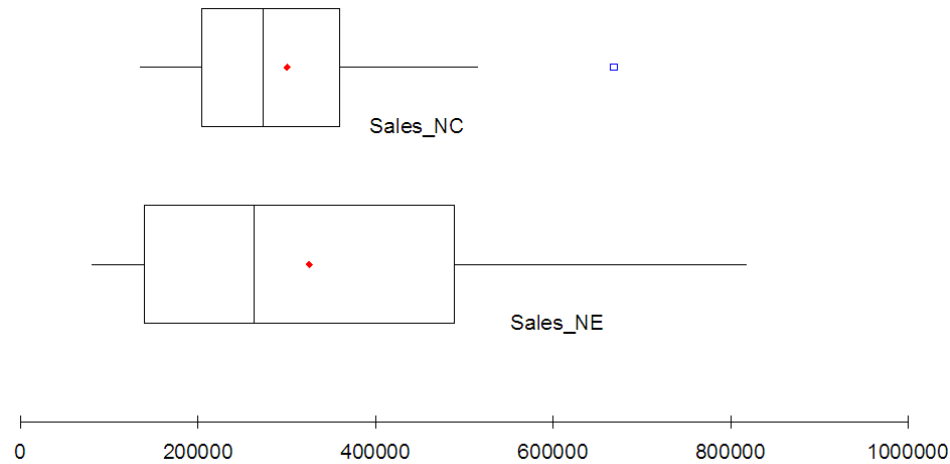


Figure 4.4: Side-by-side boxplots of sales from two regions of the Toys-For-Tots company.

As you can see, the boxplot shows that the sales in the North East region are spread over a much greater range of sales figures than the sales in the North Central region. In addition, the highest performing store in the NC region is an outlier and is not at all representative of the region's performance. However, the lowest 25% of the stores in the NE region are performing worse than all of the stores in the NC region (the minimum for the NC region is about equal to the first quartile for the NE region). By the same token, the upper 25% of stores in the NE region seem to be doing better than all the stores in the NC region (the third quartile of NE region sales is about equal to the maximum for the NC region, if we ignore the outlier). The middle of each region seems to be about the same, with the medians of the two regions almost equal. The mean sales of the NE region (indicated by the small dot) are higher than the mean sales in the MC region, but not by a very significant amount.

Given just these graphs, it might be difficult to determine which region is performing better overall. In general though, it seems that the NE region has more stores performing well than the NC region. Also, the highest performing store is in the NE region. Overall, it looks like the NE region has better sales, but we must remember that the NE region has fewer stores, so each quartile refers to fewer data points. The real question is what is causing the NE region to do better. Is it better management? Less overhead? Wealthier clients? Better marketing? Better service? Some other factor?

Example 4.7. Skewness of data

Generic University offers three sections of its math course for business majors. At the end of the semester, all sections take the same final exam. The boxplots in figure 4.5 show the results of the final separately for each section. The graphs are oriented vertically, rather than horizontally, just for variety. As you can see from the graphs, the minimum scores are the same in each section, as are the maximum scores and the means. Which section did best on the final exam?

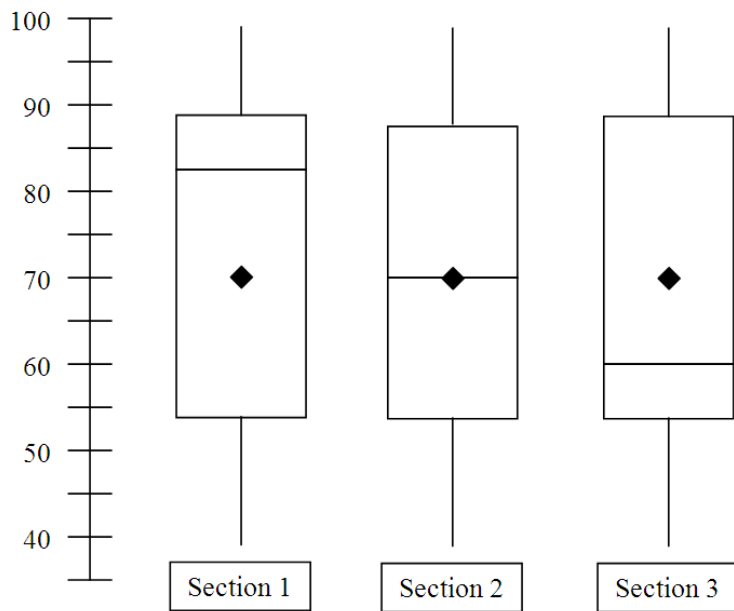


Figure 4.5: Side-by-side boxplots of final exam scores from three sections of a course.

In order to decide which section did best on the final exam, we need to picture how the data itself looks, based on the boxplots. For example, the test scores in section 1 seem to be unevenly spread throughout the range of the data (low score: 40, high score: 99). We can tell this because the median of the data is very close to the upper end of the spread. Half of the students in section 1 scored above 83 on the exam. Even so, the overall mean of this section's test scores was only 70 because the lower 50% of the class has scores from 83 on down to 40. This unevenness is referred to as skewness. When the mean is smaller than the median, we say the data is negatively skewed because the quantity (Mean - Median) would be less than zero. This is in stark contrast with section 3, where half of the students' scores are bunched together at the low end of the spread, from 40 to 60, and the top half of the class has scores ranging from 60 up to 99. In this case, the mean is larger than the median, so the data is positively skewed. What about section 2? The data for this section doesn't seem to be skewed at all; the mean and median are identical. This tells us that half of the students in section 2 scored above 70 and half of the students in section 2 scored below 70. Given all of this, it seems reasonable to conclude that section 1 had the best showing on the exam; more than half of the students in section 1 had exam scores above the median and mean scores of both sections 2 and 3.

4.2.3 Exploration 4B: Relationships Among Data, Statistics, and Boxplots

For this exploration, we'll go back to the sales data for Cool Toys for Tots in two sales regions. The data is shown above in example 3 and can be found in **C04 Tots**. To set the data up for the exploration, do the following:

- Use software to compute the mean, median, minimum, maximum, range, first quartile, third quartile and interquartile range for both regions. Place the statistics to the right of the data (not on a new worksheet, as usual).
- Create side-by-side boxplots for the two regions.

If your software allows it, place the statistics and the boxplots on the same screen with the data, so that you can explore what happens to the statistics and the boxplots as the sales information changes. You'll want to keep notes on what you observe happen (if anything) as you change the data in cells A1:B20. Explore the following questions.

1. What happens to the results (statistics and boxplots) as you change the sales figures? Be sure to keep notes on what kinds of changes you made. It may help to organize these notes into a table with three columns labeled "Change I Made", "Change in Statistics", "Change in Boxplot".
2. What happens to the results (statistics and boxplots) if you decrease the number of stores in the data? What happens if you increase the number of stores? Can you explain this behavior?
3. What happens if many of the stores in the NE region have sales above \$500,000? What if many of the stores in the NC region have sales below \$225,000?
4. What changes need to be made so that the NE and NC regions perform about the same? What changes will make the NC region perform better? (Careful! There are easy answers to these questions, but go deep and find more realistic ways to get the results.)

4.3 Homework

Mechanics and Techniques Problems

4.1. Download the data file **C04 Salaries**. This data represents salaries for employees at a small company.

1. Add in two new columns of computed data: The first column should contain the salaries of each employee after a flat \$1000 raise. The second column should contain the salaries after a 5% raise.
2. What are median and quartiles of these three different salaries? (Be sure to copy and paste these statistics from your software).
3. What happened to the median and quartiles after the \$1000 increase? Why?
4. What happened to the median and quartiles after the 5% increase? Why?
5. Describe (in words) how you think the boxplots would look different from the original boxplot for both (i) the fixed salary increase and (ii) the percent salary increase.

4.2. The boxplots below (figure 4.6) provide information about the people who tend to purchase your company's products. These data are reported as boxplots, one for the ages of the customers, one for the incomes, and one for the typical monthly credit card debt the customers carry. Use these boxplots to describe your typical customer. Make explicit reference to the quantities you can read from the boxplot directly and use these to describe your company's typical customer.

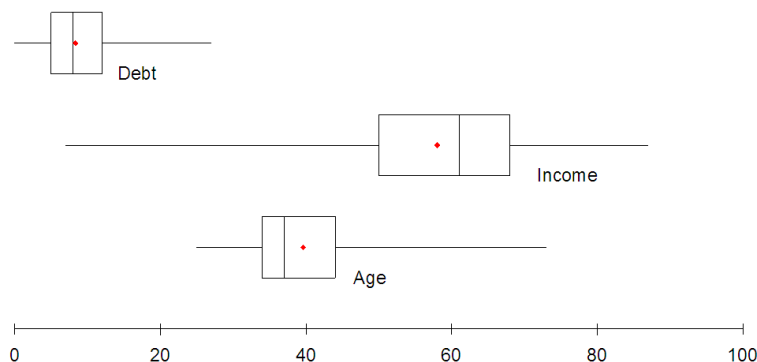


Figure 4.6: Boxplot for problem 2 showing income, age, and credit card debt distributions.

4.3. Consider the data shown in **C04 MachineParts**. This data shows the diameter of 1,000 rods manufactured on your company's assembly line. The rods must be within 0.03 inches of being 0.50 inches in diameter to fit in the structure for which they are made.

1. Create a boxplot of this data. Determine how many data points are extreme outliers and how many data points are mild outliers.
2. Sort the original data and locate all the extreme outliers. Make a new column containing all the data except these outliers. Make a boxplot of the data without the extreme outliers.
3. Are there any outliers in the reduced data from part b? If so, eliminate them and redraw the boxplot. Continue doing this until there are no outliers in the data. (Hint: This should take two more rounds of eliminating mild and extreme outliers.)
4. Compare your final boxplot (with no outliers) to the original boxplot from part a. What can you learn about the data?
5. When reporting these data, should you include the outliers? Why or why not?

Application and Reasoning Problems

4.4. Place the data $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$ along the horizontal axis of the boxplot in figure 4.7 so that it would produce a boxplot similar to the one shown. Assume that $X_1 < X_2 < X_3 < X_4 < X_5 < X_6 < X_7 < X_8 < X_9$.

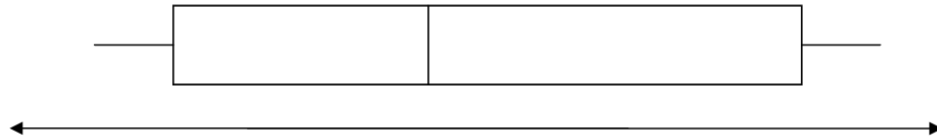


Figure 4.7: Boxplot for problem 4.

4.5. Place the data $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$ along the horizontal axis of the boxplot in figure 4.8 so that it would produce a boxplot similar to the one shown. Assume that $X_1 < X_2 < X_3 < X_4 < X_5 < X_6 < X_7 < X_8$.

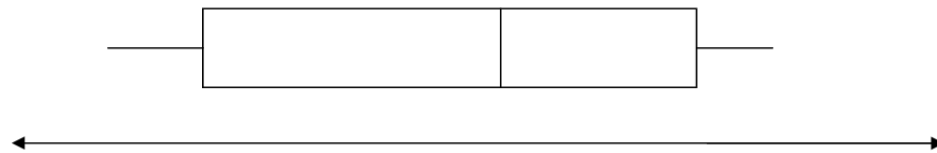


Figure 4.8: Boxplot for problem 5.

4.6. In figure 4.9, boxplot A represents the set Data 1 and Boxplot B represents the set Data 2. Given that the min and max of both data sets are the same answer the following questions:

1. In general, why is the length of box B longer than the length of box A?
2. Why is the median of B off to the right compared to A that is more central?
3. Even though the max of A and B are equal, why does the max of A appear as an isolated dot to the right of the whisker where as the max of B appears as the endpoint of the whisker? (Do not use calculations; give an eyeball explanation.)

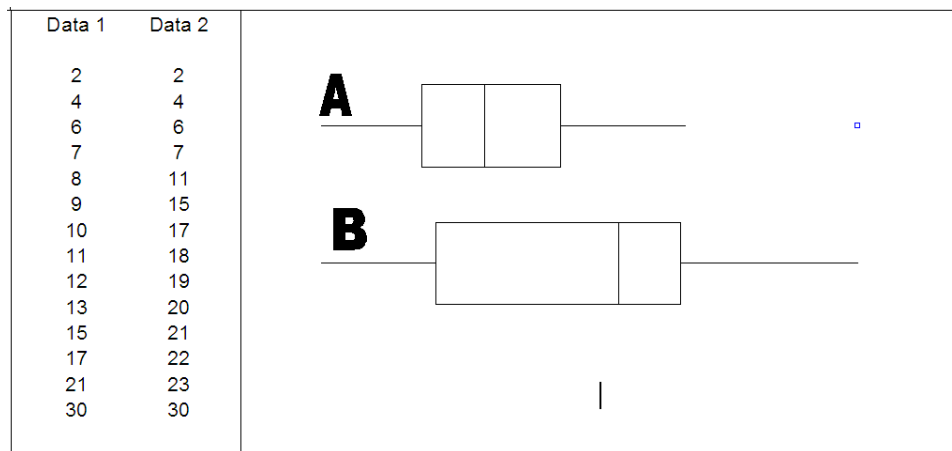


Figure 4.9: Boxplot for problem 6.

4.7. Consider the data on baseball players in 1992 that is in data file C04 Baseball1. We are interested in whether the salaries of players eligible to be free agents in 1992 were significantly different from the salaries of players who were not eligible to be free agents. Use software to make side-by-side boxplots of the stacked salary data, using the variable "Free Agent Eligibility" as the code variable. Use your boxplot to explain whether being eligible to be a free agent has an effect on your salary. In your explanation, make use of all the statistics that the boxplots show you.

4.8. Consider the data and work you did in Mechanics and Techniques, problem 1.

1. Will these results occur for any fixed (flat) salary increase? What about a salary decrease? Explain your reasoning.
2. Will these results occur for any percentage increase in salary? What about a percentage decrease in salary? Explain your reasoning.

4.9. Download the data file **C04 Sales**. The data shows sales figures from sixteen stores in our chain. We have plans to open new stores in the following cities: Honolulu, HI; Little Rock, AR; El Paso, TX; Tucson, AZ; and Hartford, CT. Generate sales figures at the five new stores that will result in the mean sales of the company exceeding the median sales of the company. To answer this problem, add five new data points to make the mean of the new data larger than the median of the new data. To demonstrate to me that your new data satisfies this criterion, copy the five data points you added, show me the box plot and summary statistics for your new set of data, and explain your thinking process for selecting these five points. In order to explain your thinking process, be sure to provide a box plot and the summary statistics for the original data in order to compare.

Organize your report as shown in figure 4.10. (NOTE: The image below is just a sample. Your graphs will probably not look like the ones below.) In addition, explain what else changed when you added the five new data points.

Keep in mind that it is unrealistic for all five of the new stores to turn in exceptional sales figures. It is more likely that the five new stores will exhibit a wide range of sales figures, with most of them falling in the inter-quartile range.

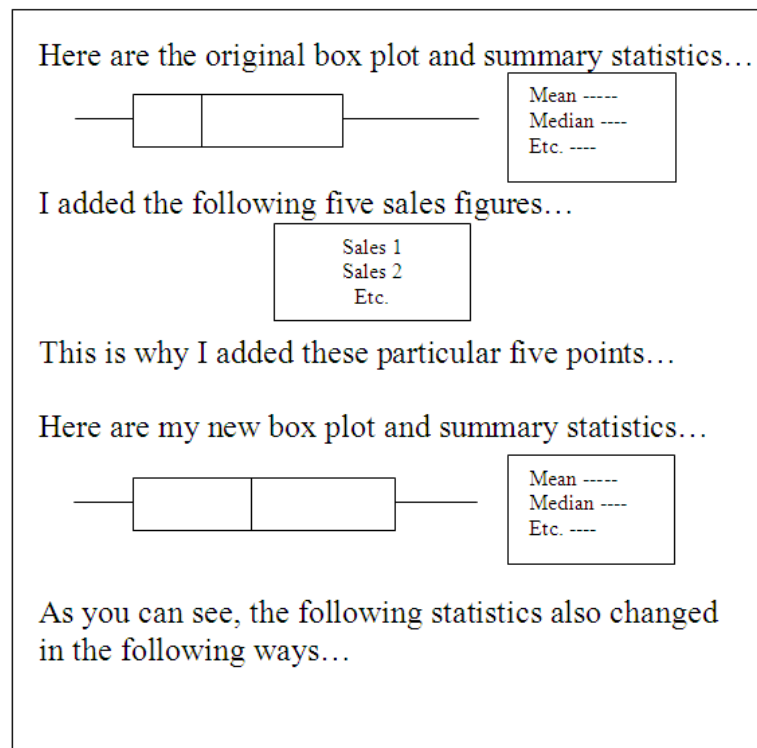


Figure 4.10: Sample report for problem 9.

CHAPTER 5

Histograms¹

This chapter focuses on looking at the spread of the data using several different tools.

- Section 5.1 introduces z-scores for each data point to abstract the notion of how spread out the data is. We can also use these to test whether the data appears to come from what is called a **normal distribution** which most randomly generated data should follow.
- Section 5.2 uses z-scores to help us build a good graphical representation of the data with a histogram. This type of graph gives a more detailed picture of the observations of a single variable and helps to classify data into one of several types. This classification then makes it easier to draw conclusions from the data.

¹©2017 Kris H. Green and W. Allen Emerson

As a result of this chapter, students will learn

- ✓ How z-scores determine a relative ranking of observations
- ✓ How z-scores allow for comparison of data that is in different units and of different sizes
- ✓ What normally distributed data is and what the "rules of thumb" are for checking it
- ✓ The difference between absolute and relative cell references
- ✓ The characteristics of each of the classic histograms: uniform, symmetric, bimodal, positively and negatively skewed
- ✓ How to check the rules of thumb using a histogram of z-scores
- ✓ The characteristics of good and bad histograms

As a result of this chapter, students will be able to

- ✓ Compute z-scores by hand or with software
- ✓ Explain why the standard deviation formula is set up the way it is
- ✓ Check the rules of thumb for the spread of data
- ✓ Read a histogram
- ✓ Interpret the information in a histogram
- ✓ Make a histogram of data either by hand or with software
- ✓ Improve on a badly made histogram in order to tease more information from it

5.1 Getting the Data to Fit a Common Ruler

Well, now we have a tool for measuring the spread of a set of data. This measuring stick, the standard deviation, is a useful tool for looking at the entire set of data. Notice that we are slowly building up more information about the data as a whole: We started with thousands of observations. Then we reduced this down to a single statistic, the mean, which measures the typical data point. Next we added a little more information by using a boxplot, which really contains seven pieces of information (minimum, first quartile, median, mean, third quartile, maximum, outliers). Then we added the standard deviation to our arsenal, giving us quite a bit of information about the typical observations and how the rest of the data is spread out.

However, these tools really only help us look at a single variable in a set of data. The measuring stick for determining the spread of the data is different for every set of data. In essence, the standard deviation is a “ruler made of rubber”; it stretches to measure the spread of data that has a large range, and it contracts to measure the spread of data with a small range. What if we want to compare two sets of data? Better yet, what if we want to compare individual observations from two different variables in our data? How can we do this when all our tools are designed to change to fit the data? Is there no standard?

As a matter of fact, there is a standard ruler that applies to all data, regardless of its size or units. Each observation in a set of data can be converted to what is called a standard score, also known as a z-score. This converts all data to a dimensionless number on a common ruler. Once this is done, we can compare z-scores for observations from different variables and we can determine which observation is farther away from the mean in an absolute sense.

It is important to realize that the concept of a z-score is fundamentally different from the other statistics we have discussed so far. The mean, the standard deviation, and the statistics shown on a boxplot are descriptive statistics for an entire set of data. On the other hand, each observation has its own z-score; thus, z-scores are more individual. At the same time, a z-score is a comparative number. Z-scores show you how a particular observation compares to the entire data set. Essentially, a z-score is a number that tells you how many standard deviations (or fractions of a standard deviation) an observation is from the mean of the data.

5.1.1 Definitions and Formulas

Z-scores or Standard Scores The z-score for an observation is a dimensionless quantity that tells how many standard deviations the observation is from the mean. To compute the z-score for observation i of a variable x , we calculate:

$$z_i = \frac{x_i - \bar{x}}{S_x}$$

Z-scores indicate the signed distance (in standard deviations) between an observation and the mean. For example, a z-score of 0 indicates that the observation is equal to the mean, while a z-score of -1.5 indicates an observation between one and two standard deviations below the mean (because of the negative sign).

Normally Distributed Data Statistically speaking, characteristics of a population (such as height, weight, or salary) are what are called normally distributed data. This is data that is symmetrically spread around the mean according to the normal distribution. The normal distribution itself is a product of a complicated-looking formula, but the basic idea is that the data should satisfy certain rules of thumb (see below).

Rules of Thumb In normally distributed data, there are approximately 68% of the observations within 1 standard deviation of the mean, 95% of the observations within two standard deviations, and 99.7% within 3 standard deviations of the mean. Thus, most of the data is fairly close to the mean, with equal amounts being above and below. In terms of z-scores, the rules of thumb would say that

Z Scores	Percentage of Observations in that Range
-3 to -2	2.35%
-2 to -1	13.5%
-1 to 0	34%
0 to 1	34%
1 to 2	13.5%
2 to 3	2.35%
Total	99.7%

Thus, very few observations (0.3%) should have z-scores larger than 3 or less than -3 if the data is normally distributed. Keep in mind however, that unless you have a lot of data (several hundred observations) the rules of thumb may not be helpful for determining whether the data came from a normal distribution.

5.1.2 Worked Examples

Example 5.1. Converting Observations to Z-scores

Let's return to the data from example 2. Remember that the mean is 6 and the standard deviation (example 3) is approximately 1.852. To calculate the z-scores, we take the observation, subtract the mean, and divide the result by the standard deviation. A few of these are done for you. Fill in the rest of the table on your own.

$$z_1 = \frac{x_1 - \bar{x}}{S_x} = \frac{3 - 6}{1.852} = \frac{-3}{1.852} \approx -1.620 \quad z_{11} = \frac{x_{11} - \bar{x}}{S_x} = \frac{7 - 6}{1.852} = \frac{1}{1.852} \approx 0.540$$

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	3	3	4	5	5	5	6	6	6	7	7	8	8	8	9
$x_i - \bar{x}$	-3	-3	-2	-1	-1	-1	0	0	0	1	1	2	2	2	3
z_i	-1.6										0.5				

From this, we see that the smallest observations in the data (x_1 and x_2) are between 1 and 2 standard deviations below the mean, since the z-scores for these observations are between -1 and -2. What do you predict will happen when you add all the z-scores up? Why? Is this generally true, or special for this set of data?

Is this data from a normal distribution? That's harder to answer. We have so few data points, that we will have a hard time matching the rules of thumb exactly. If you've calculated all the z-scores correctly then you should get 8 out of 15 observations with z-scores from -1 to +1. This accounts for $8/15 = 53.33\%$ of the data, which is a little short of expectation. If this were a normal distribution, we would expect closer to 68% of $15 = 0.68 * 15 = 10.2$ observations in this range. We have $15/15 = 100\%$ of the data with z-scores between -2 and +2, which is slightly higher than the 95% expectation from the rules of thumb, but 95% of $15 = 0.95 * 15 = 14.25$, so we are close to the right number of observations. Overall, this data does not appear to be from a normal distribution. However, to really see whether it is from a normal distribution, we need to apply some other mathematical tools. It's entirely possible that this data really is normally distributed. We simply can't tell with the tools we currently have, but we can make a guess that the data is not normally distributed: Specifically, there are not enough observations near the mean to make it normal.

Example 5.2. Converting from Z-scores back to the Data

Notice that z-scores, regardless of the units of the original data, have no units themselves. This is because units cancel out. Suppose we have data measured in dollars. Thus, the units for x_i , \bar{x} , the deviations from the mean, and σ_x are all in dollars. When we compute the z-scores, the units vanish:

$$z_i = \frac{\text{deviation}}{\text{standard deviation}} = \frac{\text{dollars}}{\text{dollars}} = \text{no units}$$

This is the power of the z-scores: they are dimensionless, so they can be used to compare data with completely different units and sizes. Everything is placed onto a standard measuring stick that is "one standard deviation long" no matter how big the standard deviation is for a given set of data.

But suppose all we know is that a particular observation has a z-score of 1.2. What is the original data value? We know it is 1.2 standard deviations above the mean, so if we take the mean and add 1.2 standard deviations, we'll have the original observation. So, in order to answer this question about a specific observation we need to know two things about the entire set of data: the mean and the standard deviation. So, if the data represents scores on a test in one class, and the class earned a mean score of 55 points with a standard deviation of 8 points, then a student with a standard score of 1.2 would have a real score of 55 points + $1.2(8 \text{ points}) = 55 \text{ points} + 12 \text{ points} = 67 \text{ points}$. If a student in another class had a z-score of 1.5, then we know that the second student did better *compared to his/her class* than the first student did, because the z-score is higher. Even if the second student is in a class with a lower mean and lower standard deviation, the second student performed better relative to his/her classmates than the first student did.

On standardized tests, your score is usually reported as a type of z-score, rather than a raw score. All you really know is where your score sits relative to the mean and spread of the entire set of test-takers.

Example 5.3. Checking Rules of Thumb for the Sales Data

Does our sales data from the Cool Toys for Tots chain follow the rules of thumb for normally distributed data? (See the new version of the data in C05 Tots.) We can ask this question

from two different perspectives: by regional comparison and by comparison across the entire chain. Let's just look at the chain as a whole and include both the northeast and north central regions. First, we'll compute the z-scores for the stores in the chain in order to convert all the data to a common ruler. For this, we'll need the standard deviation and mean of the chain, rather than the individual means and standard deviations for each region. The final result is shown in the table below:

Store	Region	Sales	Sales Z
1	NC	\$668,694.31	2.0100
2	NC	\$515,539.13	1.1483
3	NC	\$313,879.39	0.0138
4	NC	\$345,156.13	0.1898
5	NC	\$245,182.96	-0.3727
6	NC	\$273,000.00	-0.2162
7	NC	\$135,000.00	-0.9926
8	NC	\$222,973.44	-0.4977
9	NC	\$161,632.85	-0.8428
10	NC	\$373,742.75	0.3506
11	NC	\$171,235.07	-0.7887
12	NC	\$215,000.00	-0.5425
13	NC	\$276,659.53	-0.1956
14	NC	\$302,689.11	-0.0492
15	NC	\$244,067.77	-0.3790
16	NC	\$193,000.00	-0.6663
17	NC	\$141,903.82	-0.9538
18	NC	\$393,047.98	0.4592
19	NC	\$507,595.76	1.1036
20	NE	\$95,643.20	-1.2140
21	NE	\$80,000.00	-1.3020
22	NE	\$543,779.27	1.3072
23	NE	\$499,883.07	1.0603
24	NE	\$173,461.46	-0.7762
25	NE	\$581,738.16	1.5208
26	NE	\$189,368.10	-0.6867
27	NE	\$485,344.87	0.9785
28	NE	\$122,256.49	-1.0643
29	NE	\$370,026.87	0.3297
30	NE	\$140,251.25	-0.9630
31	NE	\$314,737.79	0.0186
32	NE	\$134,896.35	-0.9932
33	NE	\$438,995.30	0.7177
34	NE	\$211,211.90	-0.5638
35	NE	\$818,405.93	2.8523

To check the rules of thumb, we need to determine how many stores fall into each of the breakdowns by using the z-scores.

- Between 0 and 1 standard deviation:

There are 25 stores out of 35. This is $25/35 = 0.7143 = 71.43\%$. This value is a little higher than the rule of thumb suggests, but not by much.

- Between 0 and 2 standard deviations:

There are 33 stores in this group, giving $33/35 = 0.9429 = 94.29\%$. This is very close to the rule of thumb.

- Between 0 and 3 standard deviations:

There are 35 stores, giving $35/35 = 100\%$ of the stores in this range. This is slightly higher than the rule of thumb suggests.

Overall, this data is fairly close to satisfying the rules of thumb for being normally distributed. There seems to be one too many observations within one standard deviation of the mean, but that is generally acceptable. (Note: $71.43\% - 68\% = 0.0343\%$ and 0.0343% of $35 = 1.2$.) With only 35 data points, these results are very close to what one would expect from normally distributed data. Is the data symmetric? If it is, there should be roughly the same number of stores above the mean as there are below the mean.

5.1.3 Exploration 5A: Cool Toys for Tots

Now we have all the tools we need to look at the various stores in the two sales regions of Cool Toys for Tots (examples 5 and 3) in complete detail. Which individual stores in our chain are performing the best, relative to their regions? Which stores are performing worst in their regions? Which are performing best and worst overall? To analyze these questions, try computing z-scores for each store in two different ways: relative to each region and relative to the entire chain of stores. You can also analyze the data with and without any outliers.

The data file `C05 Tots` contains a column of identifiers (store number), a categorical variable identifying the region of each store, and the sales figures for the stores (numerical variable).

Here are some questions to guide you in this exploration:

1. How can you compute the z-scores for each store?
2. How do you compute z-scores for each store relative to only the stores in its region? (Try looking at the How to Guide for information on "How to Stack and Unstack Data".)
3. How do you identify an outlier? Is it an outlier for its region or for the entire chain of stores?
4. Are there any stores whose relative performance in their region is not reflected when it is compared to the entire chain or vice versa?
5. Are these stores performing poorly with respect to (1) the entire chain or (2) their individual regions?

5.2 Profiling Your Data

One of the most important questions a detective asks about those involved in a criminal investigation is “What did the suspect look like?” Without a physical description, detectives will have difficulty finding the suspect. Likewise, they ask about the suspect’s habits and personality. Eventually, they build a profile of the suspect. Such profiles describe the suspect physically and psychologically. They are based on statistical analyses of criminals and are extremely helpful in locating the suspect before more crimes can be committed. In order for you to study data from a business setting, you will also need to develop a profile of the data. We have begun this in chapter three with a discussion of central tendency. In chapter four we described the way the data is spread out using various tools. Along the way we got a blurry picture of the data, the boxplot. Now it’s time to sharpen the picture and get more detail. The best tool for this is called a histogram. It will help answer the question “What does your data look like?”

A histogram is basically a graph that breaks the observations of a single variable into intervals called bins. By counting the number of observations in each bin we can generate a frequency table of the data which is then turned into a type of bar chart, with one bar for each bin and the height of each bar indicating the number of observations it contains. Usually histograms have eight to twelve bins. This means that we get a more detailed picture of the data than from a boxplot. With each step, we get more information about the data to help us make decisions.

Representation	What it is	What it tells you
Raw Data	Many observations, lots of information	Hard to make sense out of
⇓		
Averages	Single number (mean, median, or mode)	Tells what is “typical”
⇓		
Boxplot	Seven pieces of information (min, Q1, median, mean, Q3, max, outliers)	Shows where the data is bunched together
⇓		
Histogram	Ten to fourteen pieces of information usually (min, bin width, frequencies)	More detailed profile of the data

Most histograms can be classified into one of five types: uniform, symmetric, bimodal, positively skewed, or negatively skewed. Each type has certain characteristics that make it easy to recognize. Being able to classify the data as one of these types helps you analyze the data in much the same way that a good profile of a suspect tells the detectives a lot about how to catch him or her. In this section, you will learn to recognize each of these classic histograms and will learn what each one tells you about the data. As you learn how to make,

read, and interpret histograms, keep in mind that real data will never exactly look like any of the “perfect examples”. Many times you will be required to make a judgment call as to which type of distribution the data fits.

Another important detail about histograms to remember: depending on what bins you use to make the histogram, the data may look different. It’s a good idea to look at the data in several ways before drawing any conclusions.

5.2.1 Definitions and Formulas

Frequency Table Sometimes it may be useful to group the data together into subgroups (called bins, see below). To do this, you simply count how many observations fall into each bin. This count is called a frequency. When you have all of the observations placed into bins, the entire list of bins and frequencies is a frequency table for the data.

Bins A bin is one of the boxes in which data are placed to make a frequency table. Typically bins are all the same size or cover the same number of categories. For example: Ages of people could be divided into bins like 10-19, 20-29, 30-39, etc. You could also divide the ages into 0-19, 20-39, 40-59, etc. Each of these intervals is a bin into which observations are placed. Think of this as making a bunch of boxes, each labeled with a range of values. If an observation falls inside that range, place a counter into the box. When you have finished doing this for all the observations, you will have a frequency count for the data.

Distribution In the sense that we are referring to it in this text, distribution refers to the way the data is spread out or bunched together.

Histogram A histogram is a graphical representation of a frequency table. It shows the bins along the horizontal axis and has bars above each bin. The height of each bar represents the number of observations that fall in that bin. Histograms can be made directly from data using most software packages, or by first creating a frequency table and generating a bar graph.

Skewness Skewness measures how far the distribution of data is from being symmetric. The actual formula for skewness uses the z-scores of the data and is a little ugly:

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n z_i^3$$

compares the data to the mean. If most of the data is less than the mean, then the skewness will be negative. If most of the data is greater than the mean, then the skewness is positive. The reason for this behavior is the exponent of three: data points far from the mean (and thus having a large deviation and a large z-score) will affect the total more than points close to the mean. In a positively skewed data set, the smallest values are much closer to the mean than the largest values, so the large positive deviations are made even larger by cubing them. The opposite happens for negatively skewed data.

Uniform Distribution A uniform distribution (figure 5.1) has roughly the same number of observations in each bin. It looks almost flat, with each bin having almost the same height:

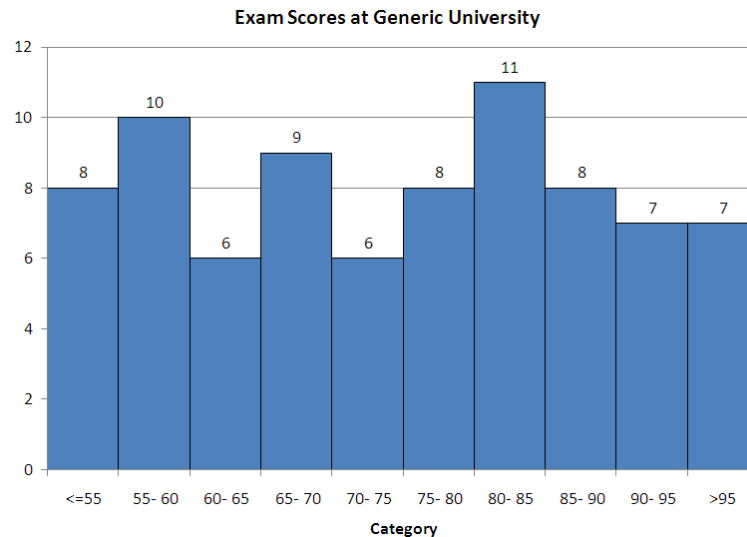


Figure 5.1: A histogram of uniform data.

Symmetric Distribution A symmetric distribution (figure 5.2) has equal amounts of data on each side of a central bin. As you move farther from the central bin in either direction, the same number of observations (approximately) can be found.

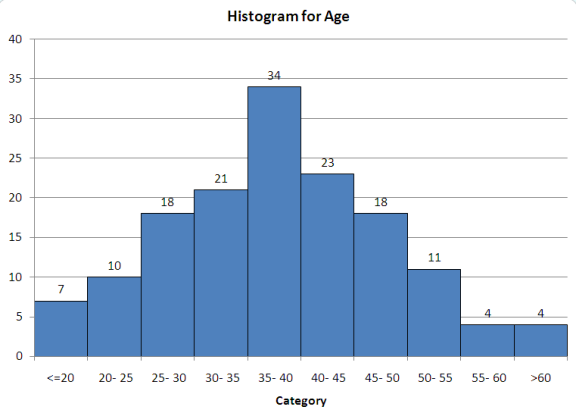
Positively Skewed Distribution A positively skewed distribution (figure 5.2) has more data on the left side of the mean. Typically, the skewness of such distributions is positive, and the median is less than the mean. The "tail" of the distribution points toward increasing values on the horizontal axis.

Negatively Skewed Distribution A negatively skewed distribution (figure 5.2) has more data on the right side of the mean. Typically, the skewness of such distributions is negative, and the median is more than the mean. The "tail" of the distribution points toward decreasing values on the horizontal axis.

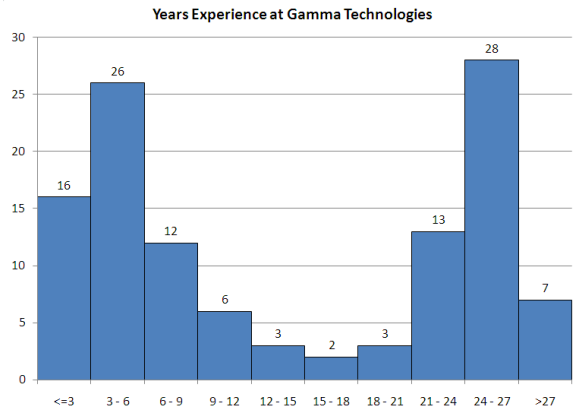
Bimodal Distribution A bimodal distribution (figure 5.2) has two major peaks in it (there are two modes to the data, hence the term bi-modal). There is usually a gap between the peaks with fewer observations. It is also possible for there to be more than two peaks in the data, leading to trimodal or multimodal distributions.

5.2.2 Worked Examples

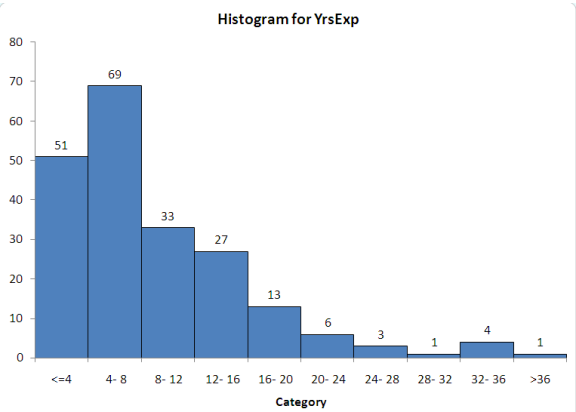
Example 5.4. Reading a Histogram



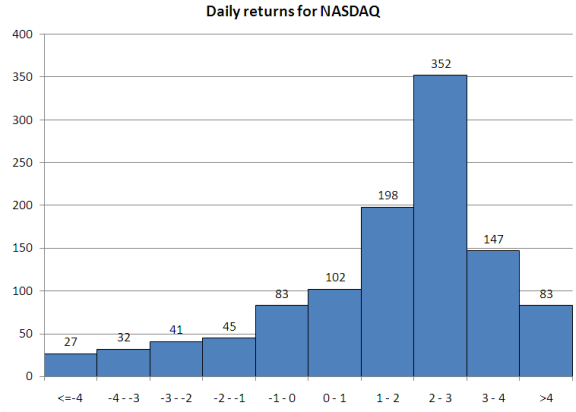
A histogram of symmetric data.



A histogram of bimodal data.



A histogram of positively, or right-skewed, data.



A histogram of negatively, or left-skewed, data.

Figure 5.2: Illustrations of the major types of distributions of data.

Look at the data shown in the histogram below. What can we learn about the data? First of all, notice the bins; they are labeled below the bars on the graph. Each bin is the same width as the others, with the exception of the two ends. These are open-ended so they can catch all the observations that are outside the main bulk of the data. When you make a histogram in many software packages, you control three things about the format of the graph: the minimum value (which is really the upper end of the first bin), the number of bins (which includes the two ends!) and the width of each bin. Consider the histogram shown in figure 5.3, which is an example of a positively skewed histogram.

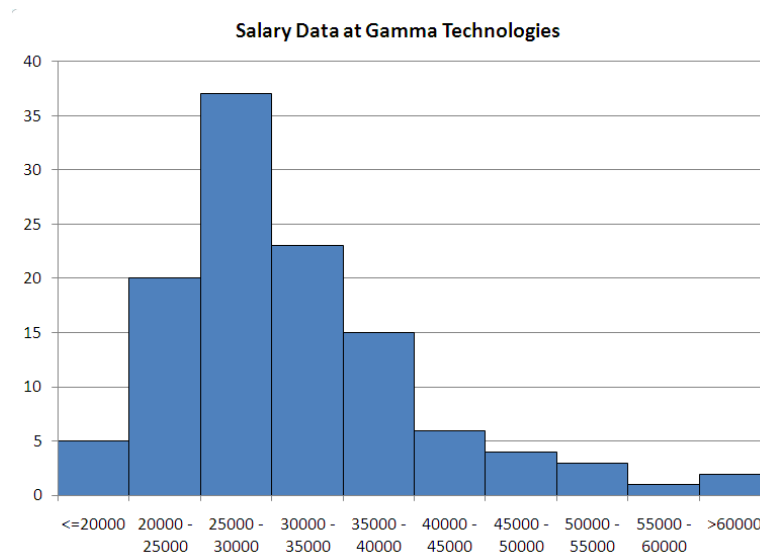


Figure 5.3: Salary distribution at Gamma Technologies, showing a distinct positive skewness.

It seems that whoever created this graph used the following settings to make the graph:

Minimum	20,000
Number of Categories	10
Category Width	5,000

Notice that even though the minimum is set at 20,000, the first bin contains all the observations less than this number. What else can we see? First off, we notice that there are a total of 116 observations in this data. The bin with the most observations is the 25,000 to 30,000 bin, containing $37/116 = 31.90\%$ of the observations. Most of the observations fall in the bins between 20,000 and 35,000, with a total of $20 + 37 + 23 = 80$ or about 68.97%. Because the data has such a long tail in the direction of increasing salaries, this graph is said to be positively skewed. This means that the mean is larger than (to the right of) the bulk of the data. Think of it like a teeter-totter. Try to find the point along the axis where the histogram would balance. Start in the 25,000 -30,000 bin. The bins on either side of it are roughly equal, so don't move the balance point. Now the next bins out ($\leq 20,000$ and 35,000 - 40,000) are very unequal, with more of the weight on the right. This pulls the mean to the right of our starting point. All of the other bins are even more unbalanced, so

the mean is pulled far to the right of the highest peak. All of this tells us that the measures of "typical" for this data are a little skewed. If we report the mean, which is approximately \$34,000, then we are ignoring the skewness of the data. More than half of the data, 85 out of 116 points, is less than the mean, while only 31 points are greater than the mean. So, in what way is the mean a measure of typical for this data? On the other hand, the median is slightly less than \$30,000, with exactly half of the data above and below it. In the next section, chapter 6B, we'll look more closely at how to read off statistics from a histogram. As it turns out, almost all of the observations in the last three bins are outliers.

Example 5.5. Using histograms to check rules of thumb

By making a histogram of z-scores, we can check to see if the data is normally distributed. First, compute the mean and standard deviation of the data. Then create a column of z-scores for the data. Now, when you make the histogram, we make it with the following options :

Minimum	-3
Number of Categories	8
Category Width	1

This will ensure that you have the following bins for the z-scores:

First Bin:	$\leq (-3)$
	-3 to -2
	-2 to -1
	-1 to 0
	0 to 1
	1 to 2
	2 to 3
Last bin:	> 3 .

The graph in figure 5.4 shows such a histogram for data taken from the stock market (daily returns of a particular stock for about two years).

Notice that since the mean has a z-score of zero, the mean of the data will always fall in the middle of a z-score histogram, between the fourth and fifth bins. Each bin is one standard deviation wide, so we can now compare the frequency counts of the data to the expected frequency counts from a normal distribution (see chapter 3B).

First, is the distribution symmetric? This graph is pretty close to being symmetric. The two central bins are close in height, as are the bins on either side of the central peak. The bin marking 2 to 3 standard deviations below the mean and the bin marking 2 to 3 standard deviations above the mean are about equal in height. The only parts that don't match are the ends. There should be very little data in these two bins anyway (less than 0.3%), and both are close to zero.

Second, does the rule of thumb hold for this data? Let's check it out.

- Within one standard deviation of the mean, there are about $194 + 217 = 411$ observations of the variable return. This is $419/552 = 74.46\%$ of the data. This is a little high when compared to the 68% rule of thumb.

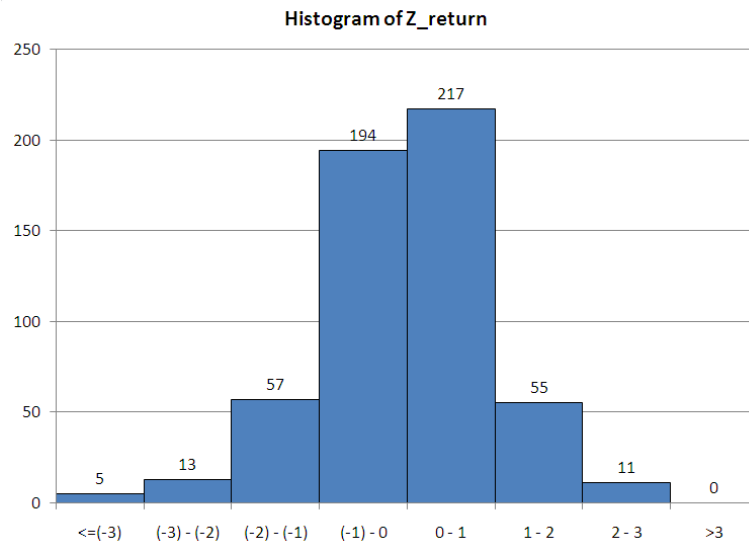


Figure 5.4: Histogram of z-scores for daily returns of a particular stock over a two-year period.

- Within two standard deviations of the mean, there are about $57 + 194 + 217 + 55 = 523$ observations of the variable return. This accounts for $523/552 = 94.75\%$ of the data. This is fairly close to the 95% rule of thumb.
- Within three standard deviations of the mean, there are about $13 + 57 + 194 + 217 + 55 + 11 = 547$ observations of return. This accounts for $547/552 = 99.09\%$ of the data. This is very close to the 99.7% rule of thumb.

Overall, this data is close to the rule of thumb, but seems to have too much data within one standard deviation of the mean, so we would probably conclude that this data is not from a normal distribution.

Now the question you've been waiting for: Why do we bother checking if the data is from a normal distribution? The main reason is related to the statement we made about normal distributions in chapter 3: "Statistically speaking, characteristics of a population (such as height, weight, or salary) are what are called normally distributed data." Suppose that we conducted a customer satisfaction survey. Part of the survey would likely contain demographic data on the ages, incomes, and so forth of the customers. This is to ensure that the conclusions we draw about their satisfaction are accurate conclusions. We would expect that, if we properly sampled the customers, the demographic data would be normally distributed around some mean with some standard deviation. If this is not the case, then we are getting too much satisfaction data from some group or groups. This could completely invalidate our conclusions. A famous case of this type of mistake occurred with *Literary Digest* in 1936. The magazine surveyed its viewers about the upcoming presidential election. The survey overwhelmingly favored the Republican candidate, who lost by a landslide in the election. The magazine failed to consider that their audience was not a good sample of the entire U.S. population: their readers were mostly high-income families in an era of

economic depression where most families could not afford the money to subscribe to a literary magazine.

Example 5.6. The Good, the Bad, and the Ugly

There are four ways for a histogram to break bad. By this, we mean that the histogram does not tell you everything it could tell you about the data. Each of these four cases is described below. When making your histograms, if you see graphs like those below, you should try to correct the problem. There are three numbers you can manipulate when building histograms: the minimum value, the number of bins/categories, and the width of each bin (sometimes called the category length). To fix the histogram, simply change one or more these values to adjust the shape. Try several combinations out before you settle on a particular graph. Use each graph you make to help choose better values for the next graph.

The five graphs in figure 5.5 illustrate histograms that are not helpful for different reasons. The data for each graph is the same, and it represents the amount of household debt accumulated by various households in a recent survey. Notice how each graph makes the data look very different, possibly leading to misunderstandings about the underlying data.

Case 1. Too much data in the end bins

This is a classic problem, typically caused by miscalculating where the two ends of the distribution are. The reason it is such a problem is that the end bins are usually "open-ended". This means that they do not cover a specific interval. Instead, they are usually labeled with " \leq some number" for the left end and " $>$ some number" for the right end. If too much data is in either of these bins, then you cannot really describe the distribution, because you do not know how far the data extends.

Case 2. Bins are too wide (lumpy)

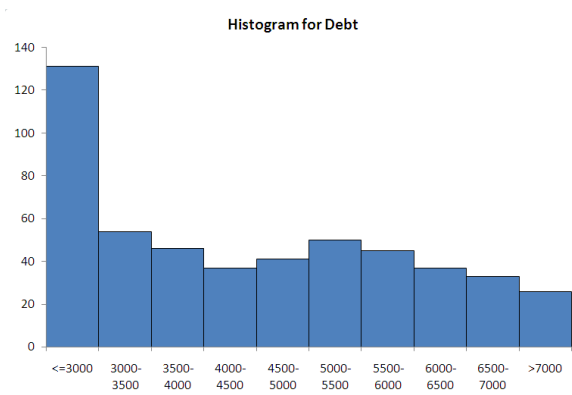
This problem is usually caused by making the bins too wide. This means that each bin covers a large range of observations and means that you can fit fewer bins into the range of the data. To see how many bins of a particular size will fit into a range of data, take the range of the data (maximum value minus the minimum value) and divide by the width of the bins. For example, if the data goes from 0 to 100, and you make each bin 25 units wide, then you will have all of the data in $(100-0)/25 = 100/25 = 4$ bins! This will not provide you with much information.

Case 3. Too few observations in each bin (compared to total number of observations)

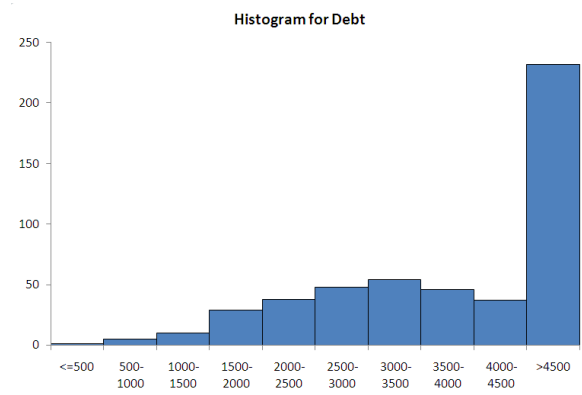
This problem is the opposite of case 2. This is caused by having so many bins that only one or two observations fall into each range. If you have 100 observations that range from 0 to 100, then it does not make sense to use 100 bins that are 1 unit wide. This isn't any better than the original data, since each bin will have, on average, only one observation. It's usually best to stick to eight to twelve bins for a histogram, unless there are many observations and the data needs more bins in order to see the detail. Typically, this is only needed with bimodal data.

Case 4. There are empty bins on the ends

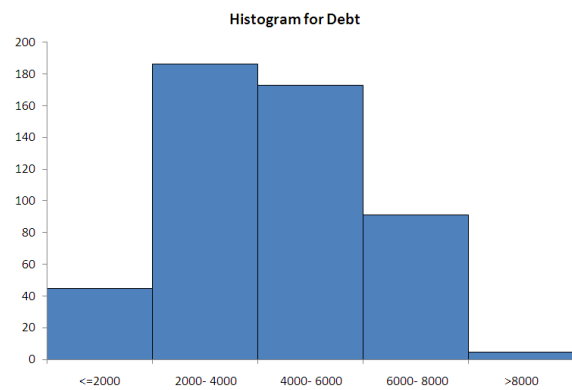
This problem is also caused by miscalculating the locations of the ends of the data. While it is not really a problem, this situation does lead to wasted space in your graph. In addition, the empty bins could be used to mislead someone about the true distribution of the data.



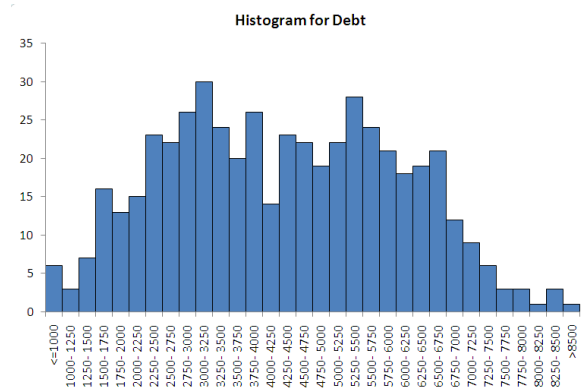
Case 1a: Too much data in first bin.



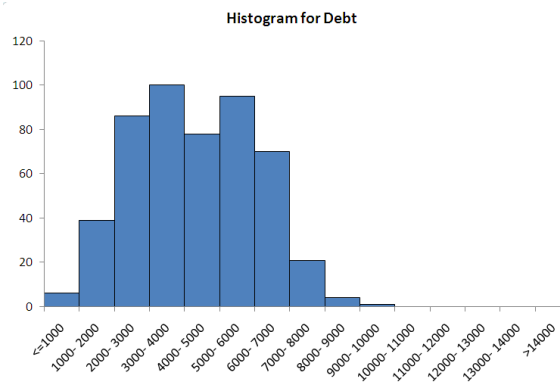
Case 1b: Too much data in last bin.



Case 2: Too lumpy.



Case 3: Too spread out.



Case 4: Too much wasted space

Figure 5.5: Illustrations of the major mistakes in displaying data with a histogram.

5.2.3 Exploration 5B: Beef n' Buns Service Times

The manager of a local fast-food restaurant is interested in improving the service provided to customers who use the restaurant's drive-up window. As a first step in this process, the manager asks his assistant to record the time (in minutes) it takes to serve 200 different customers at the final window in the facility's drive-up system. The given 200 customer service times are all observed during the busiest hour of the day for this fast-food operation. The data are in the file C05 BeefNBuns. Are shorter or longer service times more likely in this case?

STUDENT analysis: A student produces the graph shown and then states: "As the graph below shows, most of the service times are on the higher end of the graph, so we expect that there will be many customer complaints."

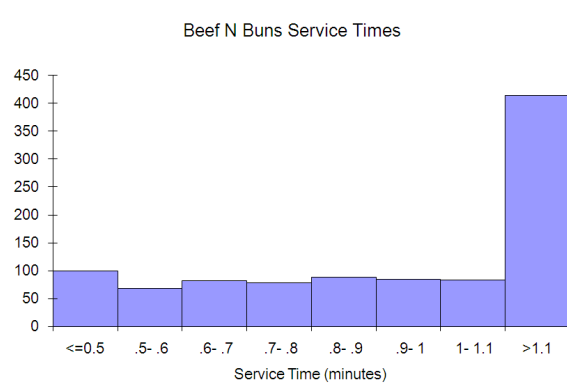


Figure 5.6: Sample histogram of service times at Beef n' Buns.

1. **OBSERVATIONS:** What does this graph tell you about the situation? How many service times were long? How many were short?
2. **INFERENCES:** What do we mean by "long service times" or "short service times"? What is wrong with the graph above? What doesn't it show?
3. **QUESTIONS:** What information that you need is left unsaid by the graph? What questions about the data do you have that a more accurate representation of the data might help you answer?
4. **HYPOTHESIS:** What will happen to the graph if we change X ?
5. **CONCLUSION:** Make an accurate sketch of what the new graph will look like.

5.3 Homework

Mechanics and Techniques Problems

5.1. The data file `C05 Homes` contains data on 275 homes that sold recently in the Rochester, NY area. Included in the data are observations of the location of the home, the annual taxes, the style of the home, the number of bedrooms, the number of bathrooms, the total number of rooms, the number of cars the garage will hold, the year in which the house was built, the lot size, the size of the home, the appraised value of the home, and the sale price.

1. Which of these variables are categorical? Which are numerical? For numerical variables, give the units and the range, for categorical variables describe the categories.
2. Make histogram of the appraised values and the selling price of the homes. Compare these distributions. In what way(s) are they similar? In what ways are they different? What does this tell you about the housing market in the greater Rochester area?

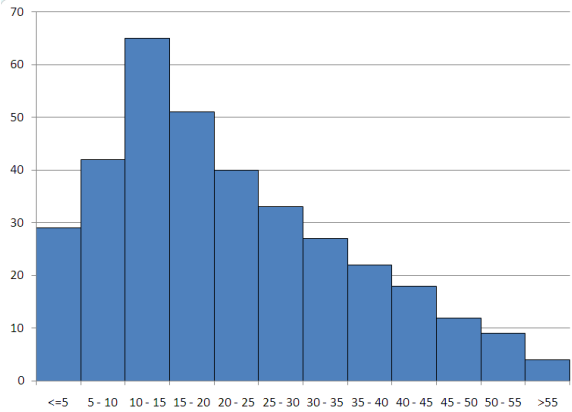
5.2. Consider the four histograms in figure 5.7 (labeled A - D). For each histogram, describe the shape.

5.3. Suppose you know that the mean of a set of data is 207.8 and the standard deviation is 43.2. If the median has a Z-score of -0.85, what is the median of this data?

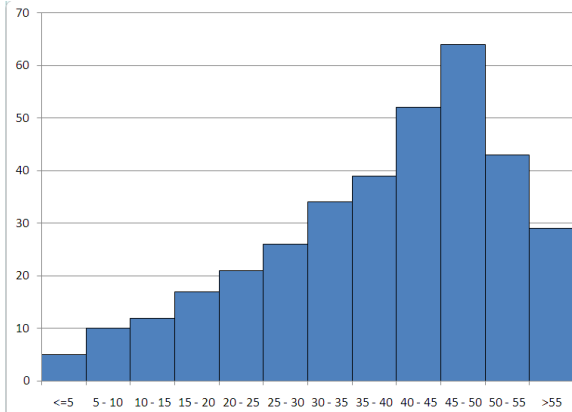
5.4. Figure 5.8 shows two histograms. Histogram A represents the ages of people attending a recent event; histogram B represents the salary ranges of employees at a company. Each of these histograms could be improved in order to provide a better picture of the underlying data. For each, explain why the given histogram is less-than-ideal, and explain what you would do to improve it. In other words, how would you go about making a better histogram?

Application and Reasoning Problems

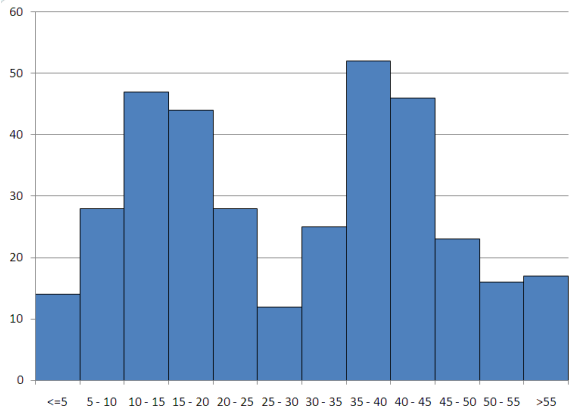
5.5. The histograms in figure 5.9 show flight arrival data for two airlines your company is considering signing a permanent contract with. Both airlines offer similar flight times, similar service, and travel to all of the same cities. For 1,000 flights on each airline last year, the data is a record of how far before or after the scheduled arrival time that the flights actually arrived. Negative times indicate that the flight landed ahead of schedule. Positive times indicate late-arriving flights. For each of these graphs, construct both a cumulative distribution graph and a boxplot. Based on these data and your graphs which airline would you be willing to pay more for? Explain your reasons.



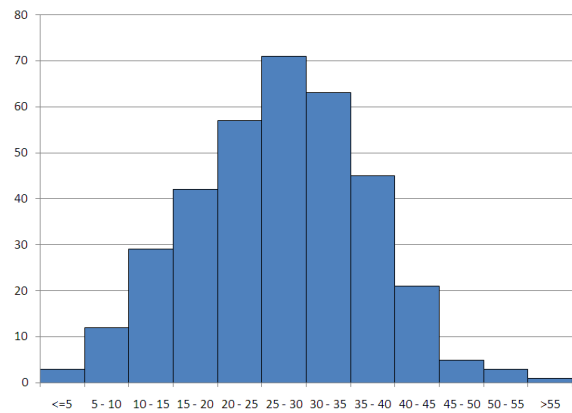
Histogram A.



Histogram B.



Histogram C.



Histogram D.

Figure 5.7: Histograms for Mechanics and Techniques problem 2.

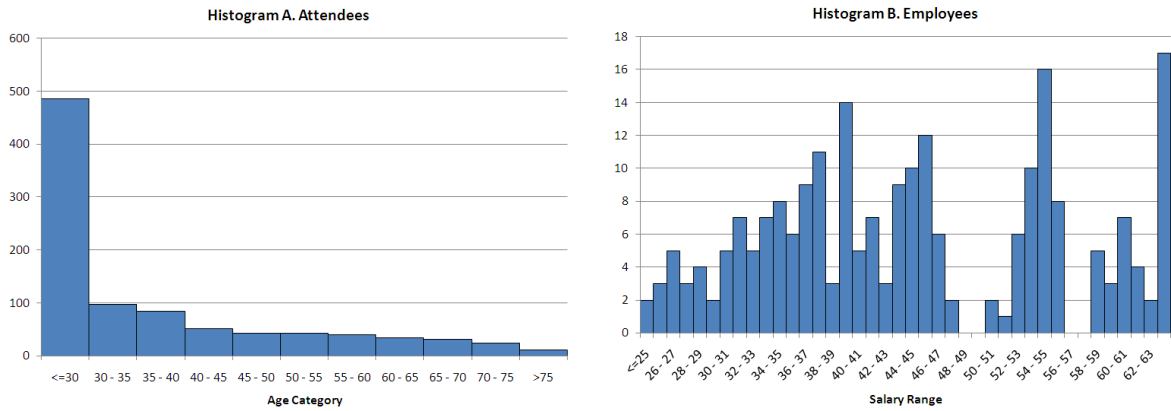


Figure 5.8: Histograms for Mechanics and Techniques problem 4.

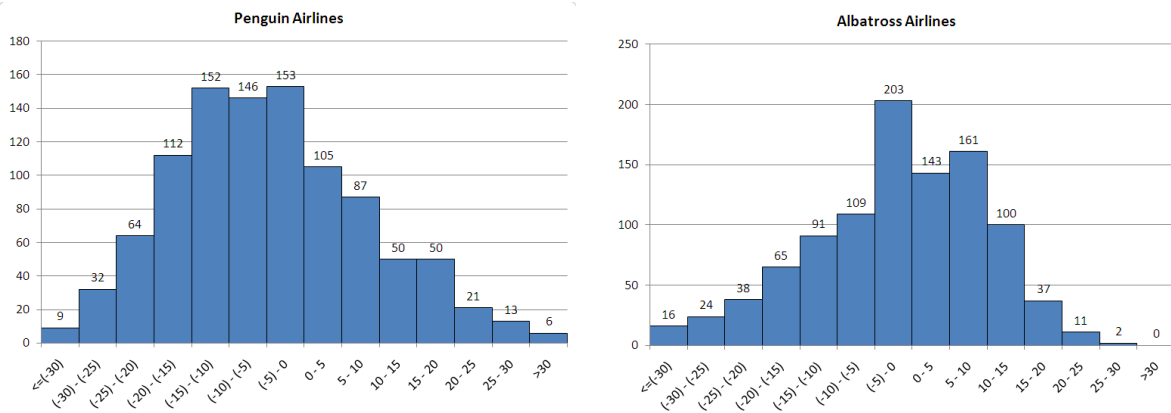


Figure 5.9: Histograms for Application and Reasoning problem 5.

5.6. Over Easy, a breakfast restaurant, has collected data on the number of customers the restaurant serves each day of the week for a full year. Assume these data are normally distributed with a mean of 135 and a standard deviation of 22. They want to use this information to determine staffing for the restaurant.

- 68% of the time, the total number of customers will fall between what two numbers (approximately)? What about 95% of the time? What about 99.7% of the time?
- Suppose a normal server can handle about 35 people in one day (at this restaurant). For each of the situations in part (a), how many servers are needed?
- How many servers would you staff? Explain your answer.

5.7. You are the manager of a local grocery store and would like to collect data on the checkout times for customers in order to help inform staffing and the arrangement of your

various express and standard checkout lines. You expect about 1000 customers each day and plan to measure the checkout time as being the total time elapsed from when the cashier starts scanning the first item to when the customer leaves with his/her groceries. Sketch a reasonable histogram of the checkout times, paying particular attention to the shape of the distribution and the units on the axes. Make sure that each axis is labeled and that each bin of the histogram is labeled with reasonable values for this situation. When you have a reasonable graph, write a short (2-3 sentence) paragraph explaining why you expect the checkout times to have the distribution you have drawn (left/right skewed, uniform, symmetric, bimodal).

5.8. Consider the salary distribution for a company named OutRun shown in figure 5.10.

1. Write a brief description (2-3 sentences) explaining what you infer about the company from looking at this distribution of salaries.
2. Now, sketch a new histogram of these same data by making the bins twice as wide. In other words, you have exactly the same data and distribution, we are just going to draw the histogram with half as many bins (each cover twice the range of salaries).
3. What does your new histogram (part b) seem to suggest about the salary distribution at the company? Write 2-3 sentences.
4. Based on your two graphs and analysis, how useful is the mean salary in representing the company? What other measures might be more appropriate?

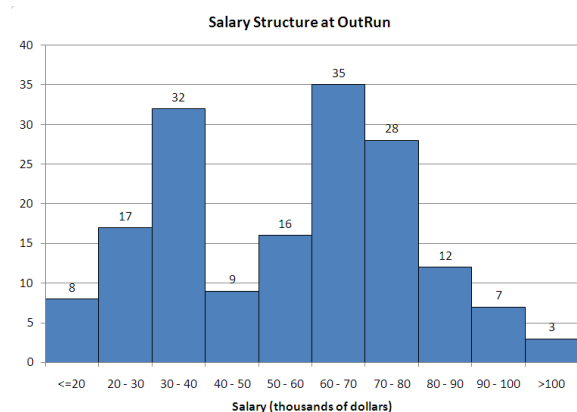


Figure 5.10: Histograms for Application and Reasoning problem 8.

5.9. Consider the histogram below, made by one of your co-workers to represent the age distribution of customers at your company. This person claims that the graph supports the idea that most of their customers are neither young nor old and do not have children

at home, which will help you decide when and where to do your advertising. Answer the following questions.

1. What does this graph tell you about your customers? What percentage of your customers is young? What percentage is old? What do you even mean by “young” or “old” in this case?
2. Why do you think your coworker claims that most of your customers do not have children at home? Do you agree? What would help you make a more informed decision about this?
3. How would you improve the graph in order to get more information without needing more data? What might the graph look like then? Sketch three possible versions of this data if the histogram were created differently (say, with more bins or different starting points).

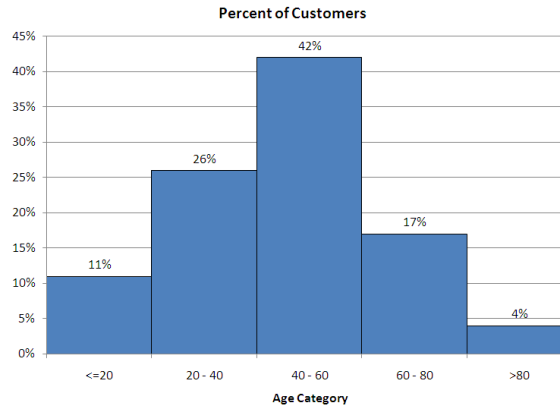


Figure 5.11: Histograms for Application and Reasoning problem 9.

Interpreting Spatial Models¹

This chapter aims to do two things.

- Section 6.1 focuses on how to estimate statistics, particularly the mean and standard deviation, from data that is only presented in summary form (like a frequency table or a histogram).
- Section 6.2 takes this one step further, by helping you connect two different ways of picturing data by relating histograms and boxplots. Both give a picture of how the data is spread out. The difference is that a boxplot takes the data and breaks it into four chunks with the same number of observations in each chunk, but with each chunk of data having a different length. Histograms are the opposite: each chunk has exactly the same length, but probably has different numbers of observations in it.

<p><i>As a result of this chapter, students will learn</i></p>	<p><i>As a result of this chapter, students will be able to</i></p>
--	---

- | | |
|---|--|
| <ul style="list-style-type: none"> ✓ Why summarized data cannot be used to compute an accurate mean or standard deviation ✓ What a percentile is ✓ What a cumulative distribution is | <ul style="list-style-type: none"> ✓ Estimate the mean from a set of summarized data ✓ Estimate the standard deviation from a set of summarized data ✓ Sketch a boxplot of the data underlying a histogram without having the data itself ✓ Sketch a rough idea of a histogram of data based only on a boxplot of the data |
|---|--|

¹©2017 Kris H. Green and W. Allen Emerson

6.1 Estimating Stats from Frequency Data

Many times we are presented with data, in newspapers, magazines, the Internet, or meetings, but these data are rarely presented in its entirety. After all, in many cases, there are thousands of observations of each variable. It is therefore more common to present summarized data in the form of tables or charts that show the number (or frequency) of observations that fall into a certain range (or bin). In the last chapter, we used this idea to create a graphical depiction of the data in the form of a histogram. But what if you are starting from the summarized data and what to know something about the original data itself?

For example, what if you wish to compute the mean of the data? This is the most frequently used measure of central tendency and is often used a model of the data. The way in which we compute this measure of central tendency is based on having all of the individual data points in the set of data. In a summarized table of data, though, we do not have the actual values to add up. One thing is certain; we cannot simply average the frequency counts, as this does nothing to account for the actual values of the data and the frequency counts are not (usually) even in the same units as the data itself. For example, in looking at the table below, we see data on salary distribution at a company. If we average the frequency counts (labeled “Number of Employees”) we get 11.8, which means that if the distribution were uniform, there would be 11.8 employees in each salary range. But this number has units of number of people. The average salary must have units of dollars. Somehow, we must estimate the mean based on both the salary ranges and the number of observations in that range.

Salary Range	Number of Employees
\$200,000 - \$250,000	1
\$150,000 - \$199,999	2
\$100,000 - \$149,999	5
\$50,000 - \$99,999	13
\$0 - \$49,999	38

Unfortunately, as we’ll discover, once you have only the summarized data, there is no way to get the actual mean of the original data. At best, you are estimating the mean, and your estimate has a great deal of possible error, depending on the size (width) of each bin into which the data has been summarized. These same ideas hold true for estimating the standard deviation of the data, especially since we must first estimate the mean in order to compute the deviations of each observation (or, in this case, each group of observations) from the mean.

And while it is true that in many cases we have the actual data and can compute the true mean of the data, this is often not true. Have you every filled in a customer satisfaction survey? Such surveys often collect demographic data, such as the age of the person filling in the form, but rarely do they ask you to write in your age. It is more common to check off a box marking a range where your age fits (for example, 31-40 years old). In situations like this, the data starts as a summarized frequency table; the company collecting the data never has the actual ages of each survey participant. So they must resort to estimating the mean if they need it for other calculations.

6.1.1 Definitions and Formulas

Summarized Data Summarized data is data not presented in raw form. Instead, the data has been grouped (or summarized) into categories. For example, rather than listing the salaries of all 250 employees at a company, a summarized presentation of this data might simply tell you the number of employees in each salary range, such as 10 employees making \$0 - \$20,000, 34 employees making \$20,001 to \$40,000 and so forth.

Weighted Average A weighted average is a type of mean where each item to be included in the average has a different weight depending on either its frequency or importance. One of the most common weighted averages is a student's GPA in college. Each class is assigned a value, based on the grade (usually a number from 0 - 4 quality points) and is assigned a weight based on the number of credit hours (3 for a three credit course, 4 for a 4 credit course, etc.) The overall GPA is then computed by weighting each grade (multiply the quality points by the weight [number of credit hours]), adding these weighted grades up, and dividing by the total number of credit hours (which is just the sum of all the weights). This means that a low grade in a high weight course (one with more credit hours) is more damaging than a low grade in a course with few credit hours. Another common use of weighted averages is to estimate the mean of a set of data given by a frequency table. In this case, the weight is determined by the frequency counts. For example, if 10% of a class scored 50 on an exam, 20% scored 60, 40% scored 70, 10% scored 80 and 20% scored 90, then the class average is

$$\frac{0.10(50) + 0.20(60) + 0.40(70) + 0.10(80) + 0.20(90)}{0.10 + 0.20 + 0.40 + 0.10 + 0.20} = \frac{71}{1} = 71.$$

More generally, if the data are given by x_i and the weights are given by w_i , the weighted average of the data is given by

$$\text{Weighted Average} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Weight Each item to be included in a weighted average is assigned a weight that identifies how much that item contributes to the overall average. The weight assigned to the i th piece of data is commonly denoted by the symbol w_i .

Estimated Mean When computing the mean of data that is given only by frequency tables, we cannot compute the actual mean since we only know the number of data points falling into each range of the frequency table. We must estimate the value for the data in each bin of the table, which results in an overall estimate of the mean. Estimated means from frequency data are computed by using the formula for the weighted average, with the weights given by the frequency counts (the number of pieces of data in the bin). The symbol \bar{x}_{est} will be used to represent the estimated mean.

Estimated Standard Deviation Estimating the standard deviation from frequency data is similar to the process of estimating the mean, but involves a few more steps. First,

of course, we need to estimate the mean of the data. Now, if each of the data points (or central points of each bin in the data, more precisely) is denoted by x_i and the frequency (number of items in that bin of the frequency table) is given by w_i , then the estimated standard deviation is

$$S = \sqrt{\frac{\sum_{i=1}^n w_i (x_i - \bar{x}_{est})^2}{(\sum_{i=1}^n w_i) - 1}}$$

Note that the sum of all the weights

$$\sum_{i=1}^n w_i = w_1 + w_2 + \dots + w_n = n = \text{Total number of data points.}$$

Sumproduct Another way to think of weighted averages is to think of lining up the data in one column and the weights in another column. By multiplying line-by-line, we find the contribution of each item to the weighted average. Adding these contributions results in calculating the top part of a weighted average. In Excel, the SUMPRODUCT function takes two lists (one is data, one is weights) and carries out this computation. This is equivalent to thinking of the items as a vector and the weights as a vector and computing the vector dot product or scalar product.

If the two lists are given by $x = (x_1, x_2, \dots, x_n)$ and $w = (w_1, w_2, \dots, w_n)$, the scalar product or SUMPRODUCT of these two lists is simply

$$\text{SUMPRODUCT}(x, w) = \sum_{i=1}^n w_i x_i = w_1 x_1 + w_2 x_2 + \dots + w_n x_n.$$

6.1.2 Worked Examples

Example 6.1. Estimating the mean

Suppose we have been presented with a table of information like that below which summarizes the salaries of the employees at a company. What is the average salary at your company (as determined by using the mean)?

Salary Range	Number of Employees
\$200,000 - \$250,000	1
\$150,000 - \$199,999	2
\$100,000 - \$149,999	5
\$50,000 - \$99,999	13
\$0 - \$49,999	38

We notice several problems immediately. First, we do not have the actual salaries of each employee; we only know a range of salaries. Second, each employee in each range could have a different salary within that range. How can we compensate for this?

The first problem is one of identifying a particular salary to represent each range. Common choices are the midpoint of the range and either endpoint. We will start with the midpoint and then repeat the analysis using the endpoints. The second problem is more interesting. Typically, we assume that all the observations in a given range of salaries are the same. Clearly, this is a poor assumption, but without it we cannot really get anywhere. This ambiguity in dealing with summarized data is why we can only claim to be estimating the mean of the data, not computing it. Based on these assumptions, we now have the following table of data to work with (after rounding the salaries off).

Salary	Number of Employees
\$225,000	1
\$175,000	2
\$125,000	6
\$75,000	13
\$25,000	38

So, is the average just $(225,000 + 175,000 + 125,000 + 75,000 + 25,000)/5 = \$125,000$? That would seem to be a bit high, wouldn't it, since only 8 employees make that much money and 51 employees are below that level. The problem with this kind of computation is that each of the salary ranges must be weighted. This means that we must put several copies of each salary into the calculation, one copy for each observation that matches that data value. Think about it this way, if we had a complete list of the salaries for computing the mean, it would look something like the table below.

225,000	75,000	75,000	25,000	25,000	25,000
175,000	75,000	75,000	25,000	25,000	25,000
175,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
125,000	75,000	25,000	25,000	25,000	25,000
75,000	75,000	25,000	25,000	25,000	

To add up all these salaries and compute the mean, we would need to include 2 copies of the \$175,000 salary, 6 copies of the \$125,000 salary and so on. Thus, we estimate the mean (measuring the data in thousands of dollars) as

$$\frac{225(1) + 175(2) + 125(6) + 75(13) + 25(38)}{1 + 2 + 6 + 13 + 38} = \frac{3,250}{60} = \$54.166.$$

Clearly this number is more reasonable for the salary. How can we estimate such means in general? First, we need to assign symbols to each quantity. Suppose we have N different groups of data. In the above example, we have 5 groups of data, so $N = 5$. Let the value of the i^{th} range be (x_i) and we will let the number of data points in the i^{th} range be given by (n_i) . With these naming conventions, the data table above would look like this:

Salary	Number of Employees	Product
x_i	n_i	$x_i n_i$
$x_1 = \$225,000$	$n_1 = 1$	225,000
$x_2 = \$175,000$	$n_2 = 2$	350,000
$x_3 = \$125,000$	$n_3 = 6$	750,000
$x_4 = \$75,000$	$n_4 = 13$	975,000
$x_5 = \$25,000$	$n_5 = 38$	950,000
Total	60	3,300,000

Now, to compute the mean, we multiply each data value (the x 's) by its weight (the n 's) and add these up. Then we divide by the total number of observations (the sum of all the n 's):

$$\bar{x}_{est} = \frac{n_1 x_1 + n_2 x_2 + n_3 x_3 + n_4 x_4 + n_5 x_5}{n_1 + n_2 + n_3 + n_4 + n_5}$$

Using the sigma notation, this becomes much easier to write down:

$$\bar{x}_{est} = \frac{\sum_{i=1}^n n_i x_i}{\sum_{i=1}^n n_i}$$

Now, if we're using Excel, the numerator can be computed as a "sum product" (or vector dot product) of two lists of numbers. One is the list of weights and the other is the list of the values associated with the data ranges. Using this, we can write the Excel version of the weighted average formula as

$$\bar{x}_{est} = \frac{\text{SUMPRODUCT(weights, values)}}{\text{SUM(weights)}}$$

Example 6.2. Effects of different midpoints

Using example the previous example, we can compute the average is several different ways, using different values for the data range. Below is a table showing the estimation of the mean for the salary data above, using the midpoint, left endpoint and right endpoint of each range of values. Keep in mind, though, that unless you have a very good reason for doing otherwise, you should probably use the midpoint to estimate the mean, since that will likely be a better representation of the data in any particular bin. Using the endpoints implies that the data within a particular bin is highly skewed, which might be the case, but would need justification.

Salary Range	Number of employees	Midpoint	Left	Right
\$200,000 - \$249,999	1	225,000	200,000	249,999
\$150,000 - \$199,999	2	175,000	150,000	199,999
\$100,000 - \$149,999	6	125,000	100,000	149,999
\$50,000 - \$99,999	13	75,000	50,000	99,999
\$0 - \$49,999	38	25,000	0	49,999
	Estimated Average	\$54,166.67	\$29,166.67	\$79,165.67

As you can see, the choice of where to place the value for each data range has a huge effect on the estimate of the mean. In fact, if each of the ranges has the same degree of spread (all of the ranges above cover \$49,999) the estimate of the mean from the left and right endpoints will differ by the spread (notice that \$79,165.67 - \$29,166.67 = \$49,999, the spread of the data in each of the ranges). Mathematically, this is easy to prove. Assume that each range has a spread of S . Then, if the left endpoints of the ranges are given by x_i , the right endpoints are given by $x_i + S$ and the estimate for the mean using the right endpoint will be (all sums are from $i = 1$ to $i = n$)

$$\bar{x}_{est} = \frac{\sum n_i(x_i + S)}{\sum n_i} = \frac{\sum n_i x_i + \sum n_i S}{\sum n_i} = \frac{\sum n_i x_i}{\sum n_i} + \frac{n_i S}{\sum n_i} = \frac{\sum n_i x_i}{\sum n_i} + S \frac{\sum n_i}{\sum n_i} = \frac{\sum n_i x_i}{\sum n_i} + S.$$

However, there is any easier way to see what happens using different estimates for the data points. Recall from exploration 4.1.3 that if we add the same amount to every single data point it will shift the mean by that amount exactly. So if we add 10 to each data point, the mean will increase by 10.

Thus, with summarized data, we can never nail down an exact value for the mean. At best, we can estimate it to fall within a particular span of values that is closely tied to the spread each range of data covers.

Example 6.3. Averaging Averages

We can also use the previous examples to understand why we cannot average several averages: Each average must be weighted by the number of data points used to compute that average. For example, if we have two sections of a course being taught and the two sections take the exact same final examination, we might want to determine the overall average on the final exam before deciding what grades to give. If one class scored an average of 82 on the final exam and the other class scored an average of 75 on the final, we cannot simply say that the overall average is $(82 + 75)/2 = 78.5$ because the two classes may have very different numbers of students. The table below uses the correct method, weighted averages, to determine the overall average for different sizes of each class. Notice that the more students there are in the class with the high average, the closer the overall average is to that class's average score.

Class 1 Size (test ave = 82)	Class 2 Size (test ave = 75)	Overall Average
10	30	76.75
15	15	78.5
30	10	80.25
35	5	81.125

Also note that when averaging averages, we are not estimating the mean; in this case we are computing the actual mean of the combined data. Algebraically, we can see why. To compute the weighted average in the above case, the numerator will be $n_1\bar{x}_1 + n_2\bar{x}_2$, but when we multiply an average by the count of data points (the n) we are getting the actual total of all the data points, since

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{N} \implies N\bar{x} = \sum_{i=1}^n x_i.$$

So in this case, the numerator is the exact total of all the data points in each class, even though we do not have the individual scores for any particular student! This is an important thing to note, since additional information about the class may eventually allow us to determine values for some of the students. For this reason, it is sometimes very difficult to release information, even summary data, while also protecting the anonymity of the people described by the data.

Example 6.4. Estimating Standard Deviation

Estimating the standard deviation for summarized data is not that much different from calculating the standard deviation normally. Recall that the formula for the sample standard deviation of a set of data given by x_1, x_2, \dots, x_n is simply

$$\sigma_{n-1} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}.$$

When we only have a summary of the data, though, we must estimate the data points and then weight the deviations based on the number of data points falling near that estimated value. Thus, the formula becomes

$$\sigma_{n-1} = \sqrt{\frac{\sum_{i=1}^m n_i (X_i - \bar{x}_{est})^2}{(\sum_{i=1}^m n_i) - 1}}.$$

where we have used the symbol X_i to refer to the estimate of the value of the data in the i th bin of the summarized data, n_i is the number of data points in that bin, and m is the total number of bins in the summary.

Using the data from the previous section, we can compute each piece of the formula and combine them into an estimate of the standard deviation in the salaries. We will use the midpoint estimate for both the mean and the standard deviation. Recall that the estimate for the mean salary was \$54,166.67.

So, the estimate for the standard deviation is

$$\sigma_{est} = \sqrt{\frac{1.26458E+11}{60-1}} = \$46,296.45.$$

Salary Range	Number of employees (n_i)	Midpoint (X_i)	Deviation ($X_i - \bar{x}_{est}$)	$n_i(X_i - \bar{x}_{est})^2$
\$200,000 - \$249,999	1	225,000	170,833	29184027778
\$150,000 - \$199,999	2	175,000	120,833	29201388889
\$100,000 - \$149,999	6	125,000	70,833	30104166667
\$50,000 - \$99,999	13	75,000	20,833	5642361111
\$0 - \$49,999	38	25,000	-29,167	32326388889
Sum	60			1.26458E+11

Computing the quartiles for this data is relatively easy. Since there are 60 data points, each quartile should contain 15 points ($60/4 = 15$). Thus, starting at the smallest value, the first quartile is between the 15th and 16th data points. This would place the first quartile

in the first bin, between \$0 and \$49,999. The second quartile would fall between the 30th and 31st data points, also in the first bin. The third quartile would lie between the 45th and 46th data points, placing it inside the second bin, between \$50,000 and \$99,999. Our estimates of the statistics relating to this data are then gathered together below.

Mean	54,166.67
Standard deviation	46,296.45
Q1	25,000
Q2 = median	25,000
Q3	75,000

With summarized data, this is about as accurately as one can estimate the standard deviation and the quartiles, since better estimates would require narrowing down the bin widths so that there is not so much possible variation in scores inside a particular bin.

6.1.3 Exploration 6A: Data Summaries and Sensitivity

Open the data file **C06 ExplorationA**. This file contains 1000 observations from five random distributions of data. Each data set contains values in the range of 0 to 100. In this exploration, you are going to investigate two aspects of the data. You will look at the error in estimating the mean and standard deviation of data from different distributions (symmetric, positively skewed, negatively skewed, bimodal and uniform) and you will explore how these errors are affected by the way the data is summarized.

1. Compute the actual values for the mean and standard deviation of each of the distributions (symmetric, positively skewed, negatively skewed, bimodal, and uniform).
2. Below are two different summaries of the data into frequency tables, one using bins of width 10, and the other with width 25. Reproduce these summary tables using your software.
3. Use the techniques of this section to estimate the mean and standard deviation from each of the data summaries. Enter the results in the table below.
4. Once you have recorded the results of your calculations in the table below, think about how our assumptions work when estimating the mean and standard deviation. For which types of data are these estimates most accurate? Why? For which are the estimates least accurate? Why? Keep in mind, these errors may seem small; typically the most error you get with these data is about 0.5 to 1.0, but that's about a 1% to 2% error in estimating these statistics! How does the bin width affect the accuracy of the estimates?

Bin	Symmetric	Positive Skew	Negative Skew	Bimodal	Uniform
0 - 10	22	13	7	10	106
10 - 20	56	72	26	125	99
20 - 30	78	152	36	278	110
30 - 40	140	249	77	97	102
40 - 50	199	201	140	12	90
50 - 60	184	134	242	11	90
60 - 70	165	90	271	124	96
70 - 80	90	53	143	226	113
80 - 90	46	24	48	106	97
90 - 100	20	12	10	11	97

Data summary with bins of width 10.

Bin	Symmetric	Positive Skew	Negative Skew	Bimodal	Uniform
0 - 25	108	158	46	287	264
25 - 50	387	529	240	235	243
50 - 75	398	255	597	247	238
75 - 100	107	58	117	231	255

Data summary with bins of width 25.

	Symmetric	Positively Skewed	Negatively Skewed	Bimodal	Uniform
Actual Mean					
Mean (bin width = 10)					
Mean (bin width = 25)					
Actual St Dev					
St Dev (bin width = 10)					
St Dev (bin width = 25)					

6.2 Two Perspectives are Better than One

Open any newspaper or magazine and you will come across graphs and representations of data that are supposed to help you make sense of some issue or help you decide whether to vote in favor of some proposition or not. You will eventually find yourself sitting in a meeting listening to a presentation with graphs and charts in it. You will probably have employees sending you reports with graphical representations of data designed to help you make a decision. However, it is relatively easy to manipulate your perceptions by presenting a particular graph. By choosing how to present the information, the writer can control the way you perceive the issue. This is true even when the writer is supposedly objective.

With a little work, though, you can look at a graph and mentally convert it to another type of graph. This will provide you with the flexibility of seeing data from multiple perspectives, gaining a much deeper insight into the way the data is structured. This, in turn, will help you make more informed decisions and will help you recognize when someone is trying to manipulate the presentation of the data toward a certain end. For this section, though, we will concentrate on the connections between boxplots and histograms, and we will develop ways to picture one graph when presented with the other type of graph.

Another key benefit to having this flexibility is that you can use a boxplot to help decide how to set up a histogram. Often, it is difficult to set up a useful histogram on the first try. Look back at the histogram of Beef N' Buns service times in Exploration 6A. If the student had first created the boxplot shown below, she might have had a better starting point for making the histogram.

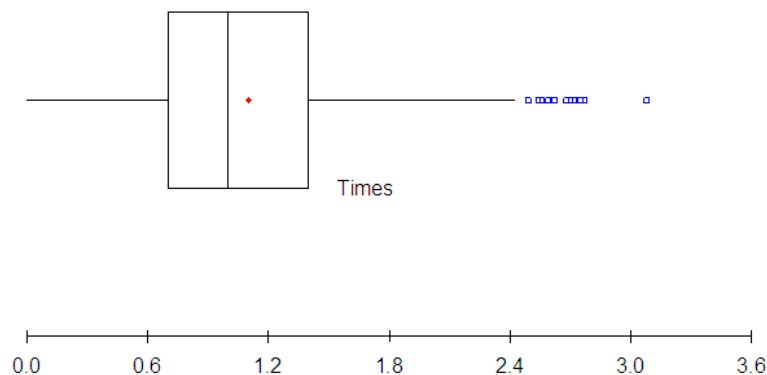


Figure 6.1: Boxplot of service times at Beef n' Buns.

Based on this, she might have set a minimum value of about 0.3 (about halfway between 0 and the first quartile). Then, using 10 bins for the histogram to cover the range from 0.3 to 2.4 would make the bin widths $(2.4 - 0.3)/10 = 2.1/10 = 0.21$ which is about 0.2 (round off to make nice bins in the graph). She could then add two bins (for the “ ≤ 0.3 ” and the open bin on the right side) making the histogram shown below. This graph clearly shows that the data is positively skewed, indicating that the bulk of the service times are below the mean service time (about 1.1 minutes, based on the boxplot).

This section will involve a lot of estimation and inferencing. Estimation involves making

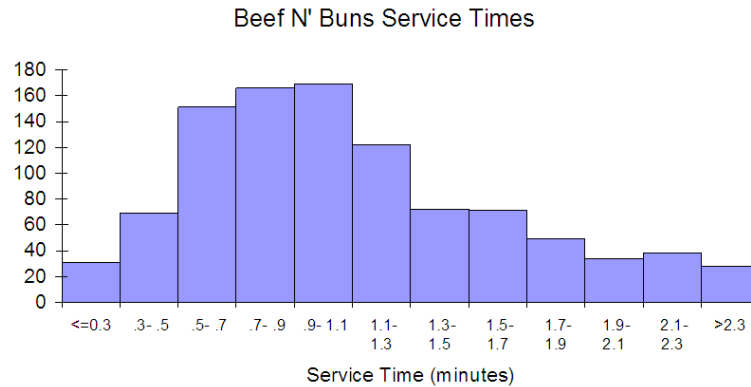


Figure 6.2: Histogram of service times at Beef n' Buns.

a rough guess at some quantity based on either scale or units. Inferencing involves drawing conclusions based on limited information. In order to inference, you will have to interpret the information you are given and “fill in the missing pieces” since you will not have complete information. As part of this process, notice that in the Beef N' Buns service times, reporting the “average service time” as 1.102 minutes (this is the mean) with a standard deviation of 0.542 minutes would misrepresent the situation. Since the data is positively skewed, we can see that most of the data falls to the left (below) the mean service time. In fact, the three largest bins of the histogram are to the left of the mean. This tells us that the mean may not be the best choice for representing the average service time. The median service time of 1 minute may be a better choice. Thus, we can use a combination of graphs to learn more about the data than we could learn from either graph individually. We might also infer that the reason the data is positively skewed has nothing to do with our service overall, but rather with specific orders. If certain orders are taking longer, but these orders do not occur that often, then we might see a few high service times (as high as 3 minutes from the boxplot!) These service times are clearly outliers, and they fall almost four standard deviations from the mean. We could even analyze the percentage of service times within one, two, and three standard deviations above and below the mean (a histogram of the z-scores for the service times would help) to determine whether we should be concerned at all with the service times at Beef N' Buns.

6.2.1 Definitions and Formulas

Percentiles Percentiles are similar to quartiles, except that the data is broken into one hundred pieces, rather than four. For comparison, the first quartile is the same as the twenty-fifth percentile, since one-fourth of 100 is 25. The median is the same as the 50th percentile, and the third quartile is the same as the 75th percentile. Percentiles are often used to break the data down even further than is possible with quartiles. The strict definition of the n th percentile is that it is the observation below which $n\%$ of the data falls. Thus, 90% of the data is less than the 90th percentile and 99% of the data is less than the 99th percentile.

Cumulative Distribution Cumulative distributions are similar to histograms. However, each bin in a cumulative distribution includes all of the observations in the bins to the left of the bin as well. Usually, the number of observations in each bin is expressed as a percentage of the total number of observations so that the right-most bin will have 100% of the observations in it.

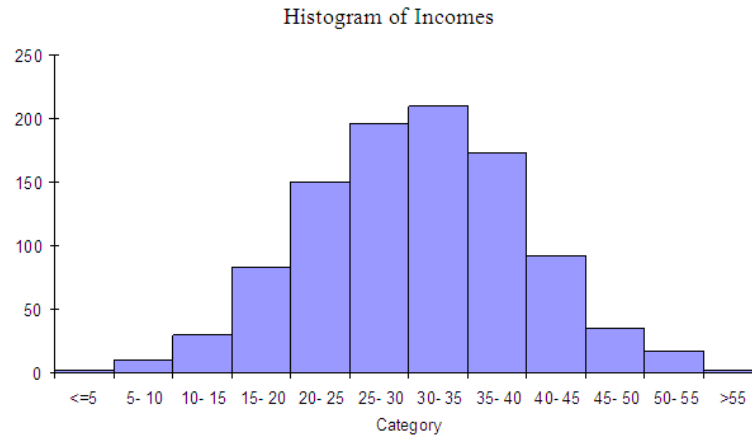


Figure 6.3: Histogram of Incomes.

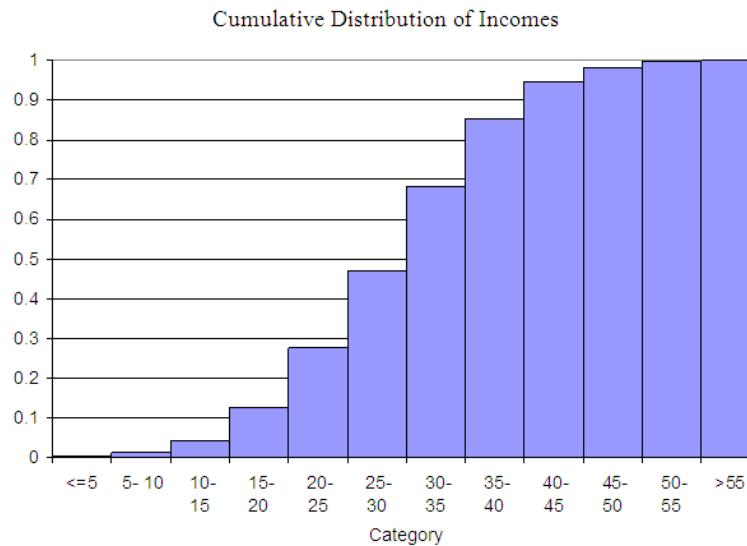


Figure 6.4: Cumulative Distribution of Incomes.

6.2.2 Worked Examples

Example 6.5. From histograms to cumulative distributions

Consider the data on family incomes in Country A from C02 **Incomes**. These are shown in a histogram below. There are 1,000 total observations in the data.

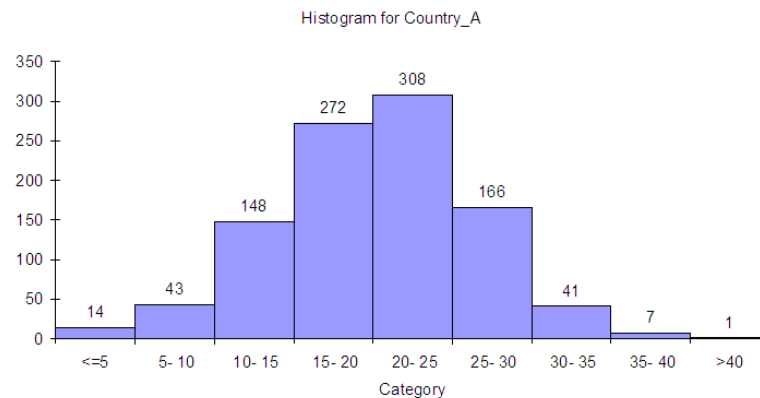


Figure 6.5: Histogram of incomes for 1,000 families in country A.

To convert this graph to a cumulative distribution, we simply start adding. In the first bin, labeled “ ≤ 5 ”, we have a total of 14 observations. In the second bin, we have 43 observations. In the cumulative distribution, the second bin will have $14 + 43$ for a total of 57 observations, since it includes all the bins to the left. The third bin of the cumulative distribution will have $148 + 57 = 205$ observations. The fourth bin “15 - 20” will have $272 + 205 = 477$. Continuing on, we get the totals shown in the graph below.

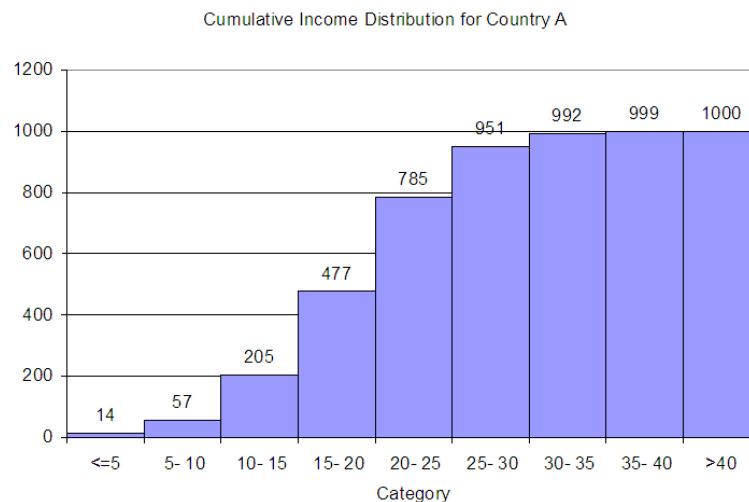


Figure 6.6: Cumulative histogram of incomes for 1,000 families in country A.

Example 6.6. Using the cumulative distribution to sketch a boxplot

Now, to generate a boxplot of the data above, we probably want the cumulative distribution graphed in terms of percentages (of the total number of observations), rather than total

amounts. This graph is shown below. Once we have this, it is relatively easy to determine in which bin each of the quartiles falls. Remember the first quartile is the same as the 25th percentile, so in the graph below, we know that the first quartile is somewhere in the bin marked “15 - 20”. Since 20.5% of the data is to the left of this bin, we can probably guess that the first quartile will be close to the left side of the “15 - 20” bin. We can also find the median; 50% of the data is less than the median, so it must be in the fifth bin, marked “20 - 25”. It is probably close to the left edge of this bin. Interestingly enough, the third quartile includes 75% of the data to its left, so it is also in the fifth bin, “20 - 25”. Q3 is probably close to the right side of the bin. Based on these estimates, then, we can sketch a boxplot on the same scale axis as the histogram. We know where the minimum and maximum are, so we can also compute whether there are any outliers in the data. For this graph, the largest the IQR could be is 10, since Q1, the median and Q3 are in the fourth and fifth bins. Thus, anything further than 15 (three bins) from the end of either side of the box must be an outlier. (The histogram itself shown only one observation in the last bin, so it is the only outlier.)

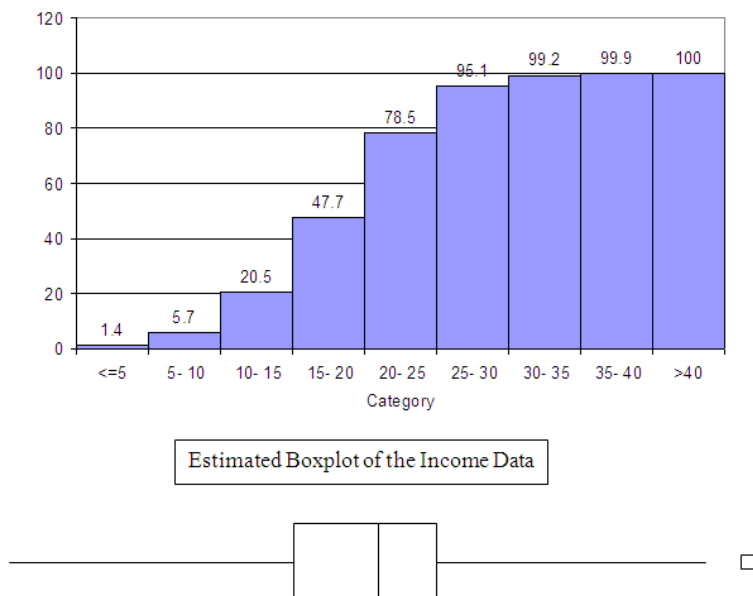


Figure 6.7: Cumulative distribution (as percents) and an estimated boxplot of family incomes.

Example 6.7. From a boxplot to a histogram

There are several ways to sketch a histogram from a given boxplot. One way is to reverse the process in the previous two examples. But there is a quicker way to sketch the histogram, based on the shape of the boxplot. Consider the graphs below, which show four basic histograms and their associated boxplots. All graphs are on the same 0 to 100 axis.

As you can see, the box of the boxplot falls in about the same place as the large bulk of the data. This means that you can start with a boxplot and sketch the bulk-part of

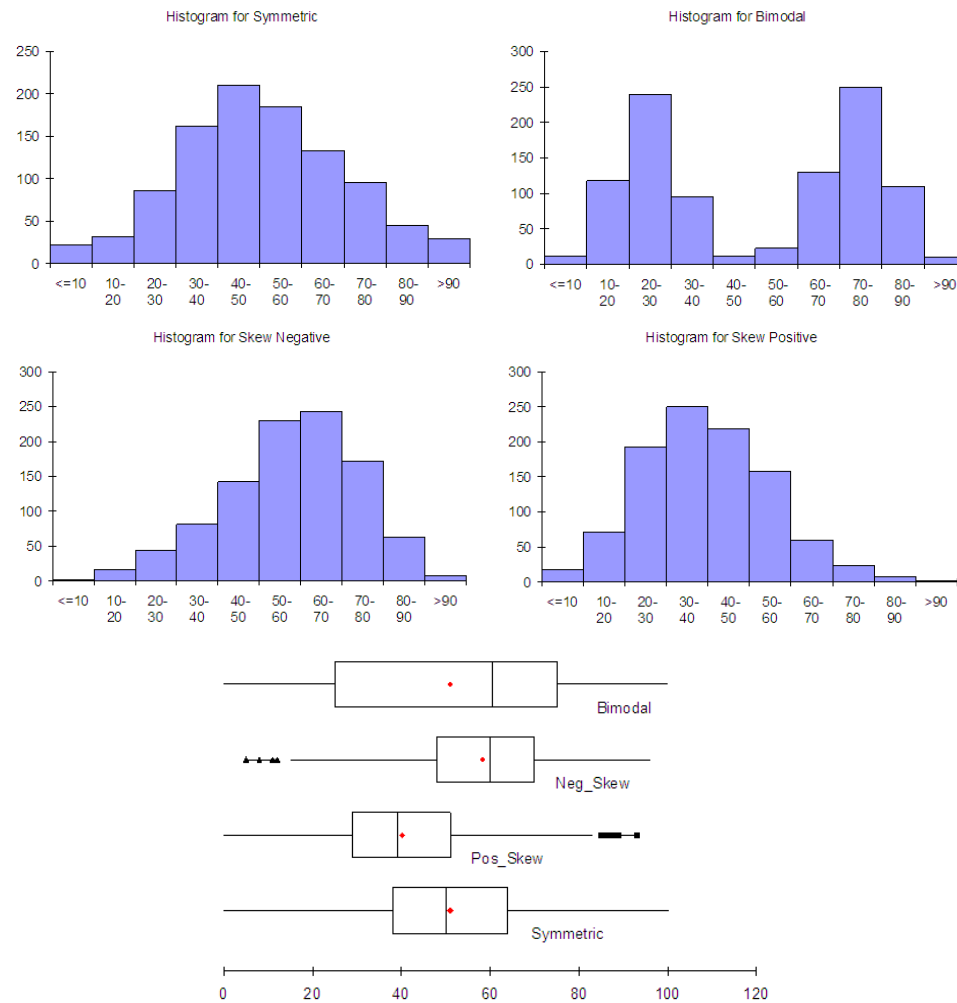


Figure 6.8: Four distributions shown as histograms and their corresponding boxplots.

the histogram where the box is if the box is fairly narrow (if it is a wide box, then the distribution is probably bimodal). Notice that the box is centered between the min and max for a symmetric distribution and for a bimodal distribution, except that a bimodal distribution has a larger spread, so the box is very long. In fact, for bimodal distributions, the box tries include both of the "humps" in the distribution. Notice that for skewed distributions, the box is still where the bulk of the data is, but it is offset from the center. For positively skewed distributions, the bulk is shifted to the left, and the mean is greater than the median. For negatively skewed distributions, the bulk is shifted to the right, and the mean is less than the median.

6.2.3 Exploration 6B: Stock Investment Decisions

You have just started working for a new company, Impressive Business Machinery. As part of the paperwork for your hiring, you have been asked to choose an investment stock for your retirement planning. Your employer offers you four choices and provides you with histograms (figure 6.9) of the daily returns for these stocks over the last 3 months. (You suspect that your employer is testing you, but you can't be sure.) For the near future, which of these stocks would you choose? Why would you choose that stock? How will you justify your decision to your family if it does not perform as well as expected?

1. Which stock did you choose? Why?
2. Discuss your ideas with a partner. Do you still agree that your original choice of stock was the best?
 - (a) If your ideas have changed, what influenced those changes?
 - (b) If your ideas have not changed, what strengthened them?
3. What makes the selection of a stock easy? What makes it difficult?

It may be helpful to sketch the cumulative distribution and a boxplot for each of the stocks. Each graph contains 96 observations (about three months worth of data). It will also be helpful to rank the four stocks from highest to lowest in terms of both the mean and the standard deviation.

Statistic	Highest	Med-High	Med-Low	Lowest
Mean				
Standard Deviation				
Minimum				
First Quartile				
Median				
Third Quartile				
Maximum				

4. After sketching your graphs and completing your estimates, has your decision as to which stock to select changed? Why or why not?
5. Does your selection of a stock depend on what your investment goals are? In what way?

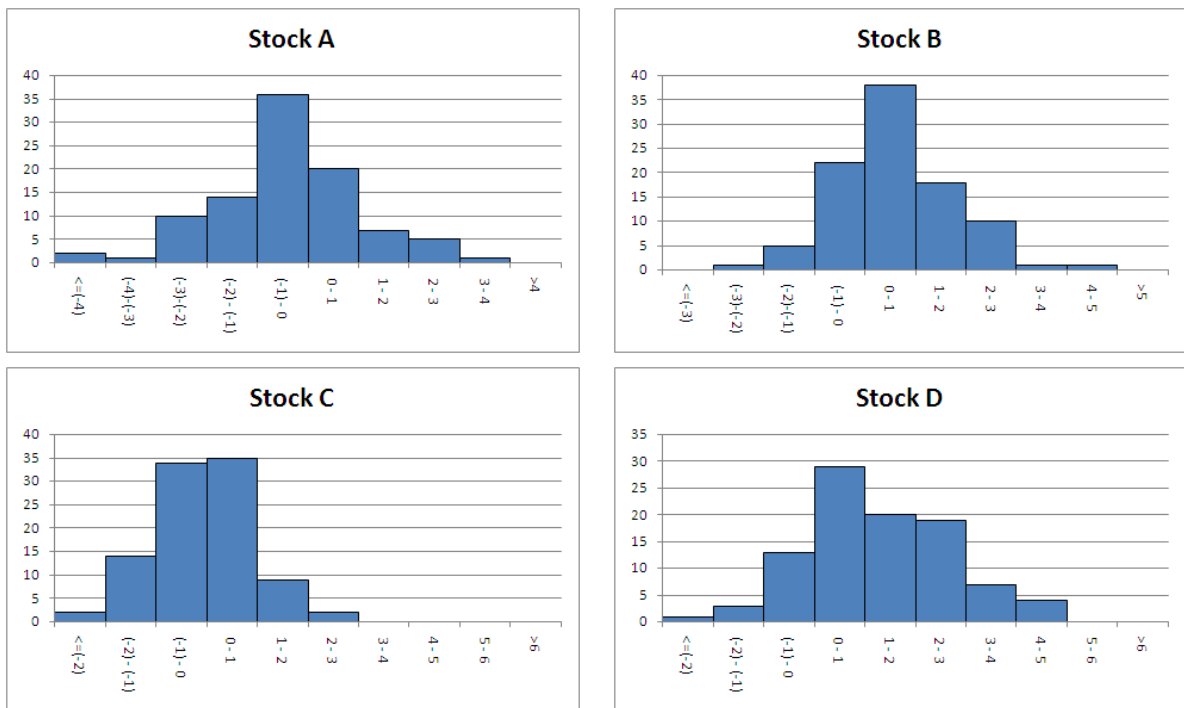


Figure 6.9: Graphs for Exploration Stock Investment Decisions

6.3 Homework

Mechanics and Techniques Problems

6.1. Consider the summarized frequency data found in file C06 Deliveries. This shows data on the unloading times for trucks at StateEx, both broken down by truck type (Semi or Van) and overall.

1. Estimate the average unloading time for semis, vans, and overall.
2. Estimate the standard deviation of the loading times for semis, vans, and overall.
3. Create a histogram of the overall unloading times. Mark the location of the mean on the histogram and add six additional markers to show the location for one, two and three standard deviations both above and below the mean. Where do the average unloading times for the semis fall? What about the vans?

6.2. EverythingRUs is an extremely diversified company, manufacturing and distributing goods as well as providing a variety of services. The table below shows the mean monthly revenue and mean monthly cost for each sector of the company, along with the percentage each sector occupies in the overall revenue and cost structure. Use this information to estimate the mean monthly revenue and mean monthly cost for the entire company. All revenue and cost figures are in thousands of dollars.

Sector	Mean Monthly Revenue	% of Total Revenue	Mean Monthly Cost	% of Total Cost
Food services	\$1,200	15%	\$380	22%
Repair services	\$2,460	18%	\$115	6%
Security services	\$875	11%	\$219	10%
Health and beauty products	\$1,620	14%	\$652	17%
Automobile parts	\$565	8%	\$95	12%
Clothing	\$3,218	13%	\$1,897	15%
Medical supplies	\$1,979	21%	\$934	18%

6.3. Match the histograms below with their cumulative distributions shown in figure 6.10. The graphs in the left-hand column (labeled A - D) are histograms. The graphs in the right-hand column (labeled E - H) are cumulative distributions. Each histogram contains the same number of total observations. The cumulative distributions are given by percentage of total, rather than actual count. Be sure to explain your reasoning.

6.4. Match the histograms (labeled A - D) below with the boxplot (labeled 1 - 4) in figure 6.11 that best matches the data and explain your reasoning.



Figure 6.10: Match the histograms (A - D) with the cumulative distributions (E - H) in problem 3.

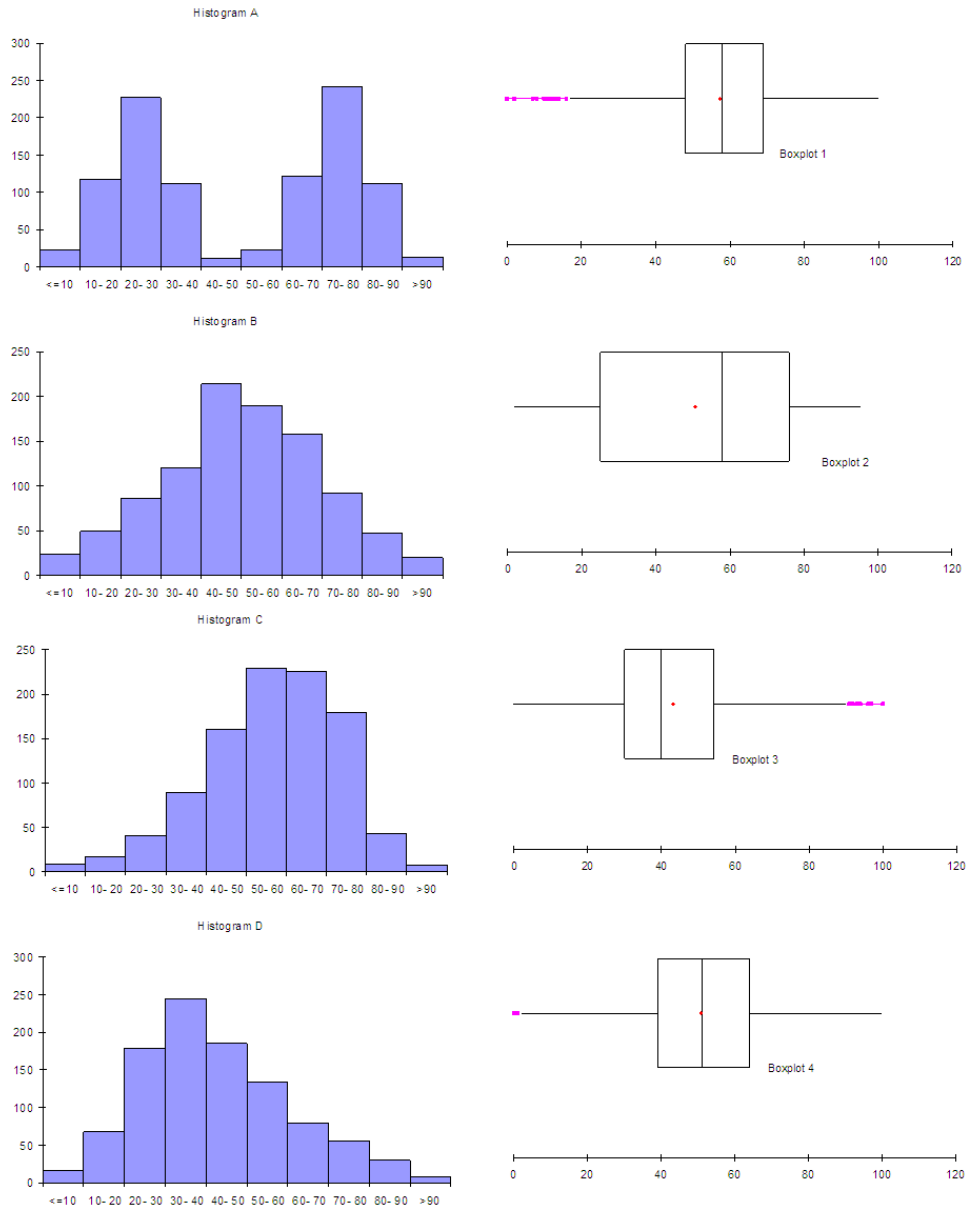


Figure 6.11: Match the histograms (A - D) with the cumulative distributions (E - H) in problem 4.

Application and Reasoning Problems

6.5. Two cumulative distributions are shown in figure 6.12. Describe the differences between the two underlying histograms from which these cumulative distributions were constructed.

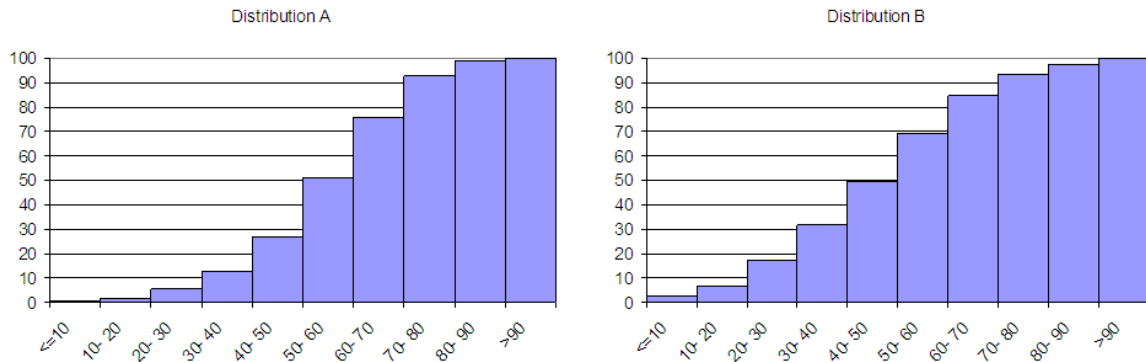


Figure 6.12: Compare these two distributions in problem 5.

6.6. During a recent meeting at your company, the group is examining the breakdown of customers, based on income (see the file **C06 Frequency**). The coworker presenting the data claimed that the company's average customer has a mean income over \$80,000. This coworker then proceeded to explain how this would impact the company. But you suspect something is not quite complete in your coworker's analysis, so at a break in the meeting, you modify the data file as shown to help you explore how the estimated mean and standard deviation change with different assumptions about the distribution of the data. At the top of the spreadsheet is a parameter labeled "Mid". This is a number between 0 and 1 (like 0.25) that represents how far from the left (as a percentage of the total bin size) you would like to position the "midpoint" of the bin for estimating the mean and standard deviation. The rest of the data table is set up similarly to the one shown in example 4 to estimate the mean and standard deviation.

1. Make your own table with three columns to summarize your exploration of the data. The first column should contain values of the parameter "Mid". The second column should contain the estimated mean for that value of the parameter, and the third column should contain the estimated standard deviation. Use at least the following five values for the parameter: 0, 0.25, 0.50, 0.75, 1.0.
2. After looking at the table you have produced, what will you tell the rest of your coworkers in the meeting when you return from the break?

6.7. The graphs in figure 6.13 show boxplots of employee salaries at four local companies. Based on these boxplots, describe the shape of the histograms of the salaries. Estimate

values for the minimum, the number of categories, and the category length that would help create a decent histogram of the salary data.

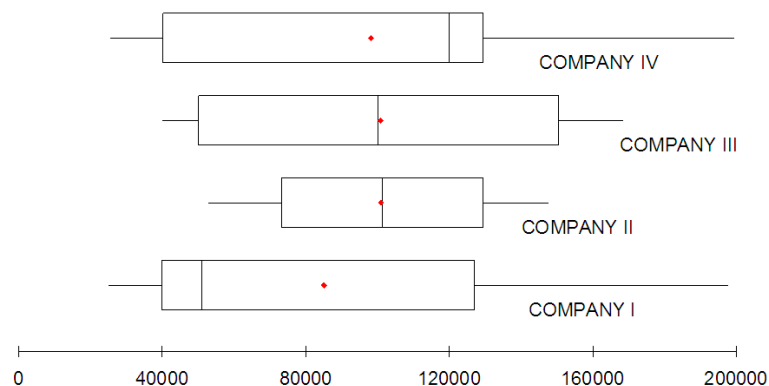


Figure 6.13: Boxplots of salaries at four different companies for problem 7.

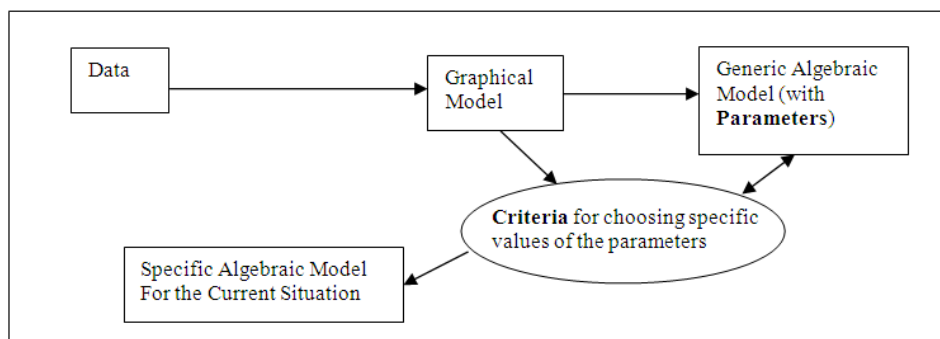
Part III

Analyzing Data Through Linear Models

We have spent some time considering problems that involve data. Usually the data contains observations of several variables, but so far, we have concentrated on understanding each variable separately with statistical methods (like the mean and quartiles) or graphical methods (like boxplots and histograms). Now we are moving into the heart of this book: analyzing the ways in which one or more variables may influence another variable. We are going to start slowly, by analyzing relationships between two variables. We will quickly see the power of graphs for showing the relationships, but we will also see the limitations of such approaches. We'll build on the graphs by developing some numerical methods of determining relationships and their strengths.

But the real heart of this book is in the application of mathematics to construct a model of the data. This means that we must eventually develop equations that embody the relationships between the variables. We will start with some data and then try to build an algebraic description of the data. Once we have this description, we can begin to make predictions, test the predictions, and use these tests to refine the model. After we have developed a reasonable model of the data, this model will enable us to understand the relationships between the variables much better than a list of numbers or even a graph.

Schematically, the method we will use to develop these equations looks something like this:



You will notice two new terms in the above diagram: **parameters** and **criteria**. In order to gain an understanding of these terms, we are going to explore many examples. As you will discover later, most of the types of algebraic models are pretty well understood. They always have the same sort of equation to represent them, but may have different numbers in the equations. These constants can be changed in order to make the general model fit the specific data you are dealing with. In order to choose good values of these parameters, though, we'll need to define criteria that let us select the best values. In the present unit, all our models will be from one category that you have probably encountered before: linear equations. These are the easiest to understand and to interpret, so they are often used as a starting point to analyzing data.

When we use mathematics to help us understand a problem situation, we are often interested in finding how one variable quantity relates to the value of another variable quantity. In particular, we would like to be able to identify which one depends on the other for its value and then, hopefully, to be able to figure out precisely what that relationship is. The variable quantity that depends for its value on the other is called the dependent variable and

the other variable quantity, the one upon which the dependent variable depends, is called the independent variable.

We often call the dependent variable the output variable (or simply output) and the independent variable the input variable (or input) because we are trying to get a precise description of some kind of causal or linking mechanism connecting the two. So we think of inputting various numbers into the mechanism and then watch the corresponding outputs. If we have correctly, or at least adequately, described the connecting link between the input and the output variables, the resulting outputs should match what we actually find in our problem situation. We can then use this input/output linkage to predict what will happen in circumstances for which we have no actual data.

Not all relationships between variables, however, are useful, particularly when they lead to ambiguous results; that is, more than one output for a given input. When this happens, we don't know which output to associate with the given input since there is more than one possibility. For example, think of the price you pay for an airline ticket and the distance you fly. It would seem reasonable that the cost should depend on the distance you fly. You probably know from experience, however, that you could pay very different prices for a flight of 400 miles from your point of origin, depending on, say, the day of the week or a special discount. An input of 400 mi. for the independent variable, distance, could produce two output prices (or more), say \$200 and \$250, for the dependent variable, cost. Our relationship between the input and output produces ambiguous results since we cannot predict what the output price of the 400-mile ticket will be².

A relationship between variables in which a single input³ produces only one output is called a function. Functions can be represented, as we have seen, as tables, graphs, or rules (equations). Here are some connections among the three ways of representing functions:

1. When we are in a problem situation in which we know what the dependent and independent variables are, but we don't know how they numerically relate to each other, we often gather experimental data and organize it into charts with each input value linked to its corresponding output value.
2. After we have a table of related pairs of input and output values, we can graph these pairs on inputs and outputs on a scatterplot. The values for the independent variable (input) are marked off along the horizontal axis, and the values for the dependent variable (output) are marked off along the vertical axis.
3. Sometimes we know how the variables are connected by a rule or an equation. Very often, we do not. If we have collected data and organized it into a chart, however, we will learn how to create an equation from the data (with the help of technology) that can replicate the data fairly accurately in most situations. Moreover, we will see that the graph of this equation fits the graph of the original data remarkably well. The analysis technique that allows us to move from data to equations is called regression analysis and is used extensively in the social and physical sciences.

²However, we can create mechanisms that accept multiple input variables (say distance, current date, date traveling, discounts, destination, starting point, etc.) and produce a single output (cost of ticket). In many cases, this will resolve the ambiguity caused by using only one independent variable.

³Or a set of inputs

Before we can begin any of the above steps, however, we have to identify what we think the independent and dependent variables might be in a given problem situation. At this preliminary stage of our analysis, a graph of the possible relationship between the two variables is an extremely simple and effective way to test whether we have identified the independent and dependent variables correctly, as well as to form a basic notion of how the independent variable relates to the dependent variable.

Key Communication Strategy:

Memo Problem: Modeling StateEx Unloading Process

To: Analysis Staff
From: Project Management Director
Date: July 23, 2017
Re: Shipping and unloading process at StateEx

Now its time to get into the real aspect of our work for StateEx: modeling the process of unloading the trucks. Attached is a data file showing how the unloading process works not only at the primary warehouse, which has a loading dock, but also at the endpoints of the delivery process. These data are aggregated, showing information from many different data collection forms in a single spreadsheet for analysis, with one row of data on each delivery. After examining the data file carefully, develop a brief explanation of what we might expect the impact of each of the explanatory variables to be on the total unloading time. The point here is not to be perfect, but to develop some intuition about the data and the relationships you expect to see. Begin your modeling by determining the most relevant variables that affect the unloading time and formulate some simple models (one variable) with the most relevant variables and explain what these models indicate about the unloading process. This will give us a starting point for the deeper analysis. Be sure to compare these models in detail.

Once youve identified the basic models and explained what they tell us, formulate a full regression model for the unloading times. Then refine this model by eliminating the insignificant variables from the model. Be sure you drop the least significant variables one at a time, since the significance of the other variables will change as you eliminate insignificant ones. Some of the important variables may be categorical, so make sure you deal with those variables correctly in your models as well. There may also be a numerical variable which could be treated categorically, so investigate this briefly. Make sure that you identify any outliers in the residual plots; we may need to investigate these further.

Your final report should contain a comparison of the models you formulate in terms of accuracy in fitting the data, accuracy in predicting other observations, and what they explain about the unloading process at StateEx. You should conclude with a recommendation for the StateEx manager. This recommendation should include how to estimate crew sizes for a particular delivery and it should contain some rules of thumb for determining if the unloading crews are slacking off.

Attachments: Data File StateEx_Deliveries

CHAPTER 7

Correlation¹

So far in this book, we have limited ourselves to looking at only one variable at a time, trying to learn as much as possible about that single variable. However, most of our data is made up of many variables, all interacting and having effects on each other. In this chapter you will explore relationships between two variables using graphical methods (scatterplots), computational methods (correlation), and algebraic methods (equations of functions).

- Section 7.1 introduces **correlation** as a tool for measuring the strength and direction of the linear relation between two variables.
- Section 7.2 gives you the tools you need to visually approximate relationships with lines and to interpret the slope and intercept.

As a result of this chapter, students will learn

- ✓ How to read and interpret a scatterplot
- ✓ How correlation describes the relationship between two variables
- ✓ The meanings of “positive” and “negative” relationships between two variables
- ✓ About the slope and y-intercept of straight lines and how to compute these

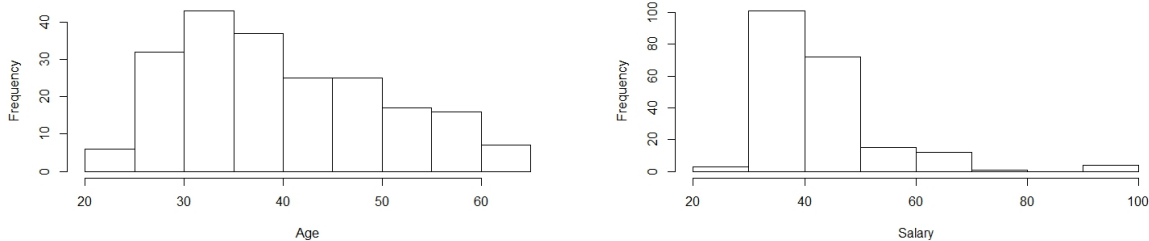
As a result of this chapter, students will be able to

- ✓ Identify variables with a positive or negative relationship using the correlation coefficient
- ✓ Construct a correlation table to determine which variable relationships are most influential
- ✓ Estimate the correlation coefficient of two variables based on a scatterplot
- ✓ Set up a scatterplot according to conventions about axes, etc.
- ✓ Add trendlines to a scatterplot

¹©2017 Kris H. Green and W. Allen Emerson

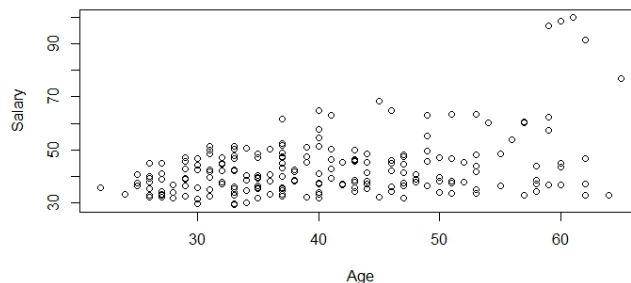
7.1 Picturing and Quantifying the Relationship Between Two Variables

In many of the previous examples in this book you have probably been tempted to go too far in your conclusions. For example, if you were to look at information about employees at a company and you learned that the salaries were negatively skewed and that the ages of your employees were also negatively skewed, you might be tempted to claim that one variable (for instance, age) influences the other variable (in this case, salary).



However, it would be dishonest to make such a claim with the tools we have discussed so far. In fact, the relationship between the two variables could be exactly the opposite of what you claim: it could be that the low salaries are all earned by employees who are older and that younger employees are making more money. It is even possible that the two variables are unrelated entirely. All of our tools up to now have been tools to analyze data one variable at a time. In order to speculate about relationships between two or more variables, we need new tools that include two variables at a time. A graphical tool for this analysis is the scatterplot. This is a two-dimensional graph made up of points where each point represents a pair of observations, one for each of the two variables you are comparing. In this way, you can quickly spot connections between variables. Such connections are called **correlations** and can also be computed numerically with a fairly simple formula based on z-scores.

Consider the employee salary example above. One could speculate that the points representing the salary and age of each employee would show that older employees tend to have higher salaries (after all, they have been working longer, have more experience and have had more opportunities for promotion). If the graph shows this, then there might be a connection between the two variables. In this case, it appears there is, at most, a weak connection.



We want to emphasize this as strongly as possible. Simply because the correlation between two variables is high does not mean that one variable is causing the changes in the other. Consider the following situation: You are interested in the performance of your stock brokers at a large investment firm. If you looked at the amount of money each broker earned for the firm and compared this to the number of cups of coffee that broker drinks each day at work, what would it mean if there were a strong positive correlation? Would that mean that drinking more coffee makes you a better broker? Clearly, this is absurd. What it does mean is that brokers who make more money for the firm also tend to drink more coffee. That's all it means. Why might this be so? There are many reasons. It could simply be that the amount of coffee consumed is a surrogate for the number of hours the broker works. More hours worked might lead to more money for the broker. But more hours worked will probably involve drinking more coffee.

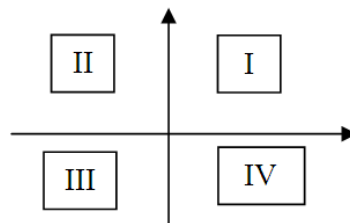
For the remainder of this book, we will be dealing with how to represent relationships among variables. Our goal is to develop these relationships into mathematical equations called **functions** that we can use in our decision-making.

7.1.1 Definitions and Formulas

Scatterplot A scatterplot is a graph that takes sets of observations of two variables and plots them as points on a graph. Each point corresponds to a single observation of both variables. The points are identified by an ordered pair, with the horizontal variable listed first. These ordered pairs are written as (x, y) . After each point in the data is plotted, the scatterplot can help determine if there is a relationship between the two variables.

Axis and axes All graphs have an axis that shows a scale and in which direction the variable being graphed is increasing. "Axes" is the plural form of the word axis.

Quadrants In a scatterplot, the horizontal and vertical axis cross at a point called the origin which has coordinates $(0, 0)$. This divides the Cartesian plane (all the possible points of the scatterplot) into four regions called quadrants. Each quadrant is numbered according to the graph shown here:



Dependent Variable The dependent variable is usually graphed on the vertical (y) axis. This is the variable that you suspect will be affected by a change in the other variable. It is also referred to as the predicted variable, response variable, or the fitted variable when you are building models.

Independent Variable The independent variable is usually graphed on the horizontal (x) axis. This is the variable that you suspect determines the value of the dependent variable. It is graphed on the horizontal axis because it is easier for the eye to scan left-to-right in picking a value for it and then scanning up the graph to determine the value of the dependent variable that corresponds to the value of the independent variable you picked. It is also referred to as the explanatory variable or predictor variable.

Direct Relationship If the cloud of points on the scatterplot seems to move upward as the eye scans across the graph from left-to-right (as shown in figure 7.1), then the relationship between the two variables is said to be a direct relationship. This means that as the independent variable increases (gets larger in value), so does the dependent variable. Such a relationship is also referred to as a positive relationship or an increasing relationship. The graph in figure 7.1 shows a strong positive relationship between two variables.

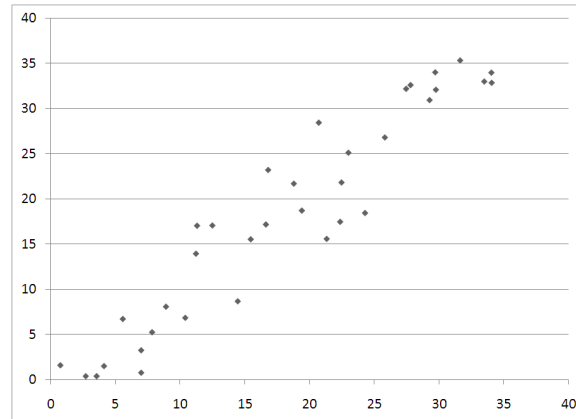


Figure 7.1: Illustration of a direct relationship between the dependent variable Y and the independent variable X .

Indirect Relationship If the cloud of points on the scatterplot seems to move downward as the eye scans across the graph from left-to-right (as shown in 7.2), then the relationship between the two variables is said to be an indirect relationship. This means that as the independent variable increases (gets larger in value), the dependent variable decreases. Such a relationship is also referred to as a negative relationship. The graph in figure 7.2 shows a strong negative relationship between the two variables graphed.

Correlation coefficient The correlation coefficient is a way of numerically determining two things:

1. Whether the relationship between two variables is direct, indirect or neither.
2. The strength of the linear relationship between two variables.

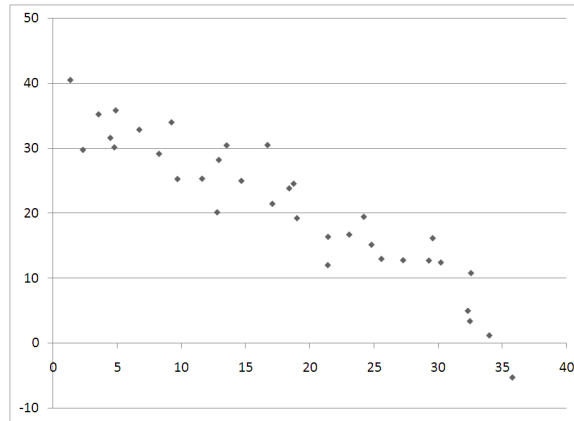


Figure 7.2: Illustration of an indirect relationship between the dependent variable Y , shown on the vertical axis as is standard, and the independent variable X on the horizontal axis.

Correlation is a number between -1 and $+1$ and is determined by the formula below, based on the z -scores of the two variables (the variables are called x and y in the formula).

$$\text{Correlation}(x, y) = \frac{1}{n-1} \sum_{i=1}^n z_{x_i} z_{y_i}$$

Notice that since this formula is based on the z -scores of the data, the overall correlation coefficient has no units. This makes it easier to interpret. Positive correlation means positive relationship, negative correlation means a negative relationship. Correlations close to $+1$ or -1 indicate strong relationships, while correlations close to zero indicate weak relationships, as shown in figure 7.3.

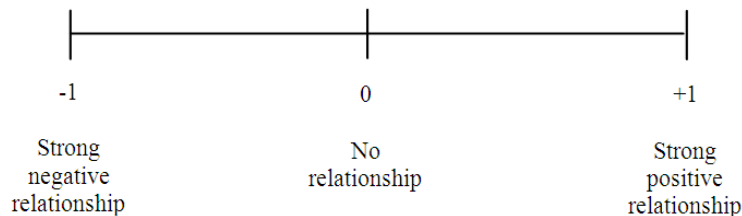


Figure 7.3: The scale of correlation, from -1 to $+1$.

Correlation Matrix A correlation matrix (see table 7.1 for an example) shows the relationships among many variables at once in a table format. Each variable is listed twice - once along the top of the table and once along the side of the table. Each cell of the matrix contains the correlation between two variables (one from the row and one from the column the cell is in). Usually such tables are only half filled in, since the correlation of x with y is the same as the correlation of y with x . Also, the diagonal entries are all $+1$, since a variable has a perfect correlation with itself.

Table of correlations	Age	Credits	WorkHours	SleepHours	GPA
Age	1.000				
Credits	0.221	1.000			
WorkHours	0.658	-0.439	1.000		
SleepHours	0.775	-0.886	-0.228	1.000	
GPA	0.342	0.669	-0.824	0.713	1.000

Table 7.1: Sample correlation matrix of relationships among the variables describing students at a large university.

Strong Relationship A strong relationship between two variables is seen in scatterplots with points that are tightly bunched together around some pattern (like a line or a curve). Strong relationships have correlations close to $+1$ or -1 . Examples of strong relationships are shown in Figure 7.1 and figure 7.2.

Weak Relationship In a weak relationship, such as that shown in figure 7.4, there is almost no connection between the two variables. Figure 7.4 shows such a situation. This might result from graphing the two variables “grade on a test” and “amount of pizza consumed”. Weak relationships have correlations close to zero.

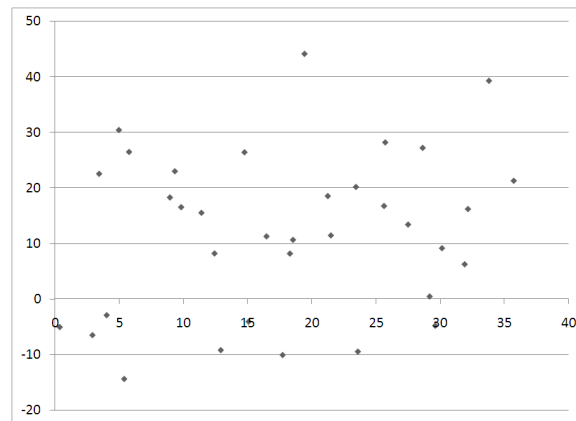


Figure 7.4: XY scatterplot showing a very weak relationship between the two variables.

7.1.2 Worked Examples

Example 7.1. Reading Variables and Relationships from a Graph

Suppose we have collected data on students taking the SAT shown in figure 7.5. If we have observations of the variables Study Time and Score, we might try to examine whether there is a relationship between the amount of time a particular student studies for the test and the score that this student receives on the test. We would then select Study Time as the independent variable, since we are guessing that study time predicts the test score. To create

the scatterplot we then draw the axes and label them Study Time on the horizontal axis and SAT Score on the vertical axis. Next, we select a scale for each axis, based on the range for each variable. (Recall that the range is the difference in the maximum and minimum observations.) Finally, for each observation, we place a dot on the graph. The values of the two variables will determine where each dot is placed. For example, if one student studied 19 hours for the test and scored 741 (on a scale of 400-1600), the dot representing her score would be located along a line passing through the 19 hour mark on the horizontal axis, and it would be lined up with the 741 mark on the vertical axis.

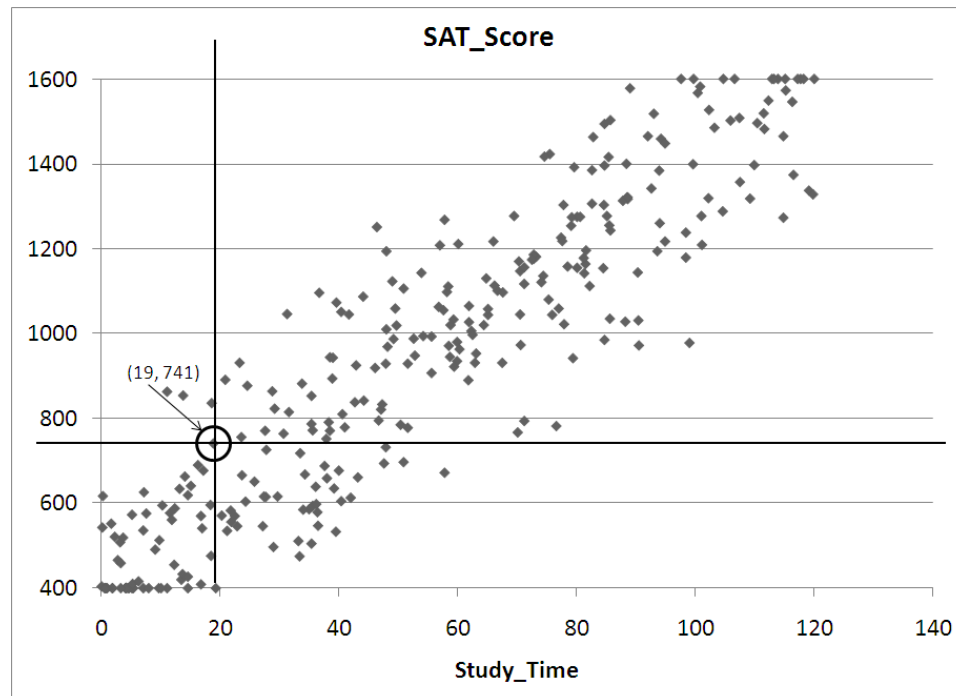


Figure 7.5: Scatterplot of SAT scores versus hours of study time.

After plotting all of the data on the graph above, it is clear that the variable Study Time has a strong influence on the final score a student receives on the SAT. The relationship looks quite strong and positive: as study time increases, students score higher on the test. Notice however, that the relationship is not perfect. There is a wide range of scores for students spending, for example, 20 hours studying for the test. In fact, all we can say for certain is that 20 hours of studying will probably get a score between 400 and 800 on the test. If we increase the amount of studying, though, the final score is quite likely to be higher. For example, 60 hours of studying seems to result in a score between 1000 and 1300.

Example 7.2. Reading a Correlation Matrix

Suppose we collect observations of several variables related to employees at Gamma Technologies: Age, Prior Experience (in years), Experience at Gamma (in years), Education (in years past high school), and Annual Salary. The matrix of correlations of such data might look like this:

Table of correlations	Age	Prior Experience	Gamma Experience	Education	Annual Salary
Age	1.000				
Prior Experience	0.774	1.000			
Gamma Experience	0.871	0.443	1.000		
Education	0.490	0.362	0.308	1.000	
Annual Salary	0.909	0.669	0.818	0.650	1.000

To read the table, simply choose two variables and look up the intersection of those two variables in the table. If we choose Age and Gamma Experience, the correlation is 0.871. This number is quite high, indicating a strong positive relationship between the Age and the Gamma Experience variables. Thus, we expect that older employees have been with the company longer. (This is not much of a discovery.) However, the strongest relationship between two variables in this study is between Age and Annual Salary. The correlation of 0.909 indicates that Age may be an excellent indicator of salary: older employees make more money.

Also, notice that the correlation between any variable and itself is always 1.000. This is because any variable is perfectly correlated with itself. (If you know an employee's Age, you can predict their Age exactly.) You may also notice that the correlation of Prior Experience with Salary (0.669) is slightly higher than the correlation of Education with Salary (0.650). This might indicate that Gamma places slightly more importance on experience over education. The last thing to notice is that part of the chart is blank. This is because the correlation of the variable Age to Prior Experience will be the same as the correlation between Prior Experience and Age. There is no need to duplicate the information.

Example 7.3. Strong and Weak Correlation Through Pictures

Note: Before reading this example, you may wish to review the material on z-scores in section 5.1.

Consider the gas mileage for cars, a topic you may have spent some time thinking about recently. We have collected data on a sample of vehicles on the road in the file C07 AutoData. The data include the gas mileage (measured in MPG or miles per gallon), the power of the engine (measured in horsepower) and the weight of the vehicle (measured in pounds). What general conclusions can we draw from the data, as represented in the graphs and charts below? As you can see from the graphs in figure 7.6, all three variables are strongly correlated. However, two of the relationships are inverse relationships: As the weight of the vehicle increases, gas mileage decreases. As the power of the engine increases, the mileage also drops. However, the positive relationship shows us that larger cars (as measured by weight) tend to have more powerful engines (by horsepower). Three graphs illustrating various relationships among variables about automobiles in figure 7.6.

Which of these relationships is the strongest? This is much harder to tell from the graphs. It appears that all three of the relationships have very similar correlations (in magnitude). To estimate the correlations, we need to know the means of the three variables.

Variable	MPG	Engine	Weight
Mean	31.50	90.84	2756.52

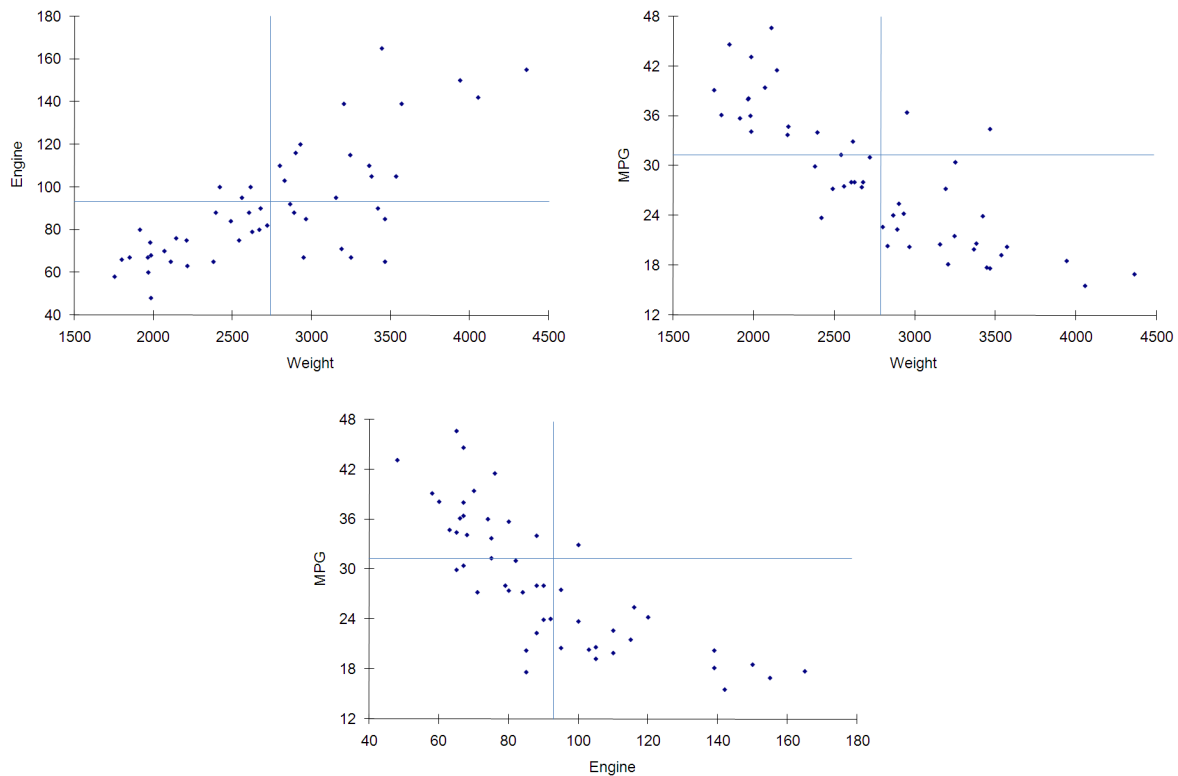


Figure 7.6: Relationships in car data among Engine (horsepower), mileage (MPG), and Weight (pounds).

Now, we can draw in the means (this has been done in figure 7.6) and use this to estimate the correlation between the variables in each graph. In the “Engine vs. Weight” graph, notice that most of the observations are in the upper-right and lower-left quadrants. This means that most of the observations will serve to increase the correlation coefficient. In the upper-right quadrant, $z_x > 0$ and $z_y > 0$ for each observation, so the product is also positive. In the lower-left quadrant, $z_x < 0$ and $z_y < 0$, so the product is also positive. However, there are a few observations in the upper-left quadrant which decrease the correlation (since the z_x scores of these observations is negative and the z_y scores are positive, this contributes a negative to the total correlation). There are quite a few observations in the lower-right quadrant which will also decrease the correlation ($z_x > 0$, but $z_y < 0$ for these). Based on this, we expect the correlation to be high and positive, but not perfect. A good estimate would be around 0.8.

Since the other graphs are similar in terms of spread, we expect their correlations to be the same magnitude as the first graph. Since they represent inverse relationships, though, these correlations must be negative. You could reasonably estimate the correlations to be about -0.8 for both graphs.

Example 7.4. How Correlation Works

Consider the data graphed on the scatterplot shown in figure 7.7. For each of the five

data points, we can fill in the table below in order to estimate the effect of each point on the overall correlation of the data.

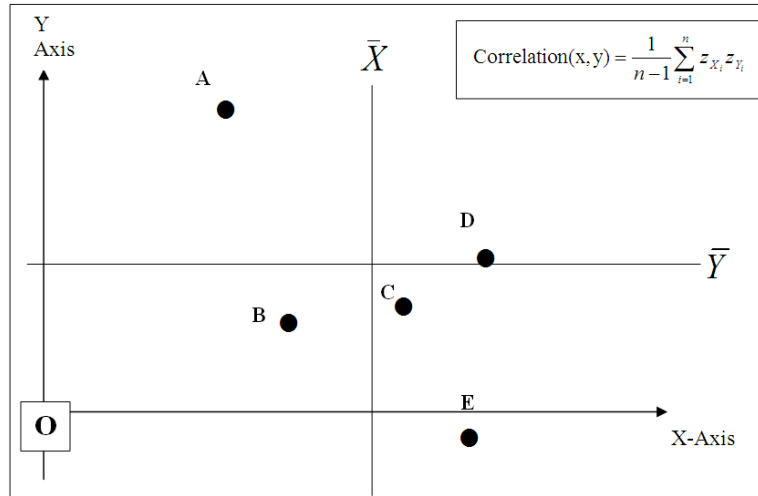


Figure 7.7: Scatterplot of points with means of X and Y shown.

Point	A	B	C	D	E
Sign of Z score of Point's z_x	Negative	Negative	Positive	Positive	Positive
Sign of Z Score of Point's y	Postive	Negative	Negative	Negative	Negative
Sign of the products of the Z Scores	Negative	Positive	Negative	Negative	Negative
Increase or Decrease Correlation	Decreases	Increases	Decreases	Decreases	Decreases
Size of effect on correlation	A lot	A little	A little	No effect	A lot

So we see that four of the five points contribute to a negative correlation, while one (B) increases the correlation. Point D has almost no effect on the correlation because the y -coordinate of D is almost equal to \bar{y} , making its z -score basically zero. Overall, these data indicate a correlation of maybe 0.7 or so.

7.1.3 Exploration 7A: Predicting the Price of a Home

Instructions: Using data file C02 Homes, answer each question below.

1. Compute the mean and standard deviation for each of the following numerical variables:

	Taxes	Year	Acres	Size	Value	Price
Mean						
Standard deviation						

2. Using the mean as a model, how much would you say the *typical* single-family home costs in this market?
3. How reliable is your estimate?
4. Using a table of correlations, calculate the correlation coefficient (r) for the following pairs of variables:

	Taxes	Style	Bath	Bed	Rooms	Year	Acres	Size	Value
Price									

5. Based on the correlation coefficients, which of the above variables seems to have the MOST effect on the PRICE of a house? Which as the LEAST effect?
6. Generate a scatterplot that describes the relationship between PRICE and SIZE. Which variable is the independent variable (should be on the x-axis)? Which variable is the dependent variable (on the y-axis)? What is the Correlation for this relationship? Your scatterplot should look something like figure 7.8.
7. Draw a vertical line on the above chart to represent the MEAN for SIZE

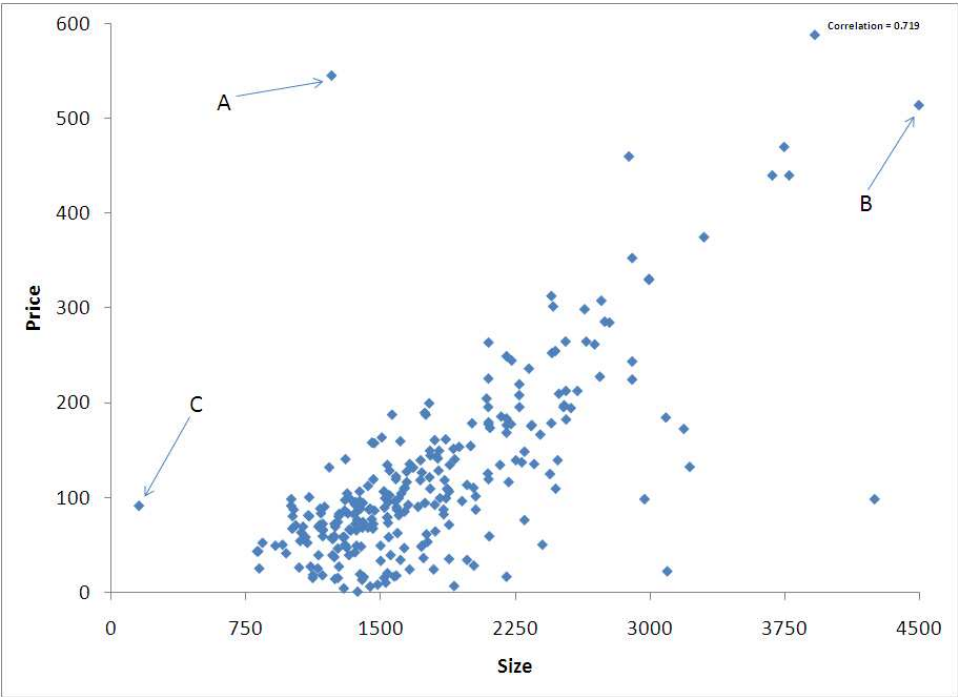


Figure 7.8: Scatterplot showing home price versus size.

8. Draw a horizontal line on the above chart to represent the MEAN for PRICE.
9. Hover your mouse over the points marked A, B and C on the chart to determine the values for PRICE and SIZE at each point. Then fill in the table below to estimate the correlation.

	SIZE	PRICE	z-score for SIZE (X)	z-score for PRICE (Y)	Total Contribution to the Numerator of Correlation
A			$z_x = \frac{(\quad)-1772}{631}$	$z_y = \frac{(\quad)-121}{94}$	
B			$z_x = \frac{(\quad)-1772}{631}$	$z_y = \frac{(\quad)-121}{94}$	
C			$z_x = \frac{(\quad)-1772}{631}$	$z_y = \frac{(\quad)-121}{94}$	

7.2 Fitting a Line to Data

The easiest relationship between two variables to model is a linear relationship. Straight lines are easy to picture, they have simple equations, and each part of a straight line equation can be easily interpreted into real-world terms. Consider the data shown in figure 7.9. The independent variable is the size of a home in hundreds of square feet and the dependent variable is the price of the home in thousands of dollars. The data were taken from a sample of fifteen homes in a single neighborhood that all sold within one year. The graph clearly indicates a strong linear relationship between the two variables: larger homes tend to have higher prices.

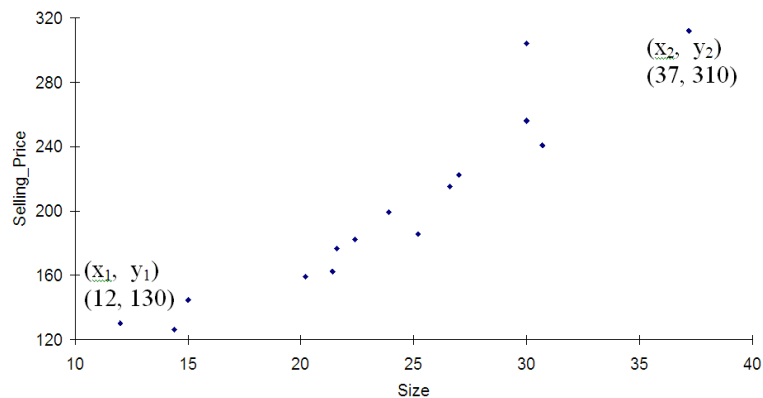


Figure 7.9: Plot of home selling prices (thousands of dollars) versus home size (square feet).

We can easily draw a straight line through this data that does a reasonable job representing the data. But what do we really mean by “representing the data”? Clearly we cannot draw a single straight line which passes through all of the data points. How then do we decide what the best line is? Each line is characterized by two numbers, **slope** and **y-intercept**. By carefully choosing these numbers we can make the line fit the data better. But how? Slope is basically the tilt of the line: larger values of the slope make the line steeper, positive slopes indicate lines that slant upward, and negative values indicate downward slanted lines. The line for the home data above must have a positive slope, since the points on the graph that indicate larger homes (those farther to the right) also tend to have higher prices. Furthermore, since the two extreme data points are about $(37, 310)$ and $(12, 130)$ we see that an increase in size of $37 - 12 = 25$ hundred square feet results in a price increase of $310 - 130 = 180$ thousand dollars. Thus, the slope of the line is approximately $180/25 = 7.2$ thousand dollars per hundred square feet of size.

Now that we have an estimate of the slope for this line, we can compute the y-intercept. The equation of any line has the form $y = A + Bx$ where A is the y-intercept and B is the slope. We can substitute the slope we calculated (7.2), and use the fact that the line must pass through one of the points we used. To do this, we just plug all the known information about that point into the equation of the line and use algebra to find the value of A that makes the line with that slope pass through that point. Suppose we use the point $(37, 310)$. This means that $y = 310$, $x = 37$. Remembering that the slope, B is 7.2, we substitute

these values into the generic equation of the line to get the equation: $310 = A + 7.2 * 37$. Solving for the missing value, A , we get $A = 310 - 7.2 * 37 = 310 - 266.4 = 43.6$. Thus, our estimated equation for the line is $y = 43.6 + 7.2x$.

Notice that we referred to this equation as an estimate of the equation of the line fitting these data. Even though our math is all correct, we started by picking two points from the data. And there's no guarantee that we picked the best two points. For now, we will save those details for the next chapter. In this section, we will explore the equations of straight lines and use them to model relationships between two variables. We will also see how these equations can be used to make predictions about data that is not part of the data set. This involves specifying a value of the independent variable and calculating the dependent variable from the equation. We will also see how to determine values of the independent variable that give rise to specified values of the dependent variable. This is usually referred to as "solving an equation."

7.2.1 Definitions and Formulas

Slope The slope of a straight line is a number that tells you exactly how much the dependent variable will increase for a given increase in the independent variable. Usually it is represented as a decimal number or a fraction and it is calculated from looking at the "rise" of the straight line between two points (this is the vertical distance between them) and comparing this to the "run" (the horizontal distance separating the two points). If the two points are labeled (x_1, y_1) and (x_2, y_2) then the slope is the change in y divided by the change in x . (Note that the Greek symbol delta, Δ , represents the phrase "change in".) The slope will have units given by the ratio of the y units to the x units.

$$\text{Slope} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

Y-intercept The y-intercept is the position on the vertical axis (possibly not shown on the graph) where the line crosses the axis. The intercept will have the same units as the y variable.

Equation of a straight line The most common way to represent the equation of a straight line is in **slope-intercept form**:

$$y = A + Bx.$$

In this equation, A is the y-intercept and B is the slope. The two other letters represent the variables: x is the independent variable, y is the dependent variable.

The equation can also be represented in **point-slope form**:

$$y - y_1 = B(x - x_1).$$

where B is again the slope and (x_1, y_1) is a point on the line. Both forms are equivalent; they are simply written in a different form to make it easier to use one or the other, depending on which two pieces of information you have. For example, if you re-arrange the point-slope form, you can produce $y = Bx + (y_1 - Bx_1)$, showing that the y-intercept $A = y_1 - Bx_1$.

Trendline A trendline is a line drawn on a graph to represent the relationship between two variables. These trendlines can take many forms. There are five basic trendline options: linear, exponential, logarithmic, power, and polynomial. Trendlines are also called **lines of best fit**, even though trendlines are not always straight lines. Perhaps they should be called curves of best fit or trendcurves?

Linear relationship A linear relationship between two variables is characterized by a constant slope. A scatterplot of the two variables shows the points lining up very closely along a straight line. The graph in figure 7.10 shows a linear relationship, a linear trendline for it, and the slope and y-intercept of that trendline.

Function A relationship between two variables (called the independent and dependent variables) in which every value of the independent variable is associated with one and only one value of the dependent variable. Functions can be represented graphically (as lines or curves on a set of axes), as a table showing sample values, by an equation, or by a verbal description in words. On a graph, the test of whether a relationship is represented with a function is called the vertical line test and consists of drawing vertical lines on the graph. If any line crosses the graph more than once, the relationship is not a function.

7.2.2 Worked Examples

Example 7.5. Estimating slope and y-intercept from a scatterplot

In the graph above (figure 7.10) we can easily make estimates of the slope and y-intercept of the trendline and use these to write down its equation. This equation could then be used to make predictions of other values.

The y-intercept appears to be about 21. It might be a little smaller, but clearly the trendline hits the y-axis above the tick mark for 20.

The slope is a little harder. We need two points on the line. Fortunately, this trendline seems to pass through several of the points on the scatterplot. (This is not always the case. The procedure for finding trendlines does not guarantee that the trendline will pass through any of the data points.) This line seems to pass through the points $(3, 17)$ and $(14, 5)$. Thus, when the run is $(14 - 3) = 11$ the line has a rise of $(5 - 17) = -12$ (notice the negative sign; it means that the relationship is indirect or decreasing). Thus, the slope of the line is approximately rise over run $= -12/11$ which is about -1.091 .

Putting this together, we get the equation of the line to be $y = 21 - 1.091x$.

Example 7.6. Using data to find the equation of a line

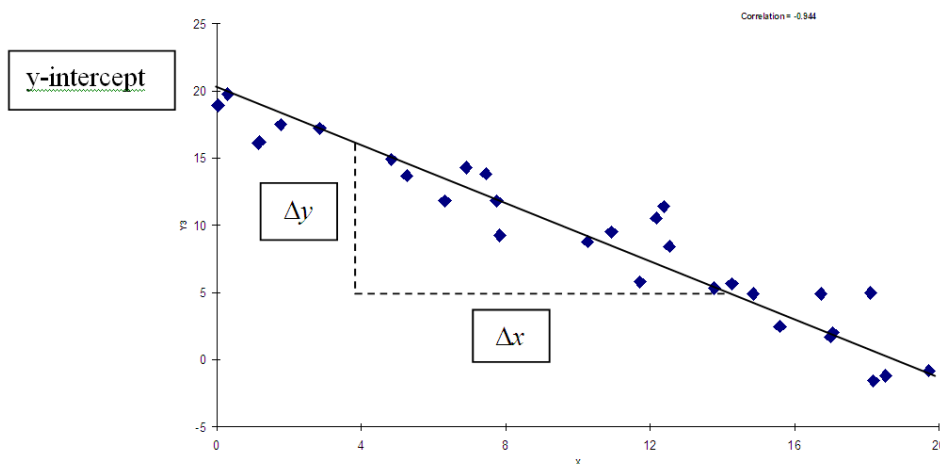


Figure 7.10: Sample linear relationship showing y-intercept and slope. In this case, we can estimate the slope. First we compute the change in y along the line (the rise) as about $5 - 17 = -12$. This is shown by the vertical dashed line. The run, or change in x , is about $14 - 4 = 10$. Thus, the slope is the ratio of these: $-12/10 = -6/5 = -1.2$. Since the line touches the vertical axis at about $y = 20$, the y -intercept is 20. Putting this together, we estimate the equation of this line as $y = -1.2x + 20$.

Suppose we have data that consists of only two points. This means that we have two ordered pairs: one for each point. The ordered pair is another way to give data. Rather than listing the variables in columns, we list the data like this: $(1, 2)$ and $(3, 6)$. These ordered pairs are given so that the first number is the value of the independent variable that is associated with the number after the comma, the dependent variable. For example, in the ordered pair $(1, 2)$, the 1 is the independent variable that gives 2 for the dependent variable. The ordered pairs listed above would be identical to the table below:

X	Y
1	2
3	6

How many straight lines are there that are a “best fit” to the data above? Do you think this will be true for any two data points? If you play around with this for a little while, you’ll discover that only one line can be drawn that passes through both points. What would the slope of this “best fit” be for the two point data set listed above? What about the y -intercept?

If we use the formulas above, the slope should be $(6-2)/(3-1) = 4/2 = 2$. This means that for every one unit we move to the right along this line, we also move two units up. Finding the y -intercept is a little trickier. Let’s use the slope-intercept form of the equation of a line. We already know the slope, so the equation must be $y = A + 2x$. To find A , just remember that we also know the point $(1, 2)$ is on the line, so $2 = A + 2(1)$. If we work with this expression, we find that $2 = A + 2$, and the only number A which works in this equation is 0, so the y -intercept must be 0. This means that the equation of the line is $y = 2x$.

Note that we could also use the point (3, 6) to find the y-intercept, A. We should get the same equation for the line using either of the two points.

Example 7.7. Calculating Values from Trendlines (Making Predictions)

In August 1997 Consumer Reports printed an article on different makes of backpacks. They measured three variables for each backpack: average price, total volume (in cubic inches), and the number of standard 5" by 7" books it could hold. A sample of the data is shown in table 7.2. (The full data set C07 Backpacks includes 30 different backpacks.)

Price	Volume	Number of Books
48	2200	59
45	1670	49
50	2200	48
42	1700	52
29	1875	52
50	1500	49
35	1950	49

Table 7.2: Data on backpacks from *Consumer Reports*.

After plotting the price of the data versus the number of books the bags hold, we get the following trendline equation (constants have been rounded to two decimal places):

$$\text{Price} = -30.68 + 1.46 * \text{Number of Books}$$

The equation tells us that we can expect the price of a backpack to increase about \$1.46 for each additional 5" x 7" book it holds. Thus, if a backpack were designed to hold 60 books, we could expect the price to be about

$$\text{Price} = -30.68 + 1.46 * (60) = \$56.92.$$

We can also ask the question another way: How many 5" x 7" books would you expect to fit into a backpack that you paid \$45 for? To deal with this question, use a little algebra (there is also a tool that can help called GOAL SEEK in spreadsheets like Excel):

$$\$45 = -30.68 + 1.46 * \text{Number of Books}$$

$$\$45 + \$30.68 = 1.46 * \text{Number of Books}$$

$$\$75.68 = 1.46 * \text{Number of Books}$$

$$\text{Number of Books} = 75.68 / 1.46 = 51.84 \text{ which is about 52 books.}$$

7.2.3 Exploration 7B: Adding Trendlines

Part I. Using data file C02 Homes, answer each question below.

1. Create a scatterplot of SIZE and PRICE, as you did in the earlier exploration in this chapter. Add a trendline to it. Sketch the trendline here.

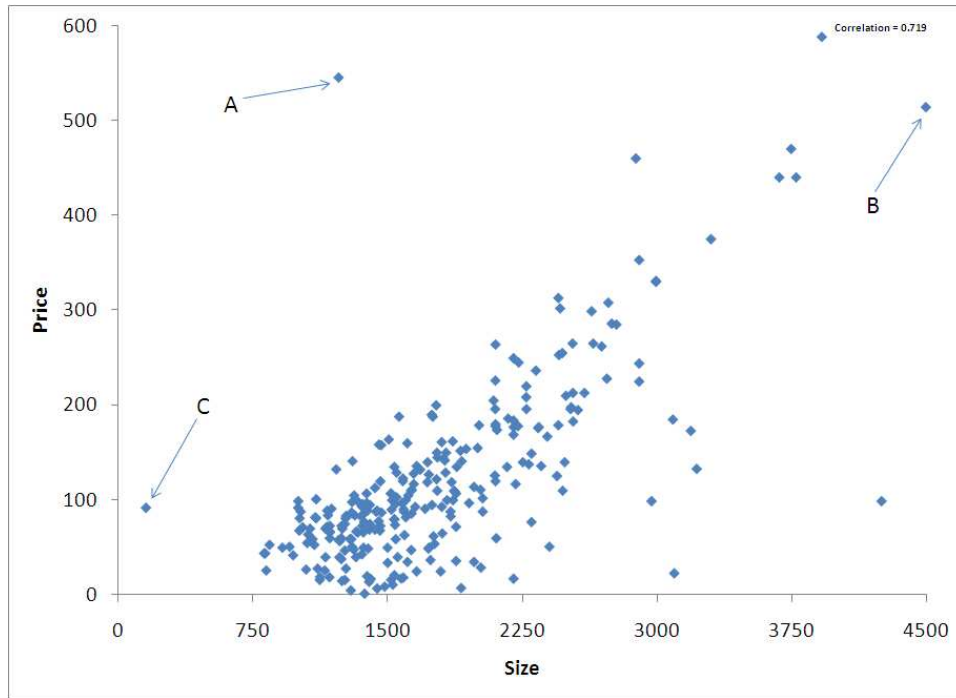


Figure 7.11: Scatterplot showing home price versus size.

2. Visually, how well does the trendline fit the data? Are there any points that seem to have undue influence on the trend or not follow the trend well?
3. What is the correlation coefficient (r) for this relationship? How does this compare with the visual evidence you considered?
4. Estimate the equation of the trendline for the relationship between the SIZE and PRICE of a home?

Part II. Create a new scatterplot between PRICE and TAXES. Be sure to do the following:

- Modify the font size
- Add units to both axis labels
- State the trendline equation in terms of the Model Variables
- Add a trendline

1. Visually, how well does the trendline fit the data? Are there any points that seem to have undue influence on the trend or not follow the trend well?
2. What is the correlation coefficient (r) for this relationship? How does this compare with the visual evidence you considered?
3. Estimate the equation of the trendline for the relationship between the PRICE and TAXES of a home?

7.3 Homework

Mechanics and Techniques Problems

7.1. Look at the data on home prices in the Rochester, NY area in 2000 found in the data file C07 Homes.

1. If you were to use this data to predict the sales price of a home, which variables would you use? Based on your intuition about homes, rank the top five most important variables in determining the price of the home in order from most influential to least influential.
2. Use the graphical and numerical tools of this chapter to determine the five variables that most influence the price of a home. Rank them in order. Compare these results with your estimates in part (a). Provide evidence for all conclusions.
3. If some of the independent variables in a data set are related to each other, you may have a problem called “co-linearity”. Are there any variables in the home data that you would expect to be related? Based on the numerical calculations (and possibly graphs) are any of the independent variables co-linear? Which ones? To what degree?

7.2. Consider the data in C07 Electricity which contains observations of total monthly electric power usage compared to the size of the home (in square feet).

1. Create a scatterplot of this data. Do you expect that a simple linear model will be a good fit to this data? Why or why not? Use the features you see in the graph to explain your answer.
2. Add a linear trendline (along with its equation) to the graph. What is the best-fit simple linear model for predicting monthly electricity usage as a function of home size? What do the slope and y-intercept mean? Do these numbers make sense? Why or why not?
3. Use the model to predict the electricity usage for the following two homes: Home #1 is 2050 square feet. Home #2 is 3200 square feet.

7.3. Suppose you have two different phone plans to select from when you make long distance calls. Plan #1 costs a flat rate of 7 cents each minute (or fraction of a minute) that the call lasts. Plan #2 costs only 3 cents per minute, but has a 39 cent connection charge for all calls, no matter how long. Which calling plan would you use for a 3 minute call? Which would you use for a 45 minute call? How can you decide ahead of time which plan to use when making a call? Explain all of your answers using trendlines and scatterplots to help. Be sure your explanation uses terms like *slope* and *y-intercept* and includes information about the units of the variables involved.

Application and Reasoning Problems

7.4. Consider two airports that are located near each other, such as the Buffalo International Airport (in Buffalo, NY) and the Rochester Airport (in Rochester, NY). Suppose you were to collect data from each airline at each airport as to what percentage of their flights arrive on time. Your data might look something like that in the data file `C07 Airports`.

1. Would you expect the two variables to be strongly or weakly correlated? Explain your answers based on an analysis of the situation, not on the actual data.
2. If you said the correlation is strong, would it be positive or negative? Explain your answer. Is this relationship causal? In other words, do more on time arrivals at one airport cause more on-time arrivals at the other airport, or is it merely a coincidence that more on-time arrivals at one airport tend to be associated with more on-time arrivals at the other airport?
3. If you said they are weakly correlated, what other variable might you measure between the two airports that would be strongly correlated?
4. How do your predictions compare with the results from the actual data?

7.5. Review the reading at the beginning of this unit, which introduces three ideas: functions, dependent variables, and independent variables.

1. Which term means the same as “explanatory variable”? Why might this be a useful description of that term?
2. Which term means the same as “response variable”? Why might this be a useful description of that term?
3. An example is given about the hypothetical relationship between the price you pay for an airline ticket and the distance being flown. Could this relationship be represented by a function (as described and defined in the reading material)? If so, what information would you need to construct the function? If not, why?

CHAPTER 8

Simple Regression¹

So far, we have encountered the idea that variables may be related to each other. Very often, we can use these relationships to determine the degree to which one variable that we have information on is related to another variable that we are interested in. To develop such relationships, we can plot the data and find an equation that relates the two variables. What we need, though, is a systematic way to decide what the best equation is to fit the data. We will start by using the simplest equations, linear models, to represent the data. The equations for the models will be developed using **least squares regression analysis**. This is a technique in which a line is assumed to exist that fits the data. By manipulating the slope and y-intercept of this line, it can be made to fit the data better. The “best fit” occurs when a certain quantity, **the total squared error**, is made as small as possible. You have already explored this concept in chapter 7 with the idea of trendlines. Most software calculates all its trendlines using least squares regression. However, we must be cautious. Regardless of whether the data appear to follow a linear pattern or not, we can pretty much always force our software to give us the equation of a best-fit line. But you need to pay attention to everything your software provides so that you do not mis-use these models. Since we can always find a “best fit” line for the data, we need some way of determining whether the linear regression equation is a good choice. To decide this, we will make use of several statistical measures and some diagnostic graphs. These will help us answer two important questions: Is the data close enough to linear to make a linear regression equation worth using? If we use the regression equation to predict information, what kind of error can we expect to have in our estimates?

- Section 8.1 is about constructing simple regression models and reading the computer output from such models.
- Section 8.2 helps you determine how well your linear model fits the data.

¹©2017 Kris H. Green and W. Allen Emerson

As a result of this chapter, students will learn

- ✓ The meaning of R^2 and S_e for regression models
- ✓ What residuals are
- ✓ What regression models are used for
- ✓ How R^2 and S_e are computed

As a result of this chapter, students will be able to

- ✓ Write down a regression model based on software output
- ✓ Use a regression model to make predictions
- ✓ Explain the meaning of a simple regression model
- ✓ Apply S_e to predictions from a model
- ✓ Use marginal analysis to interpret the slopes of a linear model

8.1 Modeling with Proportional Reasoning in Two Dimensions

At this point, we know a lot about straight lines. What we need to do now is to use this information to build models of the data. These models will allow us to predict the value of the dependent variable for different values of the independent variable. Right now, we're after models that are proportional. A proportion is simply a ratio (a fraction) between two quantities. We use the term **proportional model** interchangeably with **linear model**. These are models in which the change in the y -variable is in a fixed proportion to the change in the x -variable. What this means is that no matter what x -value you are looking at, if you increase it by a fixed amount, any fixed amount, then the change in y is fixed by the constant of proportionality. In this case, the constant of proportionality is the slope of the line.

Consider the cost of manufacturing widgets. (**Widget** is simply a word to describe something non-specific; a widget could be anything: a baseball bat, an engine part, or even a sandwich.) Normally there are fixed costs associated with manufacturing. These costs are constant, regardless of how many widgets you make. The fixed costs include things like payment for the production facilities, coverage of salaried employees, electricity, and other costs that are reasonably constant. Fixed costs look pretty much like a y -intercept on a graph. There are also variable costs in production. These include the cost of the materials to make the widgets and the wages of the employees who make the widgets. They may also include costs for quality control. Clearly, the more widgets you make, the more materials and labor you will use. It is easiest to assume that these variable costs are like the slope of a linear model, so each additional widget adds a certain amount of cost to the total manufacturing costs. Thus, we have a linear model:

$$\text{Total cost of producing widgets} = \text{Fixed Costs} + (\text{Variable Costs}) * (\text{Number of Widgets})$$

(There are certainly other ways of modeling cost, but this is the easiest to understand, so it makes a nice starting point.) Suppose your fixed costs are \$1,000 and the variable costs are \$3.50 per widget. If you make 10 widgets, it will cost you

$$\$1,000 + \$3.50 \text{ per widget} * 10 \text{ widgets} = \$1,000 + \$35 = \$1,035.$$

If you make five more widgets, for a total of 15 widgets, it will cost $\$1,000 + \$3.50 (15) = \$1,052.50$, exactly \$17.50 more. Notice that $\$17.50 = \$3.50 * 5$. In a proportional model, no matter what the current production level is (how many widgets you are making), the model always predicts the same change in y for a fixed change in x . Such models are sometimes called **level independent models**.

What this really means is that whether you are making 10 widgets or 20,000 widgets, if you make 5 more widgets it will cost an additional \$17.50. This is the reason that linear models are so useful; they are easy to interpret. Each coefficient in the equation of the model has a meaning that is easily understood in terms of the problem context. Economists and business folks often refer to marginal costs, the additional cost you will pay in order to make one more widget. For a linear model, the marginal cost is simply the slope.

For more information, see the interactive Excel workbook C08 **StepByStep**.

8.1.1 Definitions and Formulas

Linear model A linear model is one where the variables are related by a linear equation, like $y = A + Bx$. Such models are also referred to as level-independent or proportional. Level-independent refers to the fact that the slope is the same, regardless of the value of x (the level). Referring to them as proportional models emphasizes the fact that changes in y are related proportionally to changes in x .

Simple linear regression This is a process for systematically determining the equation of the best-fit line to a given set of (x, y) data. The regression equation is determined by a process called least squares regression and results in a formula to compute the slope and y-intercept of the line that will minimize the “total squared error” of the line. Based on some theoretical calculations with calculus, you can show that the slope, B , of a regression line is given by

$$B = \text{Corr}(X, Y) \frac{\sigma_y}{\sigma_x}$$

where $\text{corr}(X, Y)$ represents the correlation of the variable X with Y and the σ represent the standard deviations of the X and Y variables. Once you have the slope, the y -intercept is easy to find: $A = \bar{Y} - B\bar{X}$, where \bar{X} and \bar{Y} are the means of the X and Y variables.

Proportional Two quantities are proportional when a specific amount of change in one of the quantities results in a certain amount of change in the other quantity given by a fixed multiplicative factor. In mathematical terms, the phrase “the change in y is proportional to the change in x ” can be written as $\Delta y \propto \Delta x$. This means that $\Delta y = k\Delta x$ for some constant k that is independent of y and x .

Coefficient A coefficient is a fixed (or constant) number in an algebraic model. For example, linear equations have two coefficients: the slope and y-intercept. Coefficients are sometimes called **constants** or **parameters**. In regression output, a table of coefficients is produced. It is up to you to combine these correctly into the model equation. In the examples below, you will see how to do this.

Constant In most regression output, the y -intercept of the regression line is labeled “constant” in the table of coefficients. More generally, a constant is any value in a formula that is fixed, like the number 2 in the linear relationship $y = 2x + 5$. Sometimes constants are called parameters.

Explanatory Variable This is the variable (or variables, in later chapters) used to explain the results of the model. In simple regression, the x -variable is the explanatory variable. As you can guess, this is just another name for the independent variable.

Response Variable This is the variable that responds to the independent (or explanatory) variable. Thus, it is really the y -variable or dependent variable.

8.1.2 Worked Examples

Example 8.1. Translating Regression Output Into an Equation

Regression output from most simple regression routines will look like the results below. This regression output comes from the data on backpacks in table 7.2 from the last chapter. The data is in **C07 Backpacks**. Notice that the output is divided into three areas by headings in bold: summary measures, ANOVA table, and regression coefficients. For now, we are concerned mostly with the first few columns under the heading “Regression Coefficients.” In the next section we’ll come to understand the summary measures (which explain how good and accurate the model is) and a little of the ANOVA table (ANOVA stands for Analysis Of Variance; it is used for computing the summary measures.)

Results of simple regression for Price						
Summary measures						
Multiple R	0.7022					
R-Square	0.4931					
StErr of Est	9.8456					
ANOVA table						
Source	df	SS	MS	F	p-value	
Explained	1	2640.4857	2640.4857	27.2397	0.0000	
Unexplained	28	2714.1810	96.9350			
Regression coefficients						
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-30.6751	13.5969	-2.2560	0.0321	-58.5272	-2.8231
Number of Books	1.4553	0.2788	5.2192	0.0000	0.8842	2.0265

From this output, we can easily write down the linear equation that best estimates the price of the backpack, based on the explanatory variable “Number of Books”. We know that “Price” is the response variable from the first line of the regression output; it says “Results of simple regression for Price”. The dependent variable will always be given here. To write down the regression equation, we need to know only slope and y -intercept.

The y -intercept is the “Coefficient” next to “Constant” in the regression coefficients portion of the output. Thus, this model has a y -intercept of -30.6751. Since the Price variable is in dollars and the y -intercept will have the same units, it probably makes sense to round this to -30.67. The slope of the regression model is the coefficient next to the explanatory variable, in this case “Number of books”. So the slope here is 1.4553. Since number of books is typically between 10 and 60, we may want to round off to three decimal places (1.455), so that after multiplication by a number of books the result is a dollar amount.

The final regression equation is then

$$\text{Price (of backpack in dollars)} = -30.67 + 1.455 \cdot \text{Number of Books}.$$

Example 8.2. Interpreting Coefficients of a Regression Model

In the previous example, we developed a regression model from the regression output displayed below. But what does this model mean?

The y -intercept is usually pretty clear: it's the y -value when the x -variable is zero. So, if we were to market a backpack that couldn't hold any books (Number of Books = 0), we could expect that people would not pay any money for it. In fact, the equation predicts that we would have to pay the customer \$30.67 just to take the backpack away! After all, a backpack that doesn't hold any books isn't very useful. We'll get another interpretation of this number in the next example.

To interpret the slope, we need to use some proportional reasoning. Remember that slope is rise over run. "Run" in this case refers to the number of books the backpack will hold, while "rise" refers to the price of the backpack. Our model has slope = rise/run = (change in y)/(change in x) = 1.455. This is marginal analysis: to determine what happens to the value of the dependent variable when x changes by 1 unit. If we design an identical backpack, that can hold one more book, then the price of the backpack will increase by \$1.455. In fact, since linear models are proportional, a 2 unit increase in x will result in a $2 \times \$1.455 = \2.910 increase in the price.

Another way to see this is to analyze the units of the slope. Since slope is change in y over change in x , it must have the units " y units per x unit". Thus, our slope really means "1.455 dollars per book". For each additional book a backpack can hold, this model predicts a \$1.455 increase in the price of the backpack.

Example 8.3. Calculating the X -intercept (Solving a linear equation)

Now, the previous examples have a y -intercept that doesn't make much sense. After all, who would market a backpack that you have to pay people to use? But let's graph the equation of the model and see if it helps. We know the y -intercept is -30.67, so the line passes through the point (0, -30.67). It has a positive slope (1.455) so it is increasing; this means that the more books the backpack holds, the more it is worth.

Notice that the line passes through the x -axis. Thus, it has an x -intercept. In this case, the x -intercept appears to be around Number of Books = 20, but it's a little higher than that. Finding the x -intercept will answer the following question: What would be the least number of books the backpack would have to hold in order to be marketable? Let's try to find the x -intercept exactly.

To do this, we take the regression equation, and we plug in everything we know is true about the x -intercept. Right now, we have a guess at its x coordinate, but that's not good enough. We do know one thing for certain, though: it has a y -coordinate of 0. So we can plug in "Price = 0" to our regression equation above to get

$$0 = -30.67 + 1.455 \times \text{Number of Books}.$$

Now, we want to use algebra to rearrange the equation to find how many books it takes to make the price zero (that's what the above equation really means). To solve the equation, we simply "undo" what has been done to the number of books. First, the number of books is multiplied by 1.455, then that result is added to -30.67. So, to undo it, we first subtract



Figure 8.1: Example of a straight line predicting the price of a backpack based on the number of books it holds.

(-30.67) and then divide by 1.455. But if we do this to the right-hand side of the equation, we must do it to the left-hand side. Zero minus (-30.67) is just 30.67. Dividing this by 1.455, we get approximately 21.079. These steps are shown below.

$$\begin{aligned}
 0 + 30.67 &= -30.67 + 1.455 * \text{Books} + 30.67 \\
 30.67 &= 1.455 * \text{Books} \\
 \frac{30.67}{1.455} &= \frac{1.455 * \text{Books}}{1.455} \\
 21.079 &= \text{Books}
 \end{aligned}$$

This means that unless your backpack can hold at least 22 books (the nearest whole number greater than your x intercept of 21.079, since we cannot really have a part of a book), you cannot expect anyone to buy it.

8.1.3 Exploration 8A: Regression Modeling Practice

StateSteins is a tourist trade vendor. The company manufactures hand-crafted beer steins with state logos and images. They have facilities in each of the fifty states (and in Washington, DC). The data file `C08 Profit` contains last year's figures for profit, revenue, cost, number of steins sold, and labor for each of the separate state facilities. Your job is to investigate these data and determine which variable is the best predictor for the company's profits.

1. Formulate and estimate linear regression models to predict profit as a function of each of the four explanatory variables. Be sure that you have the routine construct the important diagnostic graphs (Fitted v. Actual, Residuals v. Fitted) to help in your analysis. (If you are using StatPro/Excel, you will need to rename the worksheets with these graphs in order to ensure that StatPro will not overwrite them each time you compute a new regression model.)
2. Interpret the slopes of each of the four regression models you created. Be sure to include the units for each slope.
3. Examine the diagnostic graphs to see what they tell you about the quality of your model. You will learn more about how to interpret these graphs in the next section, but for now, see what you can learn from them.

8.2 Using and Comparing the Usefulness of a Proportional Model

Now, it's one thing to have an equation to model data. We can always get regression equations for any data. However, the “best fit line” may not be a very good model for the data. We need a way to know not only the equation of the model, but also how good the model is. We will learn about two ways to measure “how good a model is”. The first is a direct test for whether the two variables in the model are even linearly related. This is called the coefficient of determination (R^2) and is related to the correlation between the two variables. The second measure tells us how close predictions from our model will be to the actual data. This number is called standard error of estimate (S_e) and is sort of a standard deviation, indicating how spread out the data is from the model.

These two quantities relate to the entire regression model, reducing some characteristic “error” in the model down to single numbers. There are other ways to check on the quality of the regression model, however. Most statistical packages provide diagnostic graphs for checking the regression model out. Two of the most important of these graphs are the graphs of the predicted values (also called fitted values) versus the actual response variable data and the graph of the residuals (the error in the model) versus the fitted values. A quick look at these two scatterplots can often tell you a lot about the quality of the model. Taken together with the coefficient of determination and the standard error of estimate, these are very powerful tools for determining the quality of the regression models you produce. After all, it is easy to simply point and click to produce more and more regression models; what is difficult is learning which ones are useful and to what extent they are useful.

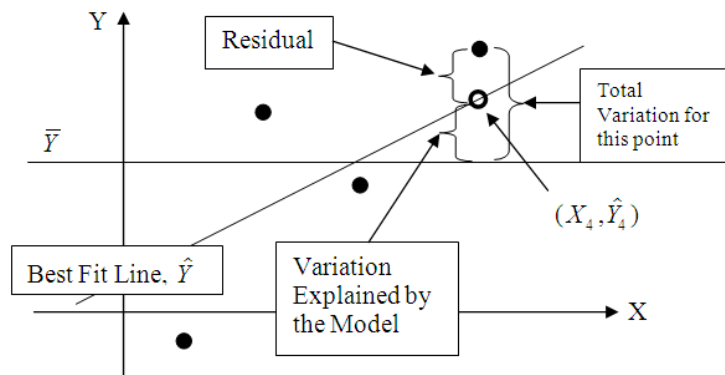


Figure 8.2: The various quantities involved in regression that are explained below.

8.2.1 Definitions and Formulas

Predicted values (fitted values) These are the predictions of the y-data from using the model equation and the values of the explanatory variables. They are denoted by the symbol \hat{y}_i .

Observed values These are the actual y -values from the data. They are denoted by the symbol y_i .

Residuals This is the part that is left over after you use the explanatory variables to predict the y -variable. Each observation has a residual that is not explained by the model equation. Residuals are denoted by e_i and are computed by

$$e_i = y_i - \hat{y}_i$$

Since these are computed from the y values, it should be clear that the residuals have the same units as the y , or response, variable.

Total Variation (Total Sum of Squares, SST) The total variation in a variable is the sum of the squares of the deviations from the mean. Thus, the total variation in y is

$$\text{SST} = \sum (y_i - \bar{y})^2$$

Unexplained variation (Sum of Squares of Residuals, SSR) The variation in y that is unexplained is the sum of the squares of the residuals:

$$\text{SSR} = \sum (y_i - \hat{y}_i)^2$$

Explained variation (Sum of Squares Explained, SSE) The total variation in y is composed of two parts: the part that can be explained by the model, and the part that cannot be explained by the model. The amount of variation that is explained is

$$\text{SSE} = \text{Total Variation} - \text{Unexplained Variation} = \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2$$

Regression Identity One will note that the Total Variation is equal to the sum of the Unexplained Variation and the Explained Variation.

$$\text{SST} = \text{SSR} + \text{SSE}$$

Coefficient of Determination (R^2) This is a measure of the “goodness of fit” for a regression equation. It is also referred to as R-squared (R^2) and for simple regression models it is the square of the correlation between the x - and y -variables. R^2 is really the percentage of the total variation in the y -variable that is explained by the x -variable. You can compute R^2 yourself with the formula

$$\begin{aligned} R^2 &= \frac{\text{Total Variation} - \text{Sum of Squares of Residuals}}{\text{Total Variation}} \\ R^2 &= \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \\ &= \frac{\text{SST} - \text{SSR}}{\text{SST}} = \frac{\text{SSE}}{\text{SST}} \end{aligned}$$

R^2 is always a number between 0 and 1. The closer to 1 the number is, the more confident you can be that the data really does follow a linear pattern. For data that falls exactly on a straight line, the residuals are all zero, so you are left with $R^2 = 1$.

Degrees of Freedom for a linear model The degrees of freedom for any calculation are the number of data points left over after you account for the fact that you are estimating certain quantities of the population based on the sample data. You start with one degree of freedom for each observation. Then you lose one for each population parameter you estimate. Thus, in the sample standard deviation, one degree of freedom is lost for estimating the mean. This leaves you with $n - 1$. For a linear model, we estimate the slope and y -intercept, so we lose two degrees of freedom, leaving $n - 2$.

Standard Error of Estimate (S_e) This is a measure of the accuracy of the model for making predictions. Essentially, it is the standard deviation of the residuals, except that there are two population parameters estimated in the model (the slope and y -intercept of the regression equation), so the number of degrees of freedom is $n - 2$, rather than the normal $n - 1$ for standard deviation.

$$S_e = \sqrt{\frac{\sum e_i^2}{n - 2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{SSR}{n - 2}}$$

The standard error of estimate can be interpreted as a standard deviation. This means that roughly 68% of the predictions will fall within one S_e of the actual data, 95% within two, and 99.7% within three. And since the standard error is basically the standard deviation of the residuals, it has the same units as the residuals, which are the same as the units of the response variable, y .

Fitted values vs. Actual values diagnostic graph This is one of the most useful of the diagnostic graphs that most statistical packages produce when you perform regression. This graph plots the points (y_i, \hat{y}_i) . If the model is perfect ($R^2 = 1$) then you will have $y_1 = \hat{y}_1$, $y_2 = \hat{y}_2$, and so on, so that the graph will be a set of points on a perfectly straight line with a slope of 1 and a y -intercept of 0. The further the points on the fitted vs. actual graph are from a slope of 1, the worse the model is and the lower the value of R^2 for the model.

Residuals vs. Fitted values diagnostic graph This graph is also useful in determining the quality of the model. It is a scatterplot of the points $(\hat{y}_i, e_i) = (\hat{y}_i, \hat{y}_i - y_i)$ and shows the errors (the residuals) in the model graphed against the predicted values. For a good model, this graph should show a random scattering of points that is normally distributed around zero. If you draw horizontal lines indicating one standard error from zero, two standard errors from zero and so forth, you should be able to get roughly 68% of the points in the first group, 95% in the first two groups, and so forth.

8.2.2 Worked Examples

Example 8.4. Interpreting the quality of a model

In the examples from section 8.1.2, we developed and explored a model for predicting the price of a backpack based on the number of 5" by 7" books that it can hold inside. Using the data graphed in figure 8.3, we can produce the regression output below.

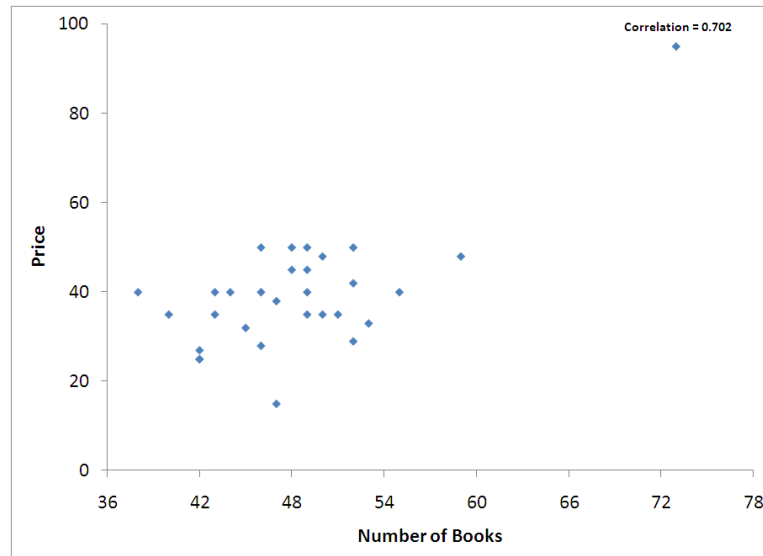


Figure 8.3: The various quantities involved in regression for the price (in dollars) of a backpack modeled by the number of books it can hold. Note that these variables have a correlation of 0.702, indicating a strong relationship.

Results of simple regression for Price

Summary measures

R-Square	0.4931
StErr of Est	9.8456

ANOVA table

Source	df	SS	MS	F	p-value
Explained	1	2640.4857	2640.4857	27.2397	0.0000
Unexplained	28	2714.1810	96.9350		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-30.6751	13.5969	-2.2560	0.0321	-58.5272	-2.8231
Number of Books	1.4553	0.2788	5.2192	0.0000	0.8842	2.0265

Now we want to ask the question: “How good is this model at predicting the price of a backpack?” We’ll start by examining the summary measures section of the output. Spoiler Alert! The model is not very good.

The first clue to the model’s poor quality is the R^2 value: It’s not terrible, but it is pretty low, being only 0.4931. This means that the number of books can only explain 49.31% of the total variation in price. That leaves over 50% of the variation in price unexplained. This could be for many reasons:

1. Some of the data could be entered incorrectly. To correct this, we could look for outliers in the data and try double-checking all the data entry.
2. The number of books may not be a good predictor for the price of a backpack. We could try making simple regression models using the other explanatory variables to see if one of them does a better job.
3. There could be additional variables that should be included in order to determine the price. We will explore multiple regression models in chapter 9.
4. It may be that a linear model is not the best choice for modeling the price of a backpack as a function of the number of books it can carry. In that case, we could use the ideas in chapter 11 to build some nonlinear regression models.

Many times, the real reason for a low R^2 value is either #3 or #4. For example, our data and model do not include any variables to quantify the style of the backpack or its comfort when wearing it. Perhaps durability or materials are important variables. Maybe the name brand is important. Perhaps certain colors sell better. Perhaps certain extra features, additional pockets or straps for keys, are desired. Our data ignores these features. Essentially, our data tries to make all backpacks that are the same size cost the same amount of money. Clearly, this is not realistic. Given all these problems, though, an R^2 of 49% is not too bad. So, we might be able to convince ourselves that the model is useful.

The standard error of estimate, S_e , for this model is about \$9.85. We can interpret this to mean that our equation will predict prices of backpacks to within \$9.85 about 68% of the time, or to within $2 \times \$9.85 = \19.70 about 95% of the time. That wide range, from \$9.85 to \$19.70, in the predicted prices is probably due to other variables that are important (see point #3 above). Notice that S_e is always measured in the same units as the y -variable. This means that there is no “hard and fast” rule for what constitutes a good S_e . Our advice is to always compare S_e with the standard deviation of the response variable, since this tells us how accurate our simplest model, the mean, is for the data and we want to do better than that with our regression model. For these data, the standard deviation of Price is \$13.59. This means that if we just used the average price (\$39.67) as our model for backpack price, we would be less accurate than if we used the regression equation which at least accounts for the size of the backpack. In general, a good model will have S_e much less than the standard deviation of the y -variable.

Example 8.5. Computing R^2 from the ANOVA Table

What does the ANOVA table tell us? It actually tells us quite a lot, but for this text, we

will only examine the first two columns of the ANOVA table. These are marked *df* and *SS*. These stand for **degrees of freedom** and **sum of squares**. Notice that the degrees of freedom that are explained is 1. This is the number of explanatory variables used in the model. *Degrees of freedom that are unexplained* is the number of observations, *n*, minus the explained degrees of freedom, minus one more for the y-intercept. Thus,

$$\text{Df (explained)} + \text{Df (unexplained)} + 1 = n$$

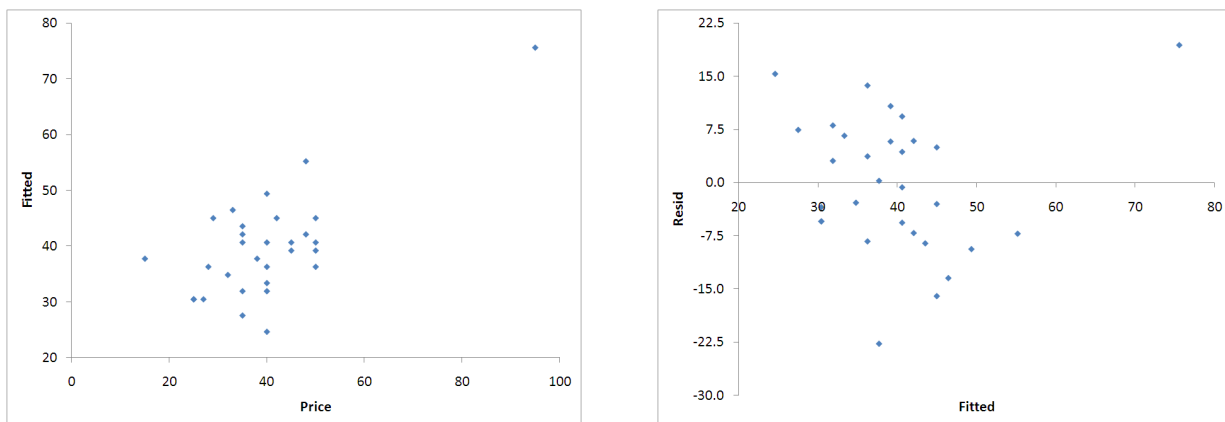
The sum of squares tells you how much of the total variation is explained and how much is unexplained. In this model, the amount of variation explained by “number of books” is given by SSE which is 2640.4857. The unexplained variation, SSR, is 2714.1810. This means that the total variation (SST) in *y* is $2640.4857 + 2714.1810 = 5354.6667$. Thus, to calculate R^2 , the fraction of the total variation in *y* that is explained by *x*, we simply compute a ratio:

$$R^2 = \frac{2640.4857}{2640.4857 + 2714.1810} = \frac{2640.4857}{5354.6667} \approx 0.4931.$$

Note that R^2 is also given by $(\text{SST} - \text{SSR})/\text{SST}$. And since $\text{SST} = \text{SSE} + \text{SSR}$, we can rewrite this as $((\text{SSR} + \text{SSE}) - \text{SSR})/(\text{SSR} + \text{SSE}) = \text{SSE}/(\text{SSR} + \text{SSE})$. So you have several ways to estimate R^2 from the ANOVA table.

Example 8.6. Reading the Diagnostic Graphs

Let’s examine the diagnostic graphs for the backpack model. There are two of these that are important. The first is the “Fitted versus Price”. This graph is a scatterplot of all the fitted or predicted data (\hat{y}_i) versus the corresponding actual data (y_i). If the model predicted 100% of the variation in backpack prices ($R^2 = 1$) then each predicted value would equal the corresponding actual value, and the scatterplot would show a perfectly straight line of points with a slope of 1 and y-intercept of 0. The further the scatterplot is from such a line, the worse a fit the regression model is for the data. In this case, we get an interesting graph. The graphs in figure 8.4 show a rough trend like this, but there is a lot of spread around the “perfect line”.



(a) Fitted vs. Actual diagnostic graph (b) Residuals vs. Fitted Values diagnostic graph

Figure 8.4: Diagnostic graphs for the backpack example.

RULE OF THUMB: When looking at the graph of fitted values versus actual values, there is no perfect explanation for how close to a straight line the graph needs to appear for the model to be “good.” Notice the graph shown on the left in figure 8.4 does not really look much like a straight line. In fact, if it were not for the lone data point far to the upper right, it might almost not look linear at all. The most important thing to do is not to make any absolute judgements about whether the model is good or bad; instead, focus on using the graph to explain potential problems with the model. Use the graph to describe the model’s features. For example, in the graph above, we see that the model is not very accurate - for the most part, the data is randomly spread around the model results, with only one point really making the model behave linearly. This one data point’s effects are called *leveraging* and this will be addressed in the exploration.

The other diagnostic graph that is important is the graph of the residuals versus the fitted values. For this graph, we want to see the points randomly splattered around with no pattern. If there is a pattern to the points on this graph, we may need to try another kind of regression model, some sort of nonlinear equation. This will all be explained in more detail in later chapters, but for now, you should at least look at these graphs and try to determine whether they indicate that the model is a decent fit or not. We want the following to hold:

1. The fitted vs. actual graph should be close to a straight line with a slope of 1 and y-intercept of 0. This is because a perfect model would exactly predict the values of the response variable. It is usually sufficient to eyeball this, rather than clutter the graph by adding trendlines.
2. The residual plot should look like a bunch of randomly scattered points with no pattern. If there is a pattern, then we can alter the model to predict the pattern, producing a better model.

8.2.3 Exploration 8B: How Outliers Influence Regression

For this exploration we are going to investigate the relationship between the appraised values of homes and the actual sale prices of the homes. (All of the homes in the data file **C08 Homes** were sold during a three-month period of time in the Rochester, NY region in 2000.)

First, construct a linear regression model for predicting the price of the home from the appraised value. Be sure that you have the routine construct the diagnostic graphs (Fitted vs. Actual and Residuals vs. Fitted). Also, make sure that you have the routine calculate the fitted values and residuals on the data worksheet. You will need all of these graphs and figures to explore the data.

1. What is the equation of your regression model? Is this model any good?
2. What does this model mean?
3. Now, on the residuals vs. fitted graph, draw horizontal lines to mark one standard error of estimate above and one below the horizontal axis (residuals = 0 along the axis). Draw similar lines to mark two, three and four standard errors. How many of the observations fall within 1, 2, 3, and 4 standard errors of the predicted values? What proportion of the observations fall in these ranges? (There are 275 observations total.)
4. Are there any outliers in the data - observations more than 4 standard errors from zero? How many are there? How do you think these outliers influence the quality of the regression model? How would the regression model change if you removed these outliers and re-ran the regression routine?
5. Next, we are going to identify the outliers and remove them from the data. To do this, we need to look at the actual data and sort it from smallest to largest residuals. Before we can do this, however, we need to delete the empty column between the data and the fitted values (this should appear in column N). To delete the column, place your cursor on the column header (N), right click, and select "Delete". Now, sort the data from smallest to largest residuals. Locate any observations with residuals more than 4 standard errors from zero. Delete these observations from the data by deleting the rows the data are in (right click on the row, select "Delete"). Now, create a new regression model to predict the price of a home from its appraised value. What is the equation of this model?
6. Compare the two models, both their equations and their quality.

8.3 Homework

Mechanics and Techniques Problems

8.1. Suppose you know statistics for X and Y shown below. You also know that the correlation of X to Y is 0.56. Use these to determine the equations of the least-squares best fit regression model to predict Y as a function of X. Produce a graph of this regression equation. Show all work.

Statistic	X-Variable	Y-Variable
Mean	15.27	107.93
Standard Deviation	7.82	38.77
First Quartile	5.3	47.1
Median	15.2	105.4
Third Quartile	22.6	160.3

8.2. The regression output below was developed from data relating the monthly usage of electricity (MonthlyUsage, measured in kilowatt-hours) to the size of homes (HomeSize, measured in square feet). One-variable statistics for each of these variables is also given below.

1. Use this information to write down the equation of the regression model.
2. Explain what each part of the regression model means, paying particular attention to the unit of the coefficients in the regression equation.
3. Analyze the quality of the regression model you wrote down, based on the summary statistics in the regression output and the statistics on the X and Y variables.
4. Based on your regression model, what is the relationship between home size and monthly usage? Does this seem realistic? (Hint: What does the model predict for bigger and bigger homes? What about smaller homes? Are there any homes for which the model predicts a monthly usage of zero?)

Results of simple regression for Monthly Usage						
Summary measures						
Multiple R	0.9120					
R-Square	0.8317					
StErr of Est	133.4377					
ANOVA Table						
Source	df	SS	MS	F	p-value	
Explained	1	703957.1781	703957.1781	39.5357	0.0002	
Unexplained	8	142444.9219	17805.6152			
Regression coefficients						
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	578.9277	166.9681	3.4673	0.0085	193.8984	963.9570
HomeSize	0.5403	0.0859	6.2877	0.0002	0.3421	0.7385

Summary measures for selected variables		
	HomeSize	MonthlyUsage
Mean	1880.000	1594.700
Median	1775.000	1641.000
Standard deviation	517.623	306.667
Minimum	1290.000	1172.000
Maximum	2930.000	1956.000
Variance	267933.333	94044.678
First quartile	1502.500	1321.250
Third quartile	2167.500	1831.000
Interquartile range	665.000	509.750
Skewness	0.893	-0.308
Kurtosis	0.340	-1.565

8.3. Pie in the Sky, Inc. runs a chain of pizza eateries (See data file C08 Pizza.) The manager has collected data from each of the stores in the chain regarding the number of pizzas sold in one month, the average price of the pizzas, the amount the store spent on advertising that month, and the average disposable income of families in the area near the store.

1. The manager wants to know how these variables are related. Specifically, he wants to know which variable is the best to use for predicting the number of pizzas that a given store will sell in a month. Develop regression models to predict the quantity sold based on each variable in the data. Use the three models you develop to determine which variable is the most influential.

2. Use your best model to determine how many pizzas will be sold if a store has an average pizza cost of \$11.00, spends \$51,000 on advertising, and is in a region with an average disposable income of \$40,000.
3. Based on the models that you developed, if a store wanted to sell 80,000 pizzas, what should the store do?

Application and Reasoning Problems

8.4. The file `C08 Hospitals` contains hospital and physician data on a number of metropolitan areas. Use this data to develop a regression model for predicting the number of general hospitals that a metropolitan region can support based on the number of general physicians in the region. Once you have the model, use the residuals and the summary measures to identify the regions which are outliers (for these purposes, an outlier is more than four standard errors away from the fitted values). You should find a total of seven outliers in the data (including both too high and too low).

Multiple and Categorical Regression¹

In this chapter, we will explore relationships that are more realistic: one variable will be dependent on several variables. This is the most common scenario in analyzing data. Consider the salary of an employee at a company. Most likely, that salary is based on a combination of factors: educational background, prior experience in a related job, job level in the company, and number of years with the company, just to name a few. Trying to separate any one of these variables out to explain salary will result in a large amount of variation in the model. This is because there are probably several employees with the same educational background (like a Bachelor's degree) but different experience. They will make different salaries. If you try to predict salary based only on education, the model will have a great deal of error caused by this spread in the data. Essentially, the problem is caused by trying to account for too much variation in salary with too few variables.

- In section 9.1, we will use multiple linear regression to model relationships in which a single response quantity is dependent on several explanatory variables at one time. Multiple regression works pretty much like simple linear regression, but has more information (more slopes to deal with) and another measure of validity, called the adjusted R^2 .
- Section 9.2 takes us back to looking at categorical data. Up till now, we've created models using only numerical variables. Many of the data sets that we are interested in, however, include categorical data. In the past, to analyze such data, we have been forced to “unstack” the data and make several graphs. One can certainly continue in this fashion, but if there are several different categorical variables of interest, the process would be time-consuming. As it happens, there is an agreed-upon method for converting categorical data into numerical data by introducing dummy variables. You will learn how to create dummy variables and how to build and interpret regression models built from them. By the end of the chapter, you will have a powerful collection

¹©2017 Kris H. Green and W. Allen Emerson

of tools for modeling data. You will be able to represent relationships with several variables, using numerical, categorical, or a combination of variable types.

As a result of this chapter, students will learn *As a result of this chapter, students will be able to*

- | | |
|--|---|
| <ul style="list-style-type: none"> ✓ How to read the measures of validity for multiple regression output ✓ What the coefficients in a multiple regression output mean ✓ How graphs can help interpret the validity of a multiple regression model ✓ How multiple regression can handle more complex problems than simple regression ✓ What dummy variables are ✓ What a reference category is ✓ How many equations are really hidden inside a single model with dummy variables | <ul style="list-style-type: none"> ✓ Set up a multiple linear regression model ✓ Write down the regression equations for a multiple regression model ✓ Analyze the accuracy of a multiple regression model ✓ Make predictions, using , from a model ✓ Determine appropriate variables to use, based on the adjusted R^2 value ✓ Create dummy variables for a set of data u(if needed, to construct models with categorical factors) ✓ Construct a model using dummy variables ✓ Identify the reference category in a model ✓ Interpret a model with dummy variables, including all the “hidden equations” |
|--|---|

9.1 Modeling with Proportional Reasoning in Many Dimensions

So far we have used only a single explanatory variable to describe the variation in our response variables. However, in real world data, there are usually complicated relationships involving many different variables. Consider the price of a home, for example. It depends on the size of the home, the condition of the home, the location of the home, the number of bedrooms, the number of bathrooms, the presence of any amenities, and many other “less tangible” qualities. If you were to use any single one of these to predict the price of the home, the model would have a very low coefficient of determination and a very high standard error (S_e) because the other variables are being ignored. In essence, a single-variable model for data like this tries to make all of the “left out” variables the same. If we choose the size (in square feet) to predict price, we are basically saying that all houses that have the same number of square feet must also have the same number of bedrooms, the same number of bathrooms, the same location, the same condition, and the same amenities. Clearly this is not the case. This means that the variation in price caused by these “left out” variables will result in a lot of spread in the observed values around the regression line.

This problem is actually related to another issue with complex data. If you want to graph the data, each variable in the problem requires a separate dimension. One explanatory variable and one response variable requires two dimensions to graph (a plane). Two explanatory variables and one response require three dimensions to graph (space). Anything more requires more dimensions that we can represent on paper or with a physical, hands-on model. Thus, as we try to build models that incorporate more variables, we lose one of our main tools for picturing the data: scatterplots. Without a scatterplot of the actual data (Y vs. all the X variables) we cannot use software to make a trendline. The only way to get the model equation is to use multiple regression.

Multiple regression produces longer, more complicated looking equations as models of the data. However, they are not more difficult to interpret than simple regression models. Suppose we use data on houses to produce a regression model that looks like

$$\begin{aligned} \text{Price (thousands)} = & 18 - 1 * \text{Age} + 27 * \text{Number of Baths} - 9 * \text{Number of Bedrooms} \\ & - 5 * \text{Number of rooms} + 0.5 * \text{Number of Acres} + 0.09 * \text{Square Footage}. \end{aligned}$$

This model shows how each variable influences the price of the home when all of the other variables are controlled for. This is another way of saying that hold all the other explanatory variables constant. Notice, however, that since each variable has different units, the coefficients do not tell us which variables are most important. Each full bathroom in the home adds \$27,000 (remember, the coefficient is in thousands of dollars, so 27 means \$27,000) to the estimated sale price, but each square foot only adds \$90. This does not mean that bathrooms are more important than size, though. In fact, an additional 300 square feet (a 15' by 20' room) also adds exactly \$27,000 to the estimated price. Without looking at the units on each coefficient, you cannot say which are more important. In this section, you will learn how to build and interpret multiple regression models like this one.

9.1.1 Definitions and Formulas

Multiple linear function This is a model much like a simple linear model, except that it includes several explanatory variables. If the explanatory variables are labeled X_1, X_2, \dots and the response variable is Y , then a multiple-linear model for predicting Y would take the form

$$Y = A + B_1X_1 + B_2X_2 + \dots + B_NX_N$$

Notice that the multiple linear function has a “ y -intercept” given by A . Each of the coefficients (the B_i ’s) is a slope associated with one of the explanatory variables.

An important difference between simple linear and multiple linear models is the graphical illustration of each. A linear function describes a line in two dimensions. A multiple linear function with two explanatory variables describes a plane in three-dimensional space. If there are more than two explanatory variables, we cannot picture the “hyperplane” that the function describes.

Multiple linear regression The process by which you can “least squares fit” a multiple linear function to a set of data with several explanatory variables.

Stepwise regression This is an automated process for determining the best model for a response variable, based on a given set of possible explanatory variables. The procedure involves systematically adding the explanatory variables, one at a time, in the order of most influence. For each variable, a p -value is determined. The user controls a cut-off for the p -values so that any variable with a p -value above the cut-off gets left out of the model.

Controlling variables This is the process by which the person modeling the data tries to account for data which may have several observations that are similar in some variables, but differ in others. For example, in predicting salaries based on education, you should control for experience, otherwise the model will not be very accurate, since several employees may have the same education, but different salaries because they have different experience.

Degrees of Freedom for Multiple Regression Models In multiple regression models, one is usually estimating several characteristics of the population that underlies the data. For each of these estimated characteristics, one degree of freedom is lost. If there are n observations, and you are estimating a multiple regression model with p explanatory variables, then you lose $p + 1$ degrees of freedom. (The “+1” is for the y -intercept.) Thus,

$$\begin{aligned} Df &= n - (p + 1) \\ &= n - p - 1 \quad [\text{Removing parentheses}] \end{aligned}$$

Also notice that in the ANOVA table for multiple regression, the degrees of freedom of the Explained $(p - 1)$ plus the degrees of freedom of the Unexplained $(n - p)$ add up to the degrees of freedom of the sum of the squares of the total variation $(n - 1)$:

$$\begin{aligned} n - 1 &= (p - 1) + (n - p) \\ SST &= SSR + SSE \end{aligned}$$

(Total Variation = Sum of Squares of Unexplained + Sum of Squares of Explained)

Multiple R^2 This is the coefficient of multiple determination used to determine the quality of multiple regression models between the responses y_i and the fitted values \hat{y}_i .

$$\text{Multiple } R^2 = \frac{SSE}{SST} = \frac{SST - SSR}{SST} = 1 - \frac{SSR}{SST}$$

SSR=Sum of the squares of the residuals (unexplained variation)

SSE=Explained amount of variation

SST=Total variation in y

A large R^2 does not necessarily imply that the fitted model is a useful one. There may not be a sufficient enough number of observations for each of the response variables for the model to be useful for values outside or even within the ranges of the explanatory variables, even though the model fits the limited number of existing observations quite well. Moreover, even though R^2 may be large, the Standard Error of Estimate (S_e) might be too large for when a high degree of precision is required.

Multiple R This is the square root of Multiple R^2 . It appears in multiple regression output under “Summary Measures.” Although many software packages report this value, it isn’t used much.

Adjusted R^2 Adding more explanatory variables to a model can only increase R^2 , and can never reduce it, because SSE can never become larger when more explanatory variables are present in the model, while SST never changes as variables are added (see the definition of multiple R^2 above). Since R^2 can often increase by throwing in explanatory variables that may artificially inflate the explained variation, the following modification of R^2 , the adjusted R^2 , is one way to account for the addition of explanatory variables: This adjusted coefficient of multiple determination adjusts R^2 by dividing each sum of squares by its associated degrees of freedom (which become smaller with the addition of each new explanatory variable to the model):

$$\text{Adj } R^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SST}$$

The Adjusted R^2 becomes smaller when the decrease in SSE is offset by the loss of a degree of freedom in the denominator $n - p$.

Full Regression Model The full regression model is the multiple regression model that is made using all of the variables that are available.

Best-fit Regression Model This is the regression model arrived after after removing insignificant variables from the full regression model. If a variable has a high p-value in the regression table (usually above 0.05) it is probably not significant and should be removed. But you should only removed the variables one at a time, since the p-values for all the others change each time a different set of variables is used.

9.1.2 Worked Examples

Example 9.1. Reading multiple regression output and generating an equation

Take a look at the data in C09 Rail System. This shows data on the commuter rail system of a large metropolitan area. Each row indicates a year's worth of ridership with data on the price of a ticket on the rail system (in dollars), the population of the region on January 1, the average disposable income (in dollars) of the people in the region, and the average price (in dollars) to park a car downtown. We could use each of these separately to model the number riders on the rail system each week (in thousands), but let's see how multiple regression gives a more complete picture of the situation. If we produce a full regression model using the numerical variables, we get the following output. But what does it mean?

Results of multiple regression for Weekly Riders						
Summary measures						
Multiple R	0.9665					
R-Square	0.9341					
Adj R-Square	0.9259					
StErr of Est	23.0207					
ANOVA table						
Source	df	SS	MS	F	p-value	
Explained	4	240471.2479	60117.8120	113.4404	0.0000	
Unexplained	32	16958.4277	529.9509			
Regression coefficients						
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-173.1971	220.9593	-0.7838	0.4389	-623.2760	276.8819
Price per Ride	-139.3649	42.7085	-3.2632	0.0026	-226.3593	-52.3706
Population	0.7763	0.1186	6.5483	0.0000	0.5349	1.0178
Income	-0.0309	0.0106	-2.9233	0.0063	-0.0524	-0.0094
Parking Rate	131.0352	33.6529	3.8937	0.0005	62.4866	199.5839

First of all, you will notice that format of the regression output is very similar to the format of the output from a simple regression. In fact, other than having more variables, it is not any harder to develop the model equation. We start with the response variable, WeeklyRiders. We then look in the "Regression Coefficients" for each coefficient and the y-intercept. The regression coefficients are in the format:

Regression Coefficients	
	Coefficient
Constant	A
X_1	B_1
X_2	B_2
X_3	B_3
\vdots	\vdots

From this, we can easily write down the equation of the model by inserting the values of the coefficients and the names of the variables from this table into the multiple regression equation shown on page 210:

$$\begin{aligned} \text{Weekly Riders} = & -173.1971 - 139.3649 * \text{Price per Ride} + 0.7763 * \text{Population} \\ & - 0.0309 * \text{Income} + 131.0352 * \text{Parking Rate} \end{aligned}$$

Example 9.2. Interpreting a multiple regression equation and its quality

The rail system model can be interpreted in the following way:

- If all other variables are kept constant (controlled), for each \$1 increase in the cost of a ticket on the rail system, you will lose 139,365 weekly riders. Notice that “weekly riders” is measured in thousands and “price per ride” is in dollars.
- Controlling for price per ride, income and parking rate, every 1,000 people in the city (“population”) will add 776 weekly riders. Notice that this does not mean that 77.6% of the population rides the rail system. Remember, “weekly riders” counts the total number of tickets sold that week. Each one-way trip costs one ticket. This means that a person who uses the rail system to get to work Monday through Friday will count as 10 weekly riders: once each way each day.
- Controlling for price, population and parking, each \$1 of disposable income reduces the number of riders by 0.0309 thousand riders, or about 31. We can scale this up using the idea of proportionality: every \$100 of disposable income will reduce the number of riders by $100 * 0.0309 = 3.09$ thousand.
- If all other variables are controlled, a \$1 increase in parking rates downtown will result in an additional 131,035 weekly riders.
- The constant term, -173.1971, does not make much sense by itself, since it indicates that if the price per ride is \$0, the population is 0, there is no disposable income, and the parking rates are \$0, there will be a negative number of weekly riders. One meaningful way to interpret this is to say that the city needs to be a certain size (population) for the rail system to be a feasible transportation system. (You can solve the equation to find out the “minimum population” for this city to maintain even minimal rail service.)

How good is this model for predicting the number of weekly riders? Let's look at each summary measure, then the p-values, and finally the diagnostic graphs. The R^2 value of 0.9341 indicates that this model explains 93.41% of the total variation in “weekly riders”. That is an excellent model. The standard error of estimate backs this up. At 23,0207, it indicates that the model is accurate at predicting the number of weekly riders to within 23,021 riders (at the 68% level) or 46,042 (at the 95% level). Given that there have been an average of 1,013,189 riders per week with a standard deviation of 84,563, this model is very accurate. The adjusted R^2 value is 0.9259, very close to the multiple R^2 . This indicates that we shouldn't worry too much about whether we are using too many variables in the model. When adjusted R^2 is more than 2-3% different from the R^2 , we should look at the variables and see if any can be eliminated. In this case, though, we should keep them all, unless either the p-values (below) tell us to eliminate a variable or unless we just want to build a simpler, easier-to-use model.

Are there any variables included in the model which should not be there? To answer this, we look at the p-values associated with each coefficient. All but one of these is below the 0.05 level, indicating that these variables are significant in predicting the number of weekly riders. The only one that has a high p-value is the y-intercept; its p-value is 0.4389, which is far above the acceptable level. This indicates that the y-intercept is not significant, and we could reasonably expect that it is actually 0. This would make a lot of sense in interpreting the model, since it would indicate, for example, that if the population is 0, there would be no weekly riders on the rail system. Given this high p-value, you could try systematically eliminating some of the variables, starting with the highest p-values, and looking to see if the constant ever becomes significant. Most regression software will also allow you to force the intercept to be zero, so you can test the model that way. For now, we'll just move on to look at some of the other analysis tools we have access to.

What about the diagnostics graphs? We have four explanatory variables, so we cannot graph the actual data to see if it is linear. Our only options involve the “Fitted vs. Actual” and the “Residuals vs. Fitted” diagnostic graphs. These graphs are shown below.

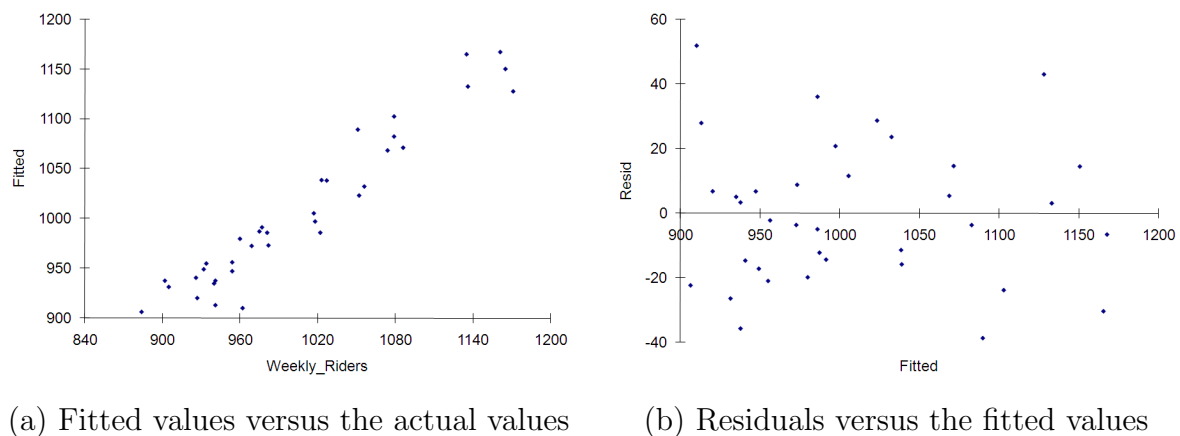


Figure 9.1: Diagnostic graphs for the full regression model of the rail system.

In the “fitted vs. actual” graph, we see that most of the points fall along a straight line has a slope very close to 1. In fact, if you add a trendline to this graph, the slope of the

trendline will equal the multiple R^2 value of the model! So far, it looks like we've got an excellent model for Ms. Carrie Allover.

In the “residuals” graph, we are hoping to see a random scattering of points. Any pattern in the residuals indicates that the underlying data may not be linear and may require a more sophisticated model (see chapters 10 and 11). This graph looks pretty random, so we're probably okay with keeping the linear model.

Example 9.3. Using a multiple regression equation

Once you have a regression equation, it can be used to either

1. predict values for the response variable, based on values of the explanatory variables in between the values in the actual data (interpolation)
2. find values of the explanatory variables that produce a specific value of the response variable (solving an equation)

You will also be tempted to use a regression model to predict values for the response variable, based on values of the explanatory variables that are outside the range included in the actual data (extrapolation). This should only be done with a great deal of caution, since technically, you do not have any information about values of the variables outside of the ranges in your data.

Suppose, for example, that Ms. Allover (see the data file **C09 Rail System**) wants to use the regression model for the rail system that we developed above to predict what next year's weekly ridership will be. If we know what the population, disposable income, parking costs, and price per ride are, we can simply plug these into the equation to calculate the number of weekly riders. Our data stops at 2002. If we know that next year's ticket price won't change, and that the economy is about the same (so that income and parking costs stay the same in 2003 as in 2002) then all we need to know is the population. If demographics show that the population is going to increase by about 5% in 2003, then we can use this to calculate next year's weekly ridership:

$$\text{Next year's population} = (1 + 0.05) * \text{Population in 2002} = 1.05 * 1,685 = 1,770$$

$$\begin{aligned} \text{Weekly Riders} &= -173.1971 - 139.3649 * (1.25) + 0.7763 * (1770) \\ &\quad - 0.0309 * (8925) + 131.0352 * (2.25) \\ &= 1,045.744 \end{aligned}$$

It is important to notice that if any of the variables change, the final result will change. Also notice that a 5% change in population while keeping all the other explanatory variables constant results in a $(1045.744 - 960)/960 = 8.9\%$ change in the number of weekly riders. If the values of the other variables were different, the change in the number of weekly riders would be a different amount.

If we wanted to solve the regression equation for values of the explanatory variables, keep in mind this rule: For each piece of “missing information” you need another equation. This

means that if you are missing one variable (either weekly riders, price per ride, population, income, or parking rates) then you can use the values of the others, together with the equation, to find the missing value. If you are missing two or more variables, though, you need more equations.

9.1.3 Exploration 9A: Production Line Data

This example is based on a data file from Albright, Winston, and Zappe (2010) *Data Analysis and Decision Making*.

The WheelRight company manufactures parts for automobiles. The factory manager wants a better understanding of overhead costs at her factory. She knows that the total overhead costs include labor costs, electricity, materials, repairs, and various other quantities, but she wants to understand how the total overhead costs are related to the way in which the assembly line is used. For the past 36 months, she has tracked the overhead costs along with two quantities that she suspects are relevant (see data file **C09 Production**):

- MachHrs is the number of hours the assembly machines ran during the month
- ProdRuns is the number of different production runs during the month

MachHrs directly measures the amount of work being done. However, each time a new part needs to be manufactured, the machines must be re-configured for production. This starts a new production run, but it takes time to reset the machine and get the materials prepared.

Your task is to assist the manager in understanding how each of these variables affects the overhead costs in her factory.

- a. First, formulate and estimate two simple regression models to predict overhead, once as a function of MachHrs and once as a function of ProdRuns. Which model is better?
- b. Would you expect that the combination of both variables will do a better job predicting overhead? Why or why not? How much better would you estimate the multiple regression model to be?
- c. Formulate and estimate a multiple regression model using the given data. Interpret each of the estimated regression coefficients. Be sure to include the units of each coefficient.
- d. Compute and interpret the standard error of estimate and the coefficient of determination. Examine the diagnostic graphs “Fitted vs. Actual” and “Residuals vs. Fitted”. What do these tell you about the multiple regression model?
- e. Explain how having this information could help the manager in the future.

9.2 Modeling with Qualitative Variables

For most statistical packages, an explanatory variable is the name of a column of data. This name usually sits at the head of its data column in the spreadsheet and appears, as we have seen, in the regression equation. A statistical package carries out regression analysis by regarding all entries in a column under a variable name as numerical data. The data listed under a categorical variable, however, may be in the form of words or letters so that the mathematical operations necessary to perform linear regression would not make any sense.

What we need is a way to convert the categories of a categorical variable into numbers. But we must do it in such a way that it makes sense and that everyone can agree on the definitions. Otherwise, the mathematics will not make sense. The key to converting categorical data into numerical data is this: Categorical data falls into two or more categories but no observation is ever in more than one category at a time. In other words, if a variable called “Style of House” has the categories “colonial”, “ranch”, “split-level”, “cape cod” or “other”, then any given house (a single observation of “style of house”) can only be one of these types.

What we cannot do to convert the categories into numbers is to simply number each category. Numerical data is, by its very nature, ordered data. It has a natural structure. In mathematics, 3 is bigger than 2, and 2 is bigger than 1. So how, ahead of time, can we know which category is “bigger” than another? How do we know which category should be numbered 1, which should be 2, etc.? Since we cannot determine this ahead of time, we must find another approach to converting the categorical data into numerical data. The problem with this approach is that we tried to do it with a single variable that has different numerical values.

In order for statistical packages to be able to create regression models, the various values in each category may have to be translated into separate, individual “dummy” variables, such as `StyleColonial`, `StyleRanch`, `StyleSplitLevel`, etc. These dummy variables can take only the values 1 or 0. For a given observation, one of the dummy variables will be equal to 1, the dummy variable named for the category that the observation fits into. The other dummy variables associated with this categorical variable will be 0, because the observation does not fall into those categories. Essentially, statistical packages, such as `StatPro`, handle categorical data as switches: either a category variable applies or it does not; it is “on” (equal to 1) or it is “off” (equal to 0).

We can then use these dummy variables (and not the original categorical variable) to build a regression equation. Each of these dummy variables will have its own coefficient. This allows us to create complex models using all sorts of data. After all, you expect categorical data to be important in most models. If you were trying to predict the cost of shipping a package, for example, the weight and its destination might be important, but so would the delicacy of the package. “Fragile” packages would cost more to ship than “durable” packages. The only way to include this characteristic in the model is through dummy variables.

9.2.1 Definitions and Formulas

Dummy variables These are variables made from a categorical variable. For each category in the variable, one dummy variable must be created. Normally, these are named by adding the category name to the end of the variable name. For a given observation, if the observation is in the category associated with a dummy variable, then the value of the dummy variable is 1 (for “yes, I’m in this category”). If the observation is not in the category associated with the dummy variable, then the dummy variable is equal to 0 (for “no, I’m not one of these”). Dummy variables are also called indicator or 0-1 variables.

Dummy variables are called “dummy” because they are artificial variables that 1) do not occur in the original data and 2) are created solely for the purpose of transforming categorical data into numerical data. In that sense, they are a special kind of **constructed variable**.

Exact multicollinearity This is an error that can occur if some of the explanatory variables are exactly related by a linear equation. For example, a new variable “total boxes” could be constructed by summing the numbers of small, medium, and large boxes. This new variable would be exactly related to all three of the underlying variables.

Reference category When creating a regression model, to avoid exact multicollinearity, it is necessary that one of the dummy variables be left out of each group that came from a single categorical variable. The dummy variable left out is the reference category to which all interpretation of the model coefficients must be compared.

9.2.2 Worked Examples

Example 9.4. Converting two-valued categorical data to dummy variables

A categorical variable must have at least two categories. Suppose a categorical variable has exactly two values. These values are used to indicate whether the category applies to a particular individual or does not. A good example of this is “Gender”. It has two values: male and female. Furthermore, since no one can be both male and female, each person is coded as either male or female (M or F, 0 or 1, etc). This means that we can create two dummy variables, one for GenderMale and one for GenderFemale. Each observation will have one of these two dummy variables equal to 1 and the other 0, since no observation can fall into multiple categories at the same time; a person falls into one or the other, but not both. So we can go down the list of data and enter 1 and 0 where we need to in order to create our dummy variables, or we could create an “If statement” in our spreadsheet or software to fill in these 1’s and 0’s for us.

Example 9.5. Converting multi-values categorical data to dummy variables

What about categorical variables with more than two categories? A good example of this is an employee’s education, which is coded with several category values (0,2,4,6,8) indicating the level of post-secondary education the employee has had, where 0 indicates no postsecondary

education, 2 indicates an associate's degree, 4 indicates a bachelor's degree, 6 indicates a master's degree and 8 indicates a doctorate. Each employee is classified according to the Education categorical variable and is assigned to one and only one of the five possible educational levels. In the end, you would wind up with the following data:

Original data

Categorical variable: Education

Has five categories: 0, 2, 4, 6, 8

Employee has	Education
No postsecondary	0
Associate's degree	2
Bachelor's degree	4
Master's degree	6
Ph.D.	8

Dummy variables

Five dummy variables (Ed#)

	Ed0	Ed2	Ed4	Ed6	Ed8
No postsecondary	1	0	0	0	0
Associate's degree	0	1	0	0	0
Bachelor's degree	0	0	1	0	0
Master's degree	0	0	0	1	0
Ph.D.	0	0	0	0	1

Example 9.6. Regression equations with dummy variables

Suppose we have a database of employee information are interested in whether “gender” has an effect on an employee’s salary. Such questions are common in gender discrimination lawsuits. (We are not saying that employers purposely compute salaries differently for male and female employees. We are merely saying that after everything is accounted for, it is possible that gender is underlying some of the salary differences in employees.) In our hypothetical data, we have three variables: gender, age, and annual salary. A sample of this data is shown below. Gender is a categorical variable with two values: “M” for male and “F” for female. Age is simply the age of the employee. We are using this as a stand-in (or surrogate) variable to include the effects of experience, education, and other time-related factors on salary. Annual salary is coded in actual dollars. We want to build a regression model to predict annual salary.

	Gender	Age	Annual Salary
Employee 1	M	55	57457
Employee 2	F	43	36345
Employee 3	F	25	23564
Employee 4	M	49	38745
Employee 5	F	52	41464
⋮	⋮	⋮	⋮

First we create dummy variables, “GenderM” and “GenderF”. Employee 1 is male, so this observation will have GenderM = 1 and GenderF = 0. Employee 2 will have GenderM = 0 and GenderF = 1, since employee 2 is female. The data now contains four variables: Gender, Age, Annual Salary, GenderM, and GenderF. To build the regression model, we select the explanatory variables that are appropriate. However, we cannot use both dummy variables. Let’s use GenderF in the equation. After all, if GenderF = 0, then we know the employee is male, so we don’t need the other dummy variable. The regression output looks exactly like multiple regression output and can read in exactly the same way. We find the full regression model to be

$$\text{Annual Salary} = 4667 - 2345 * \text{GenderF} + 845 * \text{Age}$$

When GenderF has value 0 (male employee), the salary is

$$\text{Annual Salary} = 4667 - 2345 * (0) + 845 * \text{Age} = 4667 + 845 * \text{Age}$$

When GenderF has value 1, (female employee), the salary is

$$\text{Annual Salary} = 4667 - 2345 * 1 + 845 * \text{Age} = 2322 + 845 * \text{Age}$$

We can now see that the single regression equation with dummy variables is actually two separate equations, one for each gender:

$$\text{For a female employee: } \text{Annual Salary} = 2322 + 845 * \text{Age}$$

$$\text{For a male employee: } \text{Annual Salary} = 4667 + 845 * \text{Age}$$

What do these equations mean? When we control for age, that is, when the ages of the employees are the same, the model predicts that a female employee will earn \$2345 per year less than a man. Notice that the slopes of the two equations - the rate at which salary increases based on age, is the same for both male and female employees. What is different is the starting salary, represented in these equations by the y -intercepts.

9.2.3 Exploration 9B: Maintenance Cost for Trucks

This example is based on a data file from Albright, Winston, and Zappe (2010) *Data Analysis and Decision Making*.

The data file `C09 Truck data` contains information on trucks owned by Metro Area Trucking. We are interested in predicting how all of the variables influence the maintenance costs.

1. Analyze the Variables

- What variable would be the response variable?
- Which of the explanatory variables are numerical? What are their units?
- Which explanatory variables are categorical? What are the possible categories for each?
- What dummy variables need to be created? (Notice that “location” is already coded as 0 or 1, so there is no need to create dummy variables for it.)

2. Build the models

- Create the full regression model. What is the equation of the model? How good is this model? What does it tell you about maintenance costs for each type of truck? How does location influence the maintenance cost?
- Are there any variables in the full model that should be eliminated? Why? Is there a theoretical justification for eliminating them?
- Create a model with nonessential variables eliminated. What is the model equation? How does it compare (in quality) with the full regression model? What does it tell you about the maintenance costs of each type of truck? What does this model tell you about how location affects maintenance costs?

9.3 Homework

Mechanics and Techniques Problems

9.1. A regional express delivery company has asked you to estimate the price of shipping a package based on the durability of the package. You randomly sample the packages, making sure that you get packages that are all about the same size and are being shipped about the same distance. The company rates the durability of a package as either “durable”, “semifragile” or “fragile”. Your data on fifteen packages is in the file **C09 Shipping**.

1. Formulate a multiple regression model to predict the cost of shipping a package as a function of its durability.
2. Interpret the regression coefficients and the quality of your model.
3. According to your model, what type of package is the most expensive to ship? Which is the least expensive to ship?
4. Use your model to predict the cost of shipping a semifragile package.
5. Why is it important that the packages sampled in the data are all “about the same size” and “shipped about the same distance”?

9.2. Consider the housing data in **C09 Homes**. We are going to build a model using the location and style of the home, along with some of the numerical variables, to see how these affect the price, and whether they are significant. You may want to use a table like the one below to record your work.

1. First create dummy variables for the location and the style variables.
2. Formulate a multiple regression model using the location data and the numerical variables Age, Size, Taxes, and Baths. Comment on the interpretation of this model and its quality. Finally, comment on whether this model proves the old adage “The three most important things in real estate are location, location, location.”
3. Formulate a multiple regression model using the style data and the numerical variables Age, Size, Taxes, and Baths. Comment on the interpretation of this model and its quality. Compare it to the location model you created in part b.
4. Formulate a multiple regression model using the same numerical variables as before, and using both the style and location data. How does this model compare with the previous two models?
5. Which of the models (just the numerical, numerical plus location, numerical plus style, or numerical plus style and location) would you recommend that the realtor use for making pricing decisions? Why?

Application and Reasoning Problems

9.3. Ms. Carrie Allover needs more information about the model we developed to predict the number of weekly riders on her commuter rail system. The model equation is in example

1. Recall that it predicts the number of weekly riders based on population, price per ride, parking rates, and disposable income. Ms. Allover wants more explanation of what the equation means. She has asked some very specific questions about the situation.

1. Based on the model equation, which of the following will have the largest impact on the number of weekly riders: an increase of 10,000 people in the region, a ten cent drop in the price per ticket, a ten cent raise in parking rates, or a \$100 decrease in average disposable income? Explain your answer.
2. Demographics experts suggest that the population will drop by 10% next year. The model predicts that this will change the number of weekly riders. Ms. Allover wants to ensure that the revenue (=price per ticket * number of tickets sold) remains about the same for next year as it is for this year. In order to accomplish this, the price per ticket will have to change. Should the ticket price be raised or lowered? By how much? Use the regression model and your software to help answer this.

9.4. The data file **C09 Homes** contains data on 271 homes sold in a three-month period in 2001 in the greater Rochester, NY area. A realtor has enlisted your help to develop a regression model in order to explain which characteristics of a home influence its price. You are going to build the regression model by adding one variable at a time and removing variables that do not seem to be significant. At each stage of the model building process, record the equation of the model, the R^2 , the adjusted R^2 and the standard error of estimate. You should record all of this information in a table like the one below in order to make it easier to compare the results.

1. Introduce a new variable for the age of the home. To do this, add a new column heading “Age” in cell M3. In cell M4, enter the formula “=2003 - H4” in order to calculate the age of the home based on the year in which it was built (H4). Copy this formula to all the cells in the column.
2. Develop a series of models to predict the price of the home by adding one variable at a time. Add them in this order: Size, Baths, Age, Acres, Rooms, and Taxes. Make sure that each model includes all of the previous variables. (The second model will include size and baths as explanatory variables; the third will include size, baths, and age.) Record the model equation and the summary measures indicated in the table below.
3. What do you expect to happen to each of the summary measures as you add more variables into the model? What actually happens each time? What do the differences tell you about some of the variables?
4. Based on your observations of the summary measures eliminate the variable or variables that you feel are not helpful in predicting the price of a home. Using the remaining

variables, develop your “best regression model” and compare it to the others you have developed.

Sample Table for Recording the Housing Models in Problem 2

Variable Added	Model equation	R^2	Adj. R^2	S_e
Size				
Baths				
Age				
Acres				
Rooms				
Taxes				
Best Model				

Is the Model Any Good?¹

In the last chapter we built regression models that measured the effects of several explanatory variables on a dependent variable. For example, how educational background, prior experience, years with a company, job level, or gender affect salary. We determined how each explanatory variable, whether numerical or categorical, expressed its effect on salary through its coefficient in the regression equation. The process of building such a model is a statistical one; that is, it involves determining a best-fit equation by calculating how much of the total variation is accounted for by the model. This calculation, in turn, is based on certain probabilistic assumptions concerning how the data is distributed.

- Section 10.1 concerns how confident we can be that the coefficients of our explanatory variables are trustworthy. This is critically important if we are to make decisions based on our understanding of what a model seems to be telling us. We need criteria to determine which explanatory variables are truly significant in affecting the dependent variable—and which are not—if our model is to be at all useful. This section helps us to separate the wheat from the chaff.
- Section 10.2 furthers the process of building more complex and accurate models from several explanatory variables by considering how interactions between the variables themselves might have an effect on the dependent variable. That is, some of these variables might express their effects on the dependent variable in combination with other explanatory variables. In fact, there are even cases in which an explanatory variable appears to have a significant effect only when it is combined with one or more other explanatory variables. For example, it may be that employees' gender by itself has no significant effect on salary, but gender together with job level might have a negative impact on salary. That is, the negative effect of gender on salary only has a significant impact when the employee is a female in a higher-level position: the well-known “glass-ceiling” effect. This section, then, concerns not only the effects of several

¹©2017 Kris H. Green and W. Allen Emerson

individual explanatory variables on a dependent variable, but also the effects of pairs of them on the dependent variable. You will learn in this chapter how to create multiple regression models with interaction variables built from both numerical and categorical explanatory variables and assess their significance. You will learn how to analyze and interpret these often complex models.

As a result of this chapter, students will learn

- ✓ How to determine the trustworthiness of the coefficients of a regression equation
- ✓ How to determine which coefficients should be kept in a model and which should not
- ✓ How to interpret models with complex interaction terms involving both numerical and categorical variables

As a result of this chapter, students will be able to

- ✓ To determine with 95% confidence the range of values within which regressions coefficients fall
- ✓ Create interaction terms
- ✓ Identify the reference categories of interaction variables
- ✓ Construct interaction variables from existing variables in a data set
- ✓ Construct a model using interaction terms
- ✓ How to use stepwise regression to build complex models with significant variables

10.1 Which coefficients are trustworthy?

In the last chapter, several regression models of EnPact's employee salary structure were developed in order to determine if female employees earn less than their male counterparts. These models indicate that females do earn less than their male counterparts, often many thousands of dollars a year less, depending on which variables are used in the models. As EnPact's Human Resources Director, you are aware that if females do indeed earn substantially less than males, say \$5000 a year, then EnPact could be liable for a potentially ruinous multi-million dollar law suit. But to what degree can you be confident that these models are indeed producing accurate results?

We will answer this question and related questions in this chapter, but first we need some concepts.

Suppose we have a regression equation with two explanatory variables, X_1 and X_2 , and their coefficients, B_1 and B_2 , respectively:

$$\text{dependent variable} = \text{constant} + B_1 \times X_1 + B_2 \times X_2$$

If one of the coefficients is zero, say B_1 , then X_1 makes no contribution to the dependent variable no matter what value it takes on because $0 \times X_1 = 0$ and the equation reduces to

$$\text{dependent variable} = \text{constant} + B_2 \times X_2$$

In this case, X_1 is said to be insignificant.

Just because a coefficient is nonzero, however, does not mean that the variable is necessarily significant. A statistician would warn us that regression coefficients are only estimates. Remember: the data we are working with is a *sample* rather than the entire population. If we sample the data again, we would get different values for the coefficients in the regression model. This means that some of the coefficients could wind up zero or even changing sign based on a different sample. If that were to happen, we could have a more accurate model by setting those coefficients to zero, which is the same as just eliminating them from the model completely. The question is, then, how do we identify which variables are insignificant? We have two tools to help us, and each provides a different insight into the model.

The regression output provides us with p -values and confidence intervals for each coefficient. We use these to determine the significance of each coefficient. A variable is significant if its p -value is less than 0.05. This is basically saying there is less than a 5% chance that the coefficient would be zero if we were to compute the model from a different data sample. This turns out to be equivalent to saying that zero is not in the 95% confidence interval of the coefficient. Thus, we can never be 100% certain, but we can set a level of significance and use that. Most modelers use the 95% level to determine which variables are significant and which are not. When statisticians use the phrase, "95% confident," they mean that 95% of the time we will be able to correctly identify whether a particular variable is or is not significant. Since 95% is 0.95 as a decimal, you can see where the p -value of 0.05 comes from $100\% - 95\% = 1.00 - 0.95 = 0.05$.

10.1.1 Definitions and Formulas

p-values A p-value is a probability assigned to determine whether a given hypothesis is true or not. In regression analysis, p-values are used to determine whether or not a given explanatory variable should have a coefficient of “0” in the regression model. If the p-value is above 0.05 (5%) then one can usually leave the variable out and get a model that is almost as good.

$$\begin{array}{ll} p < 0.05 & \text{Keep the variable} \\ p > 0.05 & \text{Drop the variable} \end{array}$$

Significant variable or coefficient A variable or a coefficient of a variable is significant when its p-value is less than .05. That is, there is less than a 5% chance that the coefficient is zero.

Insignificant variable or coefficient A variable or a coefficient of a variable is insignificant when its p-value is greater than .05. That is, there is more than a 5% chance that the coefficient is zero. As a general rule (there are exceptions), when a variable is found to be insignificant in a particular model, it should not be included in future models.

95% confidence interval The interval in which we can be 95% certain that a coefficient will lie, meaning that the coefficient will lie in this interval 95% of the time.

Principle of parsimony Equivalent to K.I.S.S. If we have a choice between two models, we should choose the simpler or smaller model of the two, provided that it does reasonably as well as the larger, more complicated model. (This principle is also known as Occam’s Razor: Things should not be multiplied without reason.)

10.1.2 Worked Examples

Example 10.1. Determining the significance of a variable from a p-value

We look to the last three columns of the “Regression coefficients” block in the spreadsheet below to determine if a variable is significant. This data is shown in file **C10 Enpact Data**. The variable HiJob is a dummy variable that is 1 if the employee’s job grade is 5 or 6.

We can determine if a variable, say HiJob, is significant by examining the p-value of its coefficient (third column from the right in the regression output.) Since its p-value, .0000, is less than .05, we can expect that its coefficient, 8.7389, will be zero less than 5% of the time. This means that we can expect the coefficient will not be zero 95% of the time and therefore the variable is significant at a 95% level of confidence. On the other hand, the p-value of the coefficient of the Age variable is .5670, which is greater than .05. This says that the Age variable is insignificant because we cannot be confident that its coefficient, 0.0374, is nonzero less than 5% of the time.

Example 10.2. Determining significance of a variable from a confidence interval

	A	B	C	D	E	F	G	H	I	J	K
1	Results of multiple regression for Salary										
2											
3	Summary measures										
4		Multiple R	0.8291								
5		R-Square	0.6875								
6		Adj R-Square	0.6732								
7		StErr of Est	6.4433								
8											
9	ANOVA Table										
10		Source	df	SS	MS	F	p-value				
11		Explained	9	18080.8099	2008.9789	48.3901	0.0000				
12		Unexplained	198	8220.2393	41.5164						
13											
14	Regression coefficients										
15		Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit				
16		Constant	29.7310	2.7202	10.9298	0.0000	24.3667	35.0952			
17		Age	0.0374	0.0652	0.5735	0.5670	-0.0911	0.1659			
18		YrsExp	0.7785	0.1002	7.7672	0.0000	0.5808	0.9761			
19		YrsPrior	0.2887	0.1548	1.8650	0.0637	-0.0166	0.5940			
20		EducLev_2	-0.0219	1.5669	-0.0140	0.9889	-3.1119	3.0681			
21		EducLev_3	3.8690	1.4575	2.6545	0.0086	0.9947	6.7433			
22		EducLev_4	4.9235	2.5851	1.9046	0.0583	-0.1744	10.0215			
23		EducLev_5	8.4553	1.5995	5.2862	0.0000	5.3010	11.6095			
24		Gender_Female	-3.0428	1.0644	-2.8587	0.0047	-5.1419	-0.9438			
25		HiJob	8.7389	1.6732	5.2228	0.0000	5.4393	12.0385			

Figure 10.1: Multiple regression results with p -values and confidence intervals highlighted.

Using the same regression output as the previous example, we look at the confidence intervals. The lower and upper limits of the interval are given in the last two columns of the “Regression Coefficients” section of the output. This lets us see that we can be 95% confident that the coefficient the Age variable lies somewhere between the lower-limit number, -.0911, and the upper-limit number, .1659. Since the lower limit is negative and the upper limit is positive, the coefficient, given as .0374, could very well be 0. This means that the variable is insignificant. On the other hand, if the signs of the lower and upper limits are the same, then we can be 95% confident that the associated variable (or the constant in the case of the first row) is not zero and is therefore significant at a 95% level of confidence. For example, we can be 95% confident that the variable YrsExp is significant and that its coefficient lies somewhere between .5808 and .9761.

Example 10.3. The relative advantages of using confidence intervals vs p -values

A confidence interval not only tells us whether a variable is significant or not, it also gives us a range of values within which we can be 95% confident that the coefficient will lie. A p -value only tells us whether a variable is significant or not. On the other hand, the eye can scan a single column of p -values for significance much quicker and readily than it can scan two columns of numbers looking for a sign change across them.

Example 10.4. Refining your model

The presence of insignificant variables in a model is usually means tht your model could be improved in some way. The reason is this: the presence of insignificant variables artificially raises the R^2 value of the model by introducing information in which we should not have

confidence. In other words, insignificant variables inflate the model's R^2 so that it is not a reliable indicator of how well the model fits the data. This means that we could be basing our inferences and decisions on a faulty model. And since the coefficient values and the p -values are different with a different set of variables, the more accurate model could be quite a bit different from the model that includes insignificant variables.

To avoid the problem of producing an untrustworthy model, we rerun the regression routine after leaving out all the insignificant variables. Our new reduced model will now be built with significant explanatory variables, each of which has passed the 95% confidence test. After dropping the insignificant variables from the model displayed in example 2, our reduced model will now be based on the following significant variables: YrsExp, HiJob, GenderFemale, EducLevel3, EducLevel4, and EducLevel5. The resulting reduced model is shown below:

	A	B	C	D	E	F	G	H
1	Results of multiple regression for Salary							
2								
3	Summary measures							
4		Multiple R	0.8246					
5		R-Square	0.6799					
6		Adj R-Square	0.6704					
7		StErr of Est	6.4716					
8								
9	ANOVA Table							
10		Source	df	SS	MS	F	p-value	
11		Explained	6	17882.9017	2980.4836	71.1650	0.0000	
12		Unexplained	201	8418.1475	41.8813			
13								
14	Regression coefficients							
15			Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
16		Constant	32.1066	1.4074	22.8121	0.0000	29.3313	34.8818
17		YrsExp	0.7897	0.0770	10.2544	0.0000	0.6378	0.9415
18		EducLev_3	3.2902	1.1663	2.8210	0.0053	0.9904	5.5900
19		EducLev_4	4.3646	2.4630	1.7721	0.0779	-0.4920	9.2212
20		EducLev_5	7.8043	1.3451	5.8019	0.0000	5.1519	10.4568
21		Gender_Female	-2.9490	1.0662	-2.7660	0.0062	-5.0513	-0.8467
22		HiJob	9.0251	1.6740	5.3912	0.0000	5.7241	12.3260

Figure 10.2: Regression output for Enpact data after insignificant variables are dropped.

Notice that the R^2 of our reduced model, 0.8246, is only slightly smaller than the R^2 of the full regression model. For all practical purposes, the R^2 of the original model and the reduced model are the nearly identical. Similarly, the S_e of the reduced model, 6.4716, is larger than the S_e of the full model, but only by .0283, which again, for all practical purposes, is nearly identical. Other models, however, may show much larger differences between the R^2 and S_e of the full model and a reduced one.

This example illustrates another principle of good modeling practice: the principle of parsimony. The principle of parsimony can be thought of as a principle of simplicity. If a smaller set of explanatory variables produces a model that fits the data almost as well as a

model with a larger set of explanatory variables - and with almost the same standard error - it is usually preferable to use the model with the smaller number of explanatory variables. As we shall see, each explanatory variable in a model comes with a price, not only in terms of increasing the unwieldiness of the model, but more importantly in terms of understanding or explaining how the particular variable affects the dependent variable.

Also notice that one of the variables in the original example 2, the variable `EducLevel4`, was on the border between being significant and not. Its value of 0.0583 is right about equal to the cutoff of 0.05. Because the p-values change dramatically as variables are eliminated from the model, it is important to leave such borderline variables in the model at first and see if they become more significant. In this case, the p-value got larger when we eliminated some of the variables; in the reduced model, it is definitely not significant at a p-value of 0.0779. In fact, because of the way p-values change as the variables are eliminated, it is always best to eliminate one variable at a time, making a new model as each of the variables is dropped and re-assessing which variables are significant. Often, a variable that began an insignificant can become significant.

Summary: Refining a model is both an art and a science. The general procedure is:

1. Run a full model with all the explanatory variables
2. Determine the significant explanatory variable from the results of the full model
3. Run a reduced model with the variables from 2.
4. With the principle of parsimony in mind, run models built on various subsets of significant (or nearly significant) explanatory variables until you obtain a model that you are satisfied gives the best fit to the data with the fewest explanatory variables.

10.1.3 Exploration 10A: Building a Trustworthy Model at EnPact

1. Construct a full regression model with all the explanatory variables, both numerical and categorical, of the EnPact data found in **C11 EnPact Data**. Be sure to create dummy variables of the categorical data first, if your software package requires it. And while the Job Grade and Education Level variables are ordinal, they are categorical and should be treated as such. Enter your results in the chart below.
2. Select the significant variables from the output of the full model regression in Part 1 and run the reduced model. Record your results in the chart below.
3. Use your software's stepwise regression procedure with the complete set of numerical and categorical explanatory variables. Enter your results in the chart below.

	Model	R^2	Adj R^2	S_e	List of significant variables
Part 1	Full Model				
Part 2	Reduced Model				
Part 3	Stepwise regression				

4. What do you observe about your results from Parts 2 and 3? How do you account for this?
5. Write down what you think is the most suitable model and defend your choice.
6. Interpret your model.

10.2 More Complexity with Interaction Terms

We are becoming aware that gender may have a significant impact on employees' salaries at EnPact. But is its impact isolated from that of the other variables that affect salary? Is it possible that the variable GenderFemale, for example, is somehow implicated in the impact that some other variable, say YrsExp, has on salary? If so, then a portion of the magnitude of the coefficient of YrsExp (the measurable effect of experience on salary) should actually be attributed to gender. Or, to put it another way, some of the effect of gender on salary is lost to experience. This means that our regression model is not measuring the true effect that gender has on salary. In addition, our understanding of the nature of any alleged discrimination at EnPact would be greatly increased if we could not only measure the effect that gender by itself makes on salary, but also measure the effect that the interplay or interaction between gender and years of experience makes on employees' salaries. Similarly, it would also be informative to learn, for example, that gender does not play a role in how some other variable, say education, affects salary.

These kinds of combined effects can be captured in regression models by forming new variables called interaction variables (or terms), which are created by taking the product of two variables that we believe have a combined effect on the dependent variable. The first entry in a column of data for an interaction variable $X_1 \times X_2$ is the product of the first entry of X_1 with the first entry of X_2 . The second entry of $X_1 \times X_2$ is the product of the second entry of X_1 with the second entry of X_2 , etc. When the interaction variables and the original variables are submitted to a regression routine, its computational procedure makes no distinction between variables that are interaction variables and those which are not. When the regression coefficients are computed for any set of variables, the software treats all columns of data with names at their heads the same, whether those names are GenderFemale, YrsExper, or GenderFemale*YrsExp. Most packages have a convenient routine for creating interaction terms.

The following is an example of a regression model containing interaction variables:

$$\begin{aligned} \text{Salary} = & 25 + 1.2 * \text{YrsExp} - 2.4 * \text{GenderFemale} - .80 * \text{GenderFemale} * \text{YrsExp} \\ & + 1.30 * \text{GenderFemale} * \text{EducLev3} - .42 * \text{GenderFemale} * \text{EducLev6} \end{aligned}$$

Things to know about interaction terms when building models:

1. Variables that were significant before the introduction of interaction variables may become insignificant in subsequent models containing the interaction variables
2. The reverse can also occur. That is, variables that have been insignificant may become significant when combined in new interaction terms.

10.2.1 Definitions and Formulas

Interaction variable The product of two variables, say Female and Age, that constitutes a new variable and that captures, if it proves to be significant, the combined effect of the two original variables. An interaction variable is formed by multiplying the

corresponding cells of the two variables and placing the resulting products in a new column, usually denoted, for example, by Female \times Age.

Interaction terms can be created from any two variables. Most commonly, though, they are created from interacting either two categorical variables, or a categorical variable and a numerical variable. Interaction variables created from two numerical variables really lead us away from linear models for the data and create one type of quadratic model (See chapter 13).

Base Variable These are the original “uninteracted” variables from which the interaction terms were created.

10.2.2 Worked Examples

Example 10.5. Creating and interpreting interaction terms from the EnPact data

An interaction term can be created from a numerical variable and a categorical variable:

Variable Type	Variable Name	Categories
The numerical variable	Age	N/A
The categorical variable	EducLev	EducLev1, EducLev2, EducLev3, EducLev4, EducLev5
The interaction variable	Age*EducLev	Age* EducLev1, Age* EducLev2, Age* EducLev3 Age* EducLev4, Age* EducLev5

We will interpret a rather simple model built on Age, EducLev3 and Age \times EducLev3 where EducLev1 indicates a high-school grad and has been chosen as the reference category for the categorical variable EducLev, and EducLev3 indicates a college grad.

$$\text{Model: Salary} = 12 + .56*\text{Age} + 5.2*\text{EducLev3} + .22*\text{Age}*\text{EducLev3}$$

Interpretation: When EducLev3 has the value 1, a college graduate is indicated. After substituting 1 for EducLev3 in the model equation, we have

$$\text{Salary} = 12 + .56*\text{Age} + 5.2*1 + .22*\text{Age}*1$$

After combining the Age terms, we have a college grad’s salary:

$$\text{Salary} = 17.2 + .78*\text{Age} \quad (1)$$

When EducLev3 has the value 0, a high-school graduate is indicated. After substituting 0 for EducLev3 in the model equation, we have

$$\text{Salary} = 12 + .56*\text{Age} + 5.2*0 + .22*\text{Age}*0$$

Simplifying, we have a high-school grad's salary:

$$\text{Salary} = 12 + .56 * \text{Age} \quad (2)$$

Comparing equations (1) and (2), we see that a college grad receives a bonus of \$5200 ($17.2 - 12 = 5.2$) for having a college degree plus an additional \$220 ($.78 - .56 = .220$) for each year that he or she has lived compared to a high-school grad of the same age. At age 30, for example, a high-school grad earns \$28,800 whereas a 30-year old college grad earns \$40,600. At age 60, they earn \$45,600 and \$64,000, respectively.

Example 10.6. An interaction terms created from two categorical variables

Suppose we have the variables Gender and EducLev from the previous example, and we plan to construct an interaction term using these variables.

Gender: GenderFemale, GenderMale
 Reference category: GenderMale
 EducLev: EducLev1, EducLev2, EducLev3, EducLev4, EducLev5
 Reference category: EducLev1

There are 2×5 , or 10, interaction terms involved in the interaction variable Gender*Ed. Not all 10 can be submitted to a regression routine, however. Only those interaction terms that do not contain a reference for either variable may be submitted to the regression routine. The following interaction terms are the only ones that may be submitted to a regression routine:

EducLev2*GenderFemale
 EducLev3*GenderFemale
 EducLev4*GenderFemale
 EducLev5*GenderFemale

The other interaction terms cannot be submitted to because each contains either one or both of the reference categories (in bold) from which they are created: **EdLev1* GenderMale**, **EducLev1*GenderFemale**, EducLev2***GenderMale**, EducLev3 * **GenderMale**, EducLev4 * **GenderMale**, EducLev5* **GenderMale**. This means that each of these is a reference category for the interaction variable EducLev*Gender.

We will interpret a modification of the models built above based on the variables Age, EducLev3, Age* EducLev3, GenderFemale and EducLev3*GenderFemale.

$$\begin{aligned} \text{Model: Salary} = & 13 + .52 * \text{Age} + 5.8 * \text{EducLev3} + .21 * \text{Age} * \text{EducLev3} \\ & + 4.1 * \text{GenderFemale} - 2.5 * \text{EducLev3} * \text{GenderFemale} \end{aligned}$$

Interpretation: If GenderFemale = 0 and EducLev3 = 1, we have a male college graduate. Substituting these values in the model equation, we have

$$\text{Salary} = 13 + .52 * \text{Age} + 5.8 * 1 + .21 * \text{Age} * 1 + 4.1 * 0 - 2.5 * 1 * 0$$

Combining the constants and the Age terms, we have the equation for a male college graduate

$$\text{Salary} = 18.8 + .73 * \text{Age} \quad (3)$$

If GenderFemale = 1 and EducLev3 = 1, we have a female college graduate. Substituting these values in the model equation, we have

$$\text{Salary} = 13 + .52 * \text{Age} + 5.8 * 1 + .21 * \text{Age} * 1 + 4.1 * 1 - 2.5 * 1 * 1 \quad (4)$$

In equation (4) we see that a female receives \$4100 more than a male on the basis of gender alone. But she will receive \$2500 less than a male if she has a college degree. Simplifying (4), we have the equation for a female college graduate:

$$\text{Salary} = 20.4 + .73 * \text{Age} \quad (5)$$

Comparing (3) and (5), we see that a female college graduate earns on the average of \$1600 (20.4-18.8) more than a male college graduate. The difference is larger, however, for high school graduates (EducLev3 = 0). In this case, female high-school graduates earn \$4100 a year more than male graduates. For example, comparing the salaries of 25-year old high school graduates, we have:

$$\begin{aligned} \text{Female: } \text{Salary} &= 13 + .52 * 25 + 5.8 * 0 + .21 * 25 * 0 + 4.1 * 1 - 2.5 * 0 * 1 \\ &= \$30,100 \\ \text{Male: } \text{Salary} &= 13 + .52 * 25 + 5.8 * 0 + .21 * 25 * 0 + 4.1 * 0 - 2.5 * 0 * 0 \\ &= \$26,000 \end{aligned}$$

Example 10.7. Simplifying variables in the EnPact data

When we introduce interaction variables into the EnPact gender discrimination study, we find that if we use the given variable names as they are found in **C11 EnPact** the software will create interaction variable names that are too long to be completely viewed in its multiple regression routine window. In addition, when we interact categorical variables with other variables, particularly other categorical variables, the number of possible models from which we must find an optimal model increases greatly, depending on the number of categories involved in creating the interaction terms. There are situations, therefore, in which we have to not only shorten variable names but also combine certain categories together in a meaningful way in order to reduce the number of models we have to analyze. We illustrate how to do this with the EnPact data spreadsheet:

1. Shorten the variable name “EducLev” to “Ed” by retyping directly in cell B3
2. At the top of a blank column just to the right of the Salary column, type the variable name “Female” (do not use quotes). This variable will be a discrete numerical variable with values 0 and 1 to indicate the employee’s gender. If Female has value 1, we have a female employee, whereas if Female has value 0 we have a male. We can do this in Excel by placing the following conditional statement in the first data cell of our new Female variable: =IF(F4=”Female”,1,0). Then we sweep down the column.

3. Generate one categorical/dummy variable based on the categorical variable JobGrade, so that if JobGrade is above 4, the dummy variable is scored as “True” and otherwise it is “False” similar to what is shown in figure 10.3. You may want to simplify the variable names if your software generates long variable names. For example, you could name it “HiJob” and code it as “True” or “False”. HiJob has value 1 (True) if JobGrade is 5 or 6 (this designates a higher level job) and has value 0 (False) if JobGrade is 1, 2, 3, or 4 (this designates a lower job level).
4. Convert “Ed” to a set of dummy variables, Ed1, Ed2, Ed3, and so forth. See figure 10.5.

	A	B	C	D	E	F	G	H	I	J
3	Employee	Ed	JobGrade	Age	YrsExp	Gender	YrsPrior	Salary	Female	JobGrade_GT4
4	1	3	1	26	3	Male	1	35.4	0	0
5	2	1	1	38	14	Female	1	41.6	1	0
6	3	1	1	35	12	Female	0	35.8	1	0
7	4	2	1	40	8	Female	7	34.1	1	0
8	5	3	1	28	3	Male	0	31.9	0	0
9	6	3	1	24	3	Female	0	33.1	1	0
10	7	3	1	27	4	Female	0	32.8	1	0

Figure 10.3: Steps 1, 2, and 3 of example 7 illustrated.

	A	B	C	D	E	F	G	H	I	J
3	Employee	Ed	JobGrade	Age	YrsExp	Gender	YrsPrior	Salary	Female	HiJob
4	1	3	1	26	3	Male	1	35.4	0	0
5	2	1	1	38	14	Female	1	41.6	1	0
6	3	1	1	35	12	Female	0	35.8	1	0
7	4	2	1	40	8	Female	7	34.1	1	0
8	5	3	1	28	3	Male	0	31.9	0	0
9	6	3	1	24	3	Female	0	33.1	1	0
10	7	3	1	27	4	Female	0	32.8	1	0

Figure 10.4: Step 3 of example 7 completed.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
3	Employee	Ed	JobGrade	Age	YrsExp	Gender	YrsPrior	Salary	Female	HiJob	Ed_1	Ed_2	Ed_3	Ed_4	Ed_5
4	1	3	1	26	3	Male	1	35.4	0	0	0	0	1	0	0
5	2	1	1	38	14	Female	1	41.6	1	0	1	0	0	0	0
6	3	1	1	35	12	Female	0	35.8	1	0	1	0	0	0	0
7	4	2	1	40	8	Female	7	34.1	1	0	0	1	0	0	0
8	5	3	1	28	3	Male	0	31.9	0	0	0	0	1	0	0
9	6	3	1	24	3	Female	0	33.1	1	0	0	0	1	0	0

Figure 10.5: Step 4 of example 7

10.2.3 Exploration 10B: Complex Gender Interactions at EnPact

1. Simplify the variables in the EnPact data file (**C11 EnPact Data**) until your data spreadsheet looks like the spreadsheet in Step 2 of example 7. By this simplification of our data, we now have only one categorical variable, Ed, with 5 categories. Female and HiJob are now discrete numerical variables with values 0 or 1. This is important to know when we create interaction terms in the next part. We will use Ed1, high-school graduate, as the reference category when we begin building our models.
2. Create the following interaction variables: YrsExp*HiJob, Female* YrsExp, Female* YrsPrior, Female* HiJob, Female*Ed. You may need to be careful when constructing regression models, to be sure that you avoid using any reference categories (e.g, do not select Female*Ed1 since it is the reference category for the Female*Ed categorical variable.)
3. Create a regression model using the following variables and interaction variables:
Base Variables: YrsExp, YrsPrior, Female, HiJob, Ed2, Ed3, Ed4, Ed5
Numerical-Categorical Interactions: YrsExp*HiJob, Female* YrsExp, Female* Age, Female* YrsPrior
Categorical-Categorical Interactions: Female* HiJob, Female* Ed2, Female* Ed3, Female* Ed4, Female* Ed5
4. Explain what goes into determining salary at EnPact and what role gender plays in the salary structure in terms of experience, education and job level. Then give a thumbnail description of life at EnPact for women.

10.3 Homework

Mechanics and Techniques Problems

10.1. Suppose you have a data file with one response variable, Wait Time, that measures the time for an order to be delivered to a customer at Beef 'n Buns. The data includes the following explanatory variables:

- Time: M(orning), D(ay), E(vening)
- Cost: Price of the order, in dollars
- Venue: C(ounter), D(rive-through window)
- Drinks: number of drinks included in the order

List all the possible interaction terms that could be created from two different variables. Organize your list by which base variables were used to construct each.

10.2. Bring up the data file **C10 Laptops**.

1. Change the variable names “Manufacturer” to “Manu” so that interaction terms will be short and still meaningful.
2. Form dummy variables for the categorical variable Manu.
3. Create interaction terms for Manu*Wt

10.3. Using the modified data file from the previous problem, construct a multiple regression routine with Price as the dependent variable, using the following explanatory variables:

- The numerical variable Weight
- The dummy variables for the categorical variables Manu
- The dummy variables for the interaction terms for Manu*Wt
- Let Sony be the reference category for Manu, if your software allows you to select the reference category. Reminder: this choice of reference category for Manu automatically determines the reference category for Manu*Wt.

Explain your model by breaking it apart into one model equation for each possible combination of factors.

Application and Reasoning Problems

10.4. Using full regression equation for the Price of a laptop, based on your work in the previous problems:

1. What is the predicted price of each of the following types of laptops?

Model of Laptop	Equation to Predict Price
Sony	
Compaq	
Hp	
Toshiba	

2. Explain how a computer brand's weight affects its price. Do heavier computer brands cost more? Or less?

10.5. Interpret the following model related to the laptop prices in the previous example:

$$\text{Price} = 560 + 115 * Wt + 230 * \text{ManuToshiba} * Wt$$

Part IV

Analyzing Data with Nonlinear Models

In Unit One we began to see the world as data; in Unit Two we began to ask questions of data in order to find out the story it has to tell about itself, and hence about the world from which it was extracted. In Unit Three we began to make connections between sets of data, to see how the events in the situations from which the data were extracted might be related to each other. We began to analyze the relationships between sets of data by capturing those relationships in regression models, simple linear ones at first involving a dependent variable and a single explanatory variable, and then more complex linear ones with a dependent variable and several explanatory variables. This unit investigates one of the four assumptions that underlie regression modeling and at the same time seeking to develop the relationships between even more complex sets of data.

One of the main assumptions about data when you construct a regression model is that the data is sampled from a linear relationship of some sort (either two-variable or more than two-variables). If this is not true, then your resulting regression model may seem to be okay, but it will exhibit problems of one of the following types:

1. The model may be accurate for only a small slice of data. If we apply the model to data points outside this small slice, the resulting errors from the model may become larger and larger. This is related to having too small a sample of the data to notice that it really does not exhibit linearity.
2. The regression model consistently underestimates the data in certain regions and consistently overestimates it in other regions. This resulting pattern indicates that there is a better model for the data than a linear model.

In chapter 11 we begin dealing with data that is not proportional, that is, data that violates our first regression assumption that a linear model is an appropriate fit. We will start by focusing on two-variable data and then learn how to extend this to multivariable data. Even though most real data sets are multi-dimensional, there are solid reasons for beginning our study with two-variable nonlinear data sets:

- Not all data is multidimensional - sometimes two variables are enough.
- Even in multidimensional data, we are often interested in the main effect first. That means looking at how the most significant variable relates to the dependent variable.
- In many modeling applications, the data shows one dependent variable and two independent variables with a constraint (like total cost must be less than a fixed amount). In this case, the constraint relationship between the two independent variables can be used to reduce the number of independent variables to one, making the entire data set two dimensional.
- Finally, the models we are going to discuss are easy to picture in two dimensions; in more dimensions, it is difficult to picture the models and develop an intuitive feel for what they can do. But the intuition we develop with two-variable data will help us interpret the diagnostic graphs in the regression output when we are dealing with multidimensional models.

In much the same way that straight lines have parameters that can be chosen so as to match the line closely to the data, the basic nonlinear models we will introduce have parameters that can serve the same purpose. By using these parameters to shift one of the basic models horizontally and vertically and to stretch them and flip it, we can fit this basic function to a non-proportional data set.

However, the regression routines in most software are only useful for producing linear models. In fact, we overcome this problem by transforming nonlinear data so that it becomes suitably linear and then applying our regression model to this straightened out data. Thus, chapter 12 presents the key transformations that will convert many kinds of nonlinear data into linear data. This chapter also teaches us how to evaluate the quality of models built from transformed data and then how to interpret these models. The unit closes with chapter 13 on interpreting the relationships in nonlinear models with more than one variable. We also discuss how to locate the maxima and minima of such functions.

Key Communication Strategy:

Memo Problem: StateEx Contract

To: Analysis Staff
From: Project Management Director
Date: August 9, 2017
Re: Shipping and unloading process at StateEx

As you may recall from the last project you worked on, the linear models for the unloading times at StateEx were okay, but not perfect. The manager has retained us to explore whether there are any nonlinear models that might be useful for describing the data, with particular emphasis on interpretation he wants for the model to not only be good, but also useful and sensible.

While there are many variables in the data and many possible models we could investigate, we don't have enough time to try everything, so I'm giving you some guidelines, based on my years of experience. First, focus on the numerical data. After we get that part refined, you can try to add in categorical data. Second, in each model you try, the same thing should be done to each of the explanatory variables. In other words, either leave them all linear, square them all, log them all, etc. Don't try mixing different nonlinear models; there's too many combinations for the time allotted. Third, don't forget that the response variable can also be transformed, but either log it or leave it alone. (Or, should we point them to a multiplicative model right up front and have them log everything, determine the full model, and then remove variables one at a time to get the significant ones? This would still involve a lot of work and analysis.)

Now, when it comes to refining the nonlinear, numerical multivariable model by adding categorical data, remember your earlier work on this project, as it may have all the clues you need for deciding which categorical data to include. Your final report should have equations for each model you test, analysis of the quality and predictive power of the models, and explanations of what the models mean for both the best nonlinear, numerical-only multivariate model and the best nonlinear multivariate model with categorical data.

Attachments: Data File `StateEx.Deliveries`

CHAPTER 11

Graphical Approaches to Nonlinear Data¹

The basic idea of this chapter is that not all relationships are linear. In fact, many of the most commonly occurring relationships come from other families of functions such as exponentials or polynomials. In this chapter, we'll explore the shapes of these different functions and learn how to control their shapes through the parameters of the model. Then, we will put our knowledge of nonlinear models to work in chapter 12 to build and interpret nonlinear regression models.

- Section 11.1 introduces you to some of the most useful nonlinear models: logarithms, exponentials, and power functions.
- Section 11.2 shows you how to visually shift, scale, and reflect these basic nonlinear models to fit a variety of data.

¹©2017 Kris H. Green and W. Allen Emerson

As a result of this chapter, students will learn

- ✓ Know what the parameters, constants and coefficients in a model are
- ✓ Know the basic shapes for each of the basic non-proportional models of interest (logarithmic/log, exponential, square, square root and reciprocal)
- ✓ Logs and exponentials are inverse functions to each other

As a result of this chapter, students will be able to

- ✓ Select and justify a choice of non-proportional model from among several possible candidates
- ✓ Choose an appropriate non-proportional model based on a scatterplot
- ✓ Determine something about the parameters of a model from looking at a scatterplot
- ✓ Shift the graph of a model around in order to make it better fit the data
- ✓ Stretch the graph of a model in order to make it better fit the data

11.1 What if the Data is Not Proportional

Our first assumption when modeling data using regression is that the data is based on an underlying linear relationship. Such relationships are said to be proportional: if the x data increases by a certain amount, the y data increases by a fixed constant times that same amount. The fixed constant relating the x -variable changes to the y -variable changes is called the slope of the linear model.

For many sets of data, however, the assumption of linearity is quite false. For example, the amount of electricity used in a house is related to the size of the house; larger houses are more expensive to heat or cool, so they tend to use more electricity. However, this relationship does not mean that doubling the size of the house always doubles the electricity costs. Much of the electricity use comes from lights, computers, televisions, and radios. No matter how much bigger the house, a family of four can only use so many of these devices at one time. So while the cost may increase, we might expect a more dramatic increase in electricity use when comparing a small house to a medium house, but a much less dramatic increase when comparing a medium-sized house to a large house. This implies that the slope of the model relating the electricity costs (y) to the size of the house (x) would be different for large houses than for small houses. In a linear function, this slope must be the same, regardless of the x -value being considered.

11.1.1 Definitions and Formulas

Non-proportionality Any model relating two variables (say x and y) in such a way that changes in one variable are not in a constant ratio to the changes in the second variable is said to be non-proportional. Another way of describing this is by saying that there is no constant k for which the following relation is true:

$$y_2 - y_1 = k(x_2 - x_1)$$

In the mathematical world and in the real world, most models are non-proportional.

Level-dependent Any model that is level dependent is also said to be non-proportional.

The term level-dependent emphasizes that with such models, the amount that the y variable increases for a given increase in x is different if the starting point (x value or location along the horizontal axis) is moved. In other words, you can look at different x and y values and compute their differences. When we compare them, if we find that $y_2 - y_1 = k_{12}(x_2 - x_1)$ and $y_4 - y_3 = k_{34}(x_4 - x_3)$, but the k values are different, then the model is level-dependent and represents a non-proportional relationship.

Concavity Concavity is a property of non-proportional models. It refers to the amount that the graph of the model bends. If the graph bends upward, that part of the graph is said to be “concave up”. If the graph bends downward in a certain area, then the graph is “concave down” in that area. Remember: concave up looks like a cup; concave down looks like a frown.

Basic function One of the six functions listed below as prototypes for fitting nonlinear data:

- linear,
- logarithmic,
- exponential,
- square,
- square root, or
- reciprocal.

In general, a function is a mathematical object that takes an input, usually in the form of a number or a set of numbers, and gives an output number. (There are other types of functions possible, but we will concentrate on functions that satisfy this definition.) For a relationship between two variables, say x and y , to be a function, it must satisfy the following statement:

Every x-value must be associated with one and only one y-value.

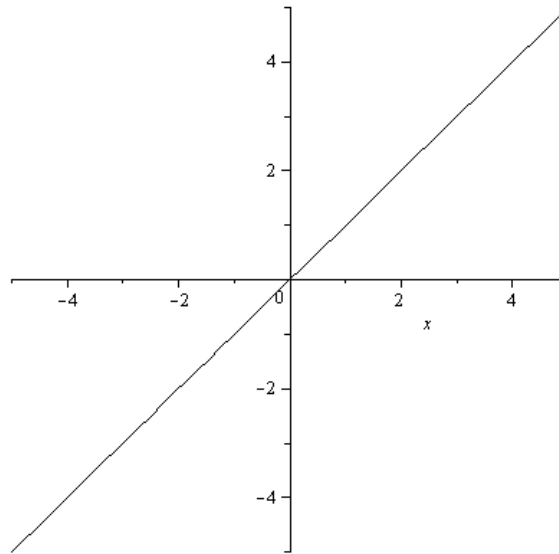
This means that if you draw a graph of the function, and draw a vertical line through any point on the graph, that line will only touch the graph once. This is sometimes referred to as the vertical line test. Generally, if the variable y is a function of the variable x , we write $y = f(x)$ to indicate this. If the variable y is a function of several variables (say x_1, x_2, x_3) then we write $y = f(x_1, x_2, x_3)$.

Linear function Graphs of linear functions (see figure 11.1) are straight lines. The prototypical, or base, form of a linear function that related y to x is given by $y = x$. You are more used (by this point) to seeing this in a more general form, involving two parameters, the slope and y-intercept: $y = A + Bx$. The graph of a linear function is shown below. Notice that linear functions are straight; they have no concavity at all.

Logarithmic function A logarithm (see figure 11.2) is a mathematical function very useful in scaling data that spans a large range of values, like from 1 to 1,000,000 (we will see this aspect of logarithms in a later chapter). In general, there are lots of different logarithmic functions. We will be using the natural logarithm of x as a function; this is written as $y = \ln(x)$. (Notice: **natural logarithm** = **nl** \rightarrow **ln**.) The graph of the basic natural logarithm is shown below. The basic logarithm is increasing and concave down everywhere.

The natural logarithmic function has several important properties to note. The natural log of 0 is undefined; in other words $\ln(0)$ does not exist. If $0 < x < 1$ then $\ln(x) < 0$, and $\ln(1) = 0$. This means that the point $(1, 0)$ is common to all basic log functions. This is actually a restatement of the fact that any base raised to the zero power is equal to 1.

Exponential function Exponential functions (see figure 11.3) are related to logarithmic functions. These can be written in two ways. The first form is as a base number

Figure 11.1: The basic linear function $y = x$.

raised to a variable power ($y = a^x$). The most common base to use is the number e , which is approximately 2.71828... In reality, e is an irrational number, like π . It shows up naturally in many situations, as we will see in example 4 from chapter 15 when examining interest rates. For now, though, the standard exponential function we will use is $y = e^x$. The second form is similar to this, but easier to type: $y = \exp(x)$. This can be read as “ y is the exponential function of x ” or “ y equals e raised to the x power.” Its graph is shown below. The basic exponential function is increasing and concave up everywhere.

In addition, since any positive number, like e , raised to a negative power is a number between 0 and 1, we know that if $-\infty < x < 0$ then $0 < e^x < 1$. Since any number raised to the zero power is 1, we also know that $e^0 = 1$ so the point $(0, 1)$ is on the graph of all basic exponential functions.

Square function You are probably familiar with the squaring function: it takes every number put into it and spits out that number raised to the second power. Thus, if we stick in the number x , we get out x^2 . Thus, the basic squaring function is $y = x^2$. The graph of this function has a special name that you may have heard before: a parabola. It looks like the letter “U”, centered at $(0, 0)$. The basic squaring function is concave up everywhere. as shown in figure 11.4.

Square root function The square root function does the opposite of what the squaring function does. This function takes in a number and spits out its square root. The square root of a number is that number which, when squared, produces the number. For example, 2 is the square root of 4, since 2×2 is 4. The square root function is usually written as $y = \sqrt{x}$. Another way to write the function reminds us of its relationship with the squaring function: $y = x^{1/2} = x^{0.5}$. (Read this as: y is x raised

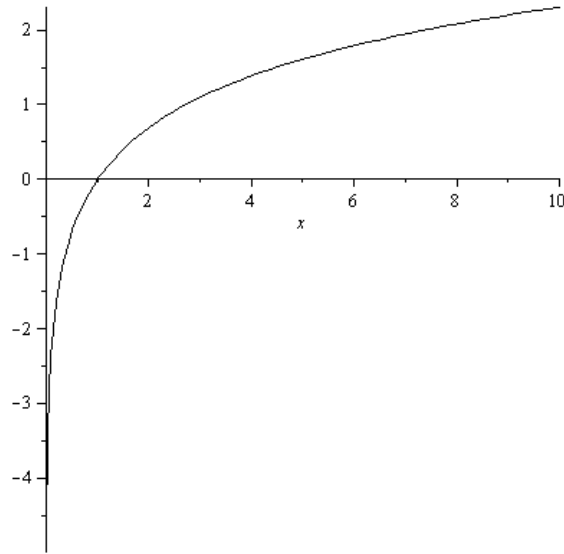


Figure 11.2: The basic logarithmic function $y = \ln(x)$.

to the one-half power or x to the 0.5 power.) The basic square root function is concave down everywhere. In figure 11.5, the square root function is not graphed for values of x less than 0, since the square root of a negative number is an imaginary quantity.

Reciprocal function The reciprocal function takes a number and returns one divided by that number: $y = \frac{1}{x}$. This function also has an alternative form in which x is raised to a power: $y = x^{-1}$. Notice that the reciprocal function shown in figure 11.6 has several interesting features: It has different concavity on the left and the right; it does not even exist at $x = 0$ since any number divided by zero is undefined; in fact, the reciprocal function never crosses either axis.

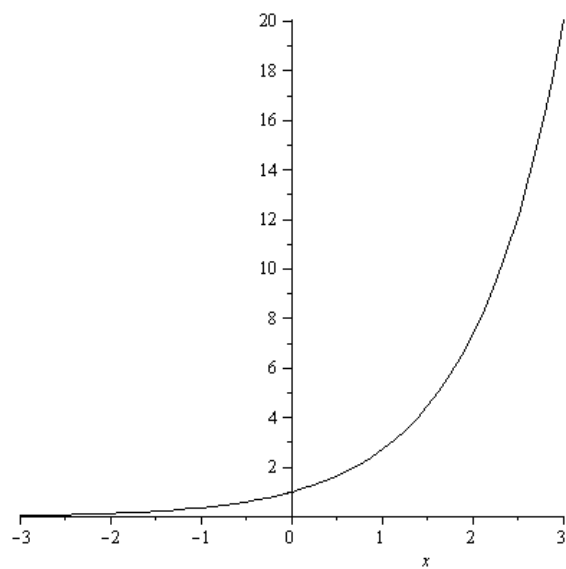


Figure 11.3: The basic exponential function $y = \exp(x)$.

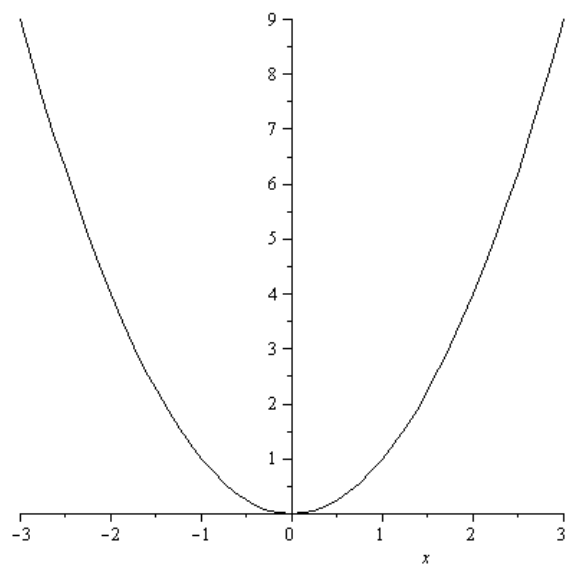


Figure 11.4: The basic squaring function $y = x^2$.

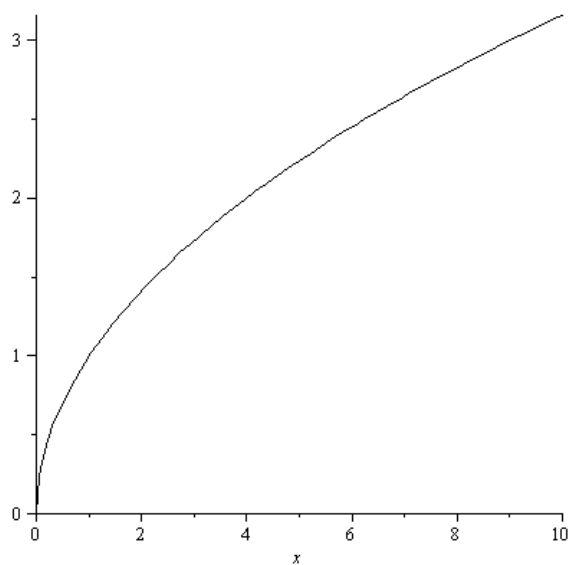


Figure 11.5: The basic square root function $y = \sqrt{x}$.

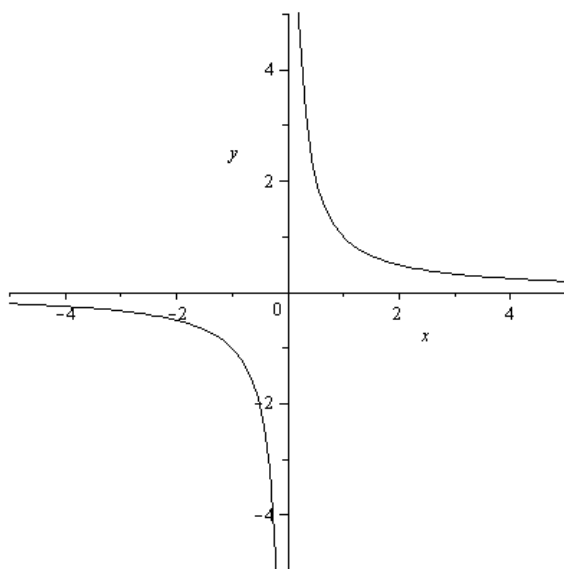


Figure 11.6: The basic reciprocal function $y = x^{-1} = \frac{1}{x}$.

11.1.2 Worked Examples

Example 11.1. Using a graph of the data to see nonlinearity

Consider the data graphed below. What can we say about it? It appears that as x increases, the y values decrease. It also looks like the data is bending upward. Mathematicians call this behavior “concave up”. Let’s see what happens when we apply a linear regression to these data.

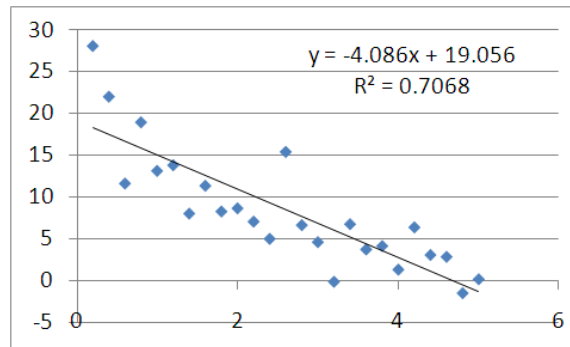


Figure 11.7: Does a linear function fit these data well?

It looks like a good candidate for fitting with a straight line, and the R^2 value is acceptable for some applications, but notice there are distinct patterns in the data, when compared to a best fit line. On the left, the data is mostly above the line, in the middle, the data is mostly below the line, and on the right, the data is mostly above the line. Patterns such as this indicate that changes in the Y data are not proportional to changes in the x data. To put this another way, the rate of change of y is not constant as a function of x , which means there is no single “slope number” that is the same for every point on the graph. Instead, these changes are level-dependent: as we move our starting point to the right, the y change for a given x change gets smaller and smaller. In a straight line, this is not the case: regardless of starting point, the y change for a given x change are the same. If the data are best represented by a linear model, we would not see any patterns in the data points when compared to the model line; the points should be spread above and below the best-fit line randomly, regardless of where along the line we are. For the graph above, though, we do see a pattern, indicating that these data are not well suited to a linear model.

Notice that R^2 by itself would not have told us the data is nonlinear, because the data is tightly clustered and has little concavity. Clearly, the more concave the data is, the worse R^2 will be for a linear fit, since lines have no concavity and cannot capture information about concavity.

Example 11.2. Comparing logarithmic models and square root models

You may have noticed that two of the functions above, the square root and the logarithmic, look very similar. Why do we need both of them? After all, the two graphs (see figure 11.8) have very similar characteristics. For example, both start off very steep for small values of x

and then flatten out as x increases. Both graphs continue to increase forever. Neither graph exists for negative values of x .

However, the graphs are actually quite different. For instance, consider the origin. The point $(0,0)$ is a point on the square root graph (since the square root of zero is zero,) but it is not a point on the logarithmic graph. In fact, if you try to compute the natural log of zero, you will get an error, no matter what tool you use for the calculation! The logarithmic function has what is called a “vertical asymptote” at $x = 0$. This means that the graph gets very close to the vertical line $x = 0$, but never touches it. This is quite different from the square root graph which simply stops at the point $(0,0)$. Furthermore, the square root has a horizontal intercept of $x = 0$, while the logarithmic graph crosses the x -axis at $(1,0)$.

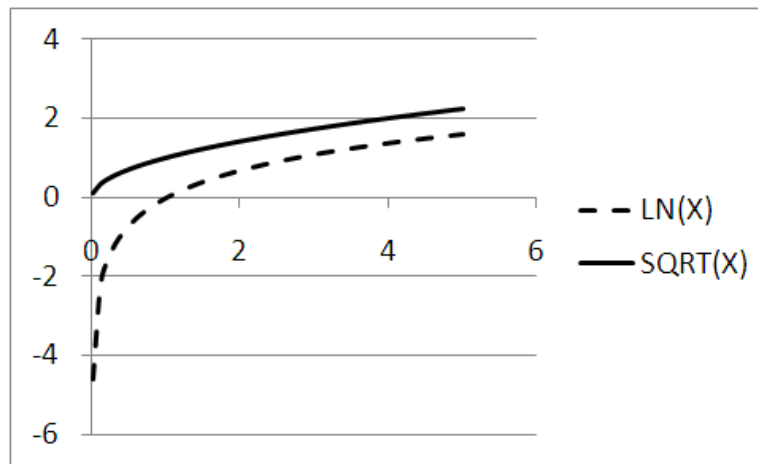


Figure 11.8: Comparison of standard log and square root functions.

You might be tempted to think that we could simply “move” the logarithmic graph over so that they both start at the same place, $(0,0)$. Figure 11.9 shows what happens if we pick up the graph of $y = \ln(x)$ and move it to the left one unit, so that both graphs pass through $(0,0)$. Notice that the square root graph rises sharply and then flattens, while the natural log graph rises more gradually. It also appears that the slope of the square root graph is larger and that it continues to grow larger, widening the gap between the two functions. In fact, the natural log grows so slowly that the natural log of 1,000 is only 6.9 and the natural log of 1,000,000 is 13.8! Thus, if the x -values of your data span a large range, over multiple orders of magnitude, a natural log may help scale these numbers down to a more reasonable size. This property of logs makes them useful for measuring the magnitude of an earthquake (the Richter scale) or the loudness of a sound (measured in decibels). Compare this growth to the square root function: The square root of 1,000 is about 31; the square root of 1,000,000 is 1000. This is a much larger increase than the natural log. In fact, of all the basic functions, the natural log is the slowest growing function; in a race to infinity, it will always lose.

Example 11.3. Comparing exponential models and square models

You may have also noticed that, for positive values of x , the graphs of the exponential

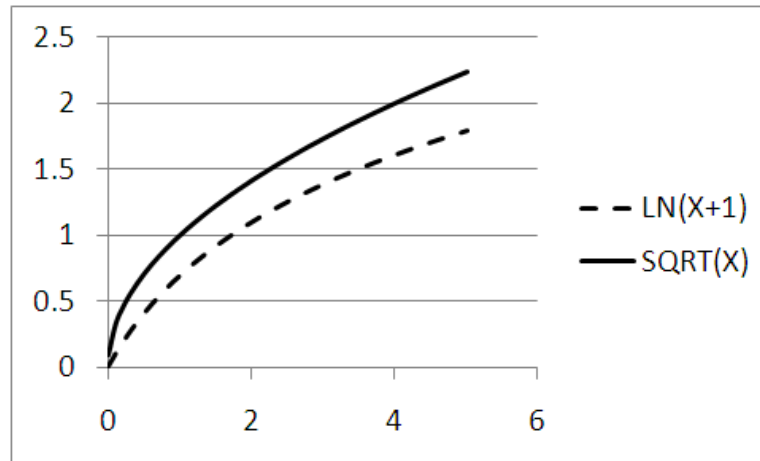


Figure 11.9: Comparison of horizontally shifted log and square root function.

function and the square function are very similar. Both are increasing. Both are growing large at a faster and faster rate, which shows in the graphs from the increasing steepness of each graph as x grows larger. This property of getting larger at an increasing rate is referred to as being “concave up.” This makes the graphs bend upward, away from the x -axis so that it looks like a cup that could hold water. Both graphs also start rather flat near the origin.

Here is where the similarities end, though. The square function has a vertical and horizontal intercept at $(0, 0)$. The exponential function, on the other hand, has a vertical intercept of $y = 1$, but no horizontal intercept at all. Much like the logarithmic function (see the previous example) the exponential function has an asymptote. In this case, though, it is a horizontal asymptote at $y = 0$, rather than a vertical intercept at $x = 0$. In addition, when we look at the graphs for negative values of x , we see that the exponential function is always increasing, while the square function is decreasing for $x < 0$. This means that the square function has a minimum, or lowest, point.

These properties are also easy to see numerically from working with the functions themselves. If I take a negative number and square it, I get a positive number. Thus, $(-3)^2 = +9$, $(-2)^2 = +4$, $(-1)^2 = 1$, etc. Notice that as the negative values of x get closer to 0, the output of the square function is decreasing. For an exponential function, we notice negative exponents are really a shorthand way of writing “flip the function upside down and raise it to a positive power.” Thus, to compute e^{-2} , we compute $1/e^2 \approx 1/7.3891 \approx 0.1353$. This is where the asymptotic nature of the exponential function shows through; for large negative powers, we are really computing one divided by e raised to a large positive power. Since e to a large positive power is a large positive number, one over this number is very small and close to zero.

As it turns out, the exponential function is the fastest growing of all the basic functions. In a race to infinity, it will always win.

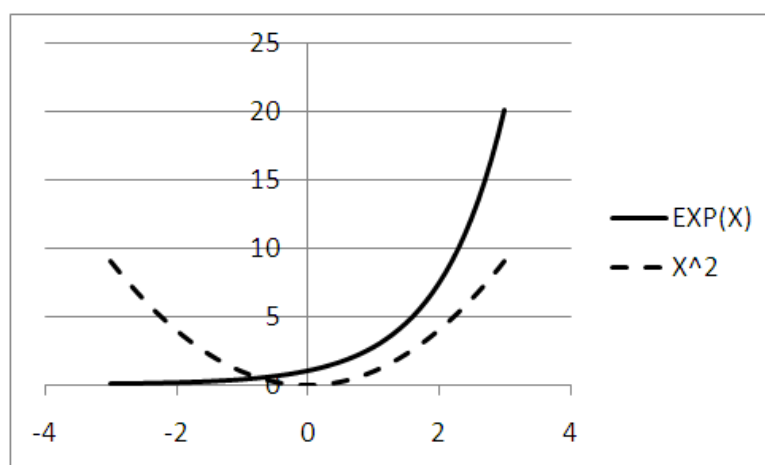
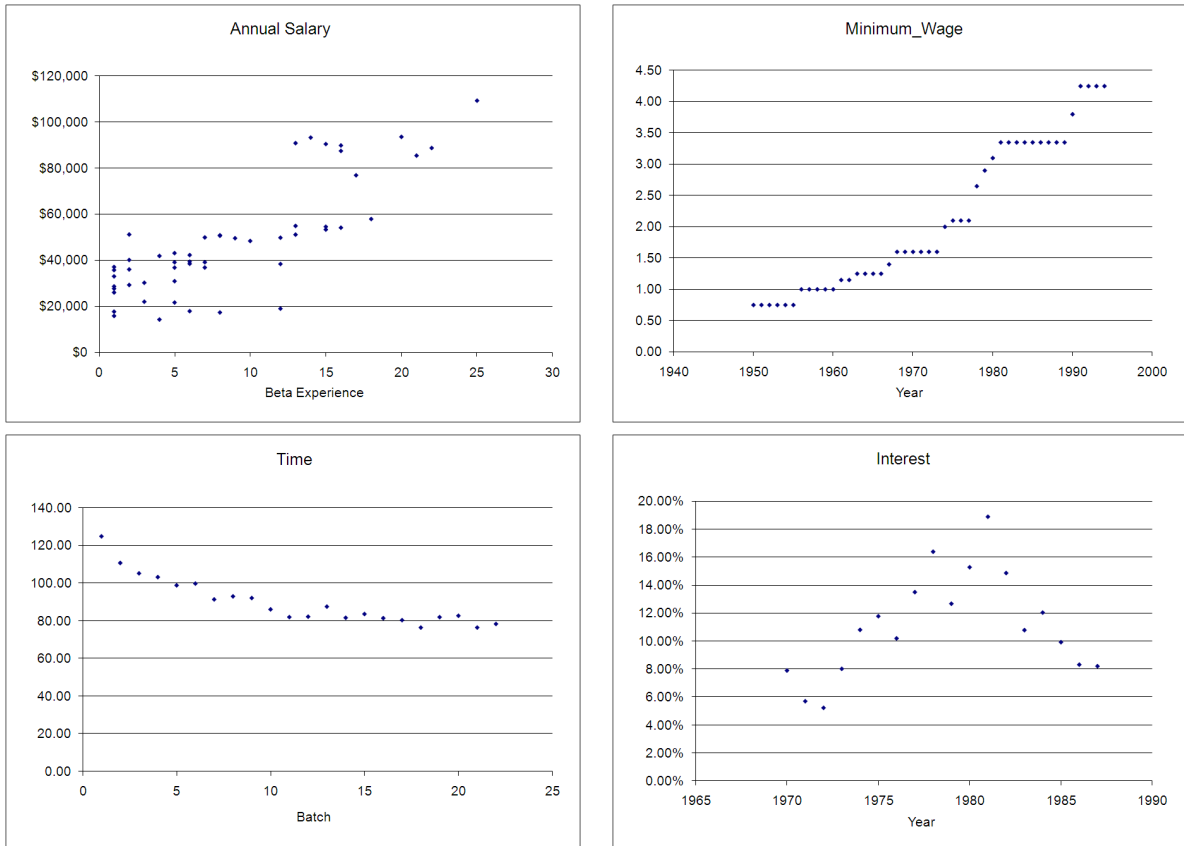


Figure 11.10: Comparison of exponential and squaring functions.

11.1.3 Exploration 11A: Developing our intuition about data that is non-proportional

Four graphs are shown below. For each, consider which of the basic functions you think would fit the data best. For your function, then describe what you think should be done to it in order to make it fit the data best. It might need to be shifted left or right, sifted up or down, flipped vertically or horizontally, stretched or squashed, or some combination of these.



Data	Best choice for basic function	How to alter the function to fit
Annual Salary		
Minimum Wage		
Time		
Interest		

Now open the file **C11 Exploration2**. Test each of the possible trendlines (linear, logarithmic, exponential, power and polynomial of order 2 - do not use the moving average or higher order polynomials.) Be sure you display the equation and R^2 value for each of the possible models. Write down the equation of the best fitting model and record its R^2 in the chart below.

Data	Best Fit Trendline Equation	R^2
Annual Salary		
Minimum Wage		
Time		
Interest		

11.2 Transformations of Graphs

The basic functions introduced in the last section are very useful. With these, we can fit a lot more data than we could with just straight lines. However, we often find that the data matches the shape of one of these basic functions, but not the specific location and specific points that the basic function passes through. Consider the data shown on the left scatterplot below. The data shows the number of products (in this case motors) returned from a production line as a function of the amount of money spent on inspections in a given month (in thousands of dollars). On the right is a graph of this same data, but with the graph of the basic square function superimposed on it. They have the same shape, but are not exactly the same: The square function starts too low and rises much more quickly than the actual data, and they are not in the same place on the graph.

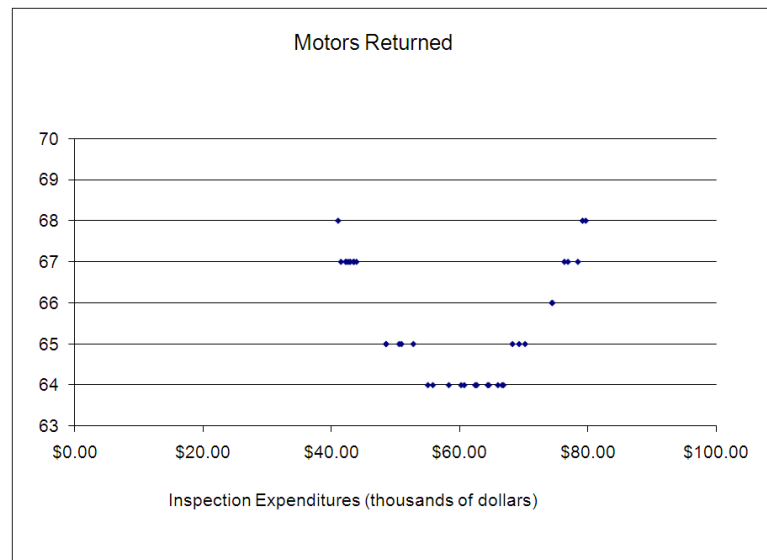


Figure 11.11: Graph of number of motors returned versus inspection expenditures.

The way to fix this is to “move the basic function around” until it fits more closely. This is very much like what we did with straight lines before: we knew we wanted a straight line, but we needed to change the slope and y-intercept until the theoretical line matched up better with the data. For the data above, it looks like we need to “squash” the square function down and move the starting point over and up. In this section, we’ll explore the mathematical way of doing this. When we are done, we will have developed formulas for the basic functions that are more general and contain several parameters (almost always two). These parameters are, in general, much harder to interpret than the parameters in a linear model, but once we understand how they affect the graphs, we’ll be able to out some meaning to them.

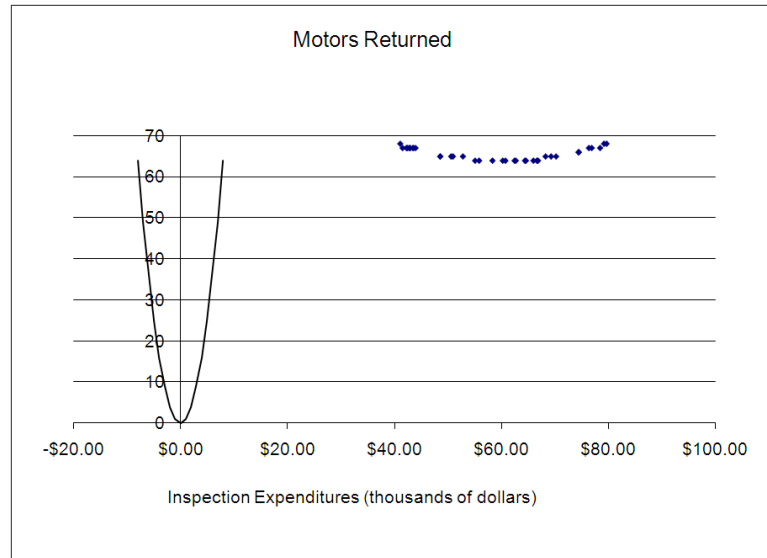


Figure 11.12: Graph of inspection expenditures with poorly fitting square model added.

11.2.1 Definitions and Formulas

Parameters A parameter is a number in the formula for a function that is constant. Changing a parameter will change the entire behavior of the function. The two parameters you are most familiar with are the slope and y-intercept of a linear function. If the slope parameter is changed, the line is more or less tilted; it may even change the direction of the tilt. If the y-intercept is changed, the graph crosses the y-axis at a different point. Most functions come in families of functions that all have the same formula, but the formula has parameters in it. Thus, linear functions of the form should really be called the “family of linear functions” since there are two parameters in the formula. To get the equation of a specific member of the family, we need to substitute in values for each of the two parameters, A and B. (Just like you need a first and last name to find a specific person in your family; you may sometimes need even more information about the person if more than one person in the family has the same name. Some functions also need more than two parameters. See quadratics below for such an example.)

Power functions This is a broad family of functions. The general form of a basic power function is where b is a number. Thus, this family includes the squaring function ($b = 2$), the square root function ($b = \frac{1}{2}$), the reciprocal function ($b = -1$), and the basic linear function ($b = 1$). This family is called the family of power functions because the independent variable, x , is always raised to a power. The shape of a power function depends on whether the power, b , is even or odd. Even power functions look something like a “U” when graphed. Odd power functions (with $b > 1$) look more like chairs: on the left they drop off; on the right they rise up high; in the middle they are relatively flat. All basic power functions pass through the origin $(0, 0)$ and the point $(1, 1)$. This is because zero raised to any power is zero and 1 raised to a power is always 1.

Polynomials A polynomial is a function made from adding together a bunch of power functions that all have whole number powers. (A whole number is a number like 5, 2, 0. Negative numbers and numbers with decimals and fractions are not allowed.) Each power function in a polynomial is multiplied by a coefficient and then they are all added together:

$$y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_2 x^2 + a_1 x^1 + a_0$$

Notice that since anything raised to the zero power is 1, there is no need to write x^0 in the last term. Each of the individual combinations of a coefficient and a power function in a polynomial is called a term. Polynomials include several well-known families of functions: the quadratics (see below) and the linear functions:

$$y = A + Bx.$$

The n in a power function gives the highest power in the polynomial. It is called the order of the polynomial. The shape of a polynomial function is highly dependent on the order of the polynomial, since this determines the leading power function in the polynomial. The following general statements can be made:

If n is even, then the polynomial function does the same thing on both sides of the y -axis: it either rises up on both sides or drops down on both sides. If n is odd, then the polynomial does the opposite on both sides: one side will rise, the other will drop. The order also determines two other properties: the maximum possible number of times the polynomial crosses the x -axis (the number of zeros) and the number of time the graph changes direction (either from increasing to decreasing or vice versa):

- Maximum number of zeros = n
- Maximum number of turning points = $n - 1$

Quadratics A quadratic function is a second-order polynomial that produces a “generalized squaring function”. It is usually written in the following way:

$$y = Ax^2 + Bx + C.$$

You may have seen the famous quadratic formula. This is a formula for finding the roots of a quadratic equation. Roots are places where the function crosses the x -axis, so these points all have $y = 0$. Thus, they are solutions to the equation:

$$0 = Ax^2 + Bx + C.$$

Using the quadratic formula, we can find the x -coordinates of these crossing points:

$$x = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

Most software can add quadratic trendlines to a graph; however, it refers to them by their more proper name as “polynomials of order 2”.

Vertical shifting Sometimes the data we are trying to fit looks exactly like a basic function, but moved up or down. We can fix this by adding in a vertical shift to the equation. If the graph of a function has been vertically shifted, the graph has the same exact shape, only every single y -value has been increased by the same amount or every y -value has been decreased by the same amount. Effectively, this moves the entire graph of the function either up or down the y -axis. Thus, a vertical shift by k will move the y -intercept up by k units.

Horizontal shifting Sometimes, the data is moved right or left of the basic function that it is most similar to. We can compensate by adding a horizontal shift to the equation of the graph. If a graph has been horizontally shifted, the graph has been moved to the right or the left. Thus, if the graph is moved to the right h units, then the zeros of the function (if any) will all move to the right h units.

Translation This is the general term to refer to any type of shift (vertical or horizontal).

Vertical scaling It is sometimes necessary to stretch a graph out or compress the graph of a basic function so that it will match up better with the data. This can easily be done by multiplying the entire function by a scaling factor.

11.2.2 Worked Examples

Example 11.4. Vertical shift

Consider the data shown in the table below for $y = f(x)$. If we make a new function by adding the same amount, say 10, to each of these y values, then we will be creating the function $y = f(x) + 10$; each y value will be 10 more than it would be without the increase. This will result in the graph of the function being shifted up by 10 units at each data point. It's just like we picked up the graph and slid it up the y -axis 10 units.

x	$y = f(x)$	$y = f(x) + 10$
0	10	20
1	15	25
2	12	22
3	3	13
4	6	16
5	11	21
6	15	25
7	19	29
8	25	35
9	23	33
10	22	32

Example 11.5. Horizontal shift

We can also shift a graph to the left or right. In the last example, we added a value to all

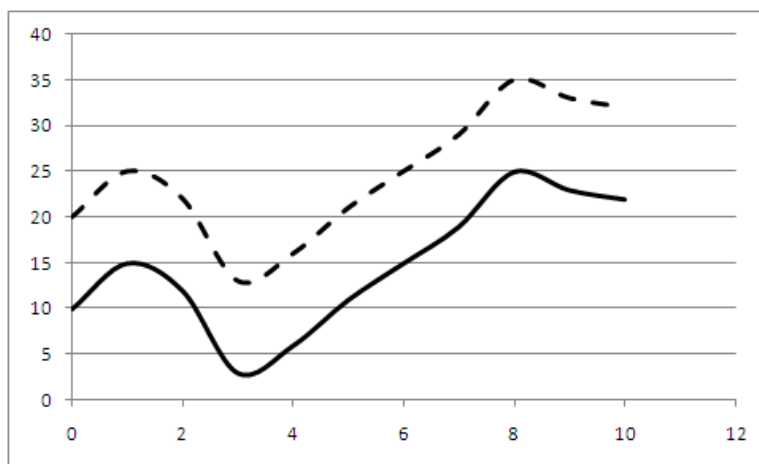


Figure 11.13: Graph of $y = f(x)$ (solid line) and $y = f(x) + 10$ (dashed line).

the y values in order to shift the graph up or down the y -axis. To shift left and right, we need to add or subtract from the x values. For example, suppose we wanted to move the graph four units to the right. The old graph would have the point (x, y) corresponding to the statement that $y = f(x)$. The new graph should have the point $(x + 4, y)$. So if the old graph had the point $(3, 3)$, the new graph should have the point $(3 + 4, 3)$ or $(7, 3)$. Here's the catch, though, the function will only give 3 for y if we plug in a value of 3 for x . We want to plug in 7 for x and get 3 out. Thus, we need to subtract 4 from each x value in order to make sure the function gives the right output. This means that to shift the function to the right 4 units, we need to plot the graph of $y = f(x - 4)$. This is shown in the data table below and the graph beside it.

x	$y = f(x)$	$y = f(x - 4)$
0	10	?
1	15	?
2	12	?
3	3	?
4	6	10
5	11	15
6	15	12
7	19	3
8	25	6
9	23	11
10	22	15
11	?	19
12	?	25
13	?	23
14	?	22

Example 11.6. Vertical scaling

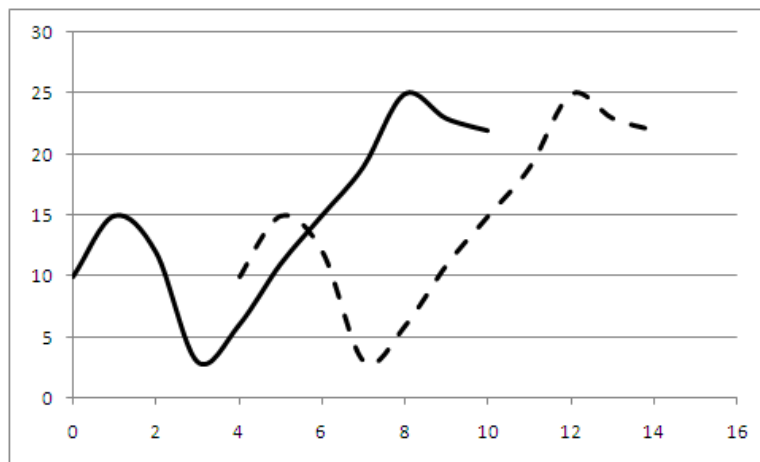


Figure 11.14: Graph of $y = f(x)$ (solid line) and $y = f(x - 4)$ (dashed line).

We can also stretch the shape of a graph out. Suppose that we have a set of data that looks parabolic, so we want to use the square function. But suppose that the data contains the points $(1, 1)$, $(2, 8)$, $(3, 18)$, and $(4, 32)$. For a basic squaring function, this can't happen; the shape is right, but $2 \times 2 = 4$, not 8; $3 \times 3 = 9$, not 18; and $4 \times 4 = 16$, not 32. But notice, that each of the actual data values is simply twice what the squaring function would give. Thus, we want to graph $y = 2x^2$. This means that we should take each value of x , compute x^2 , and then multiply the result by 2. This stretches the graph out to fit the data.

We can also compress a graph, squashing it down flatter instead of stretching it up taller. Suppose that the data contains the points $(1, 0.5)$, $(2, 2)$, $(3, 4.5)$ and $(4, 8)$. Each of the y values is half of what we would expect from the squaring function, so we want to graph $y = 0.5x^2$. We see that the general form to scale the graph of $y = f(x)$ is $y = a \times f(x)$ where a is a constant. The graphs below show the basic squaring function and the two functions we have just created.

But what happens if we let a be a negative number? This will simply take each of the old y values from the function and put a negative sign in front of them. This flips the graph over the x -axis, creating a mirror reflection of the original graph. Thus, the graph of $y = -f(x)$ is the same as the graph of $y = f(x)$ except that it is flipped over the x -axis. In a similar way, multiplying x by a factor can scale the graph horizontally, and negating x flips the graph horizontally over the y -axis.

Example 11.7. Combination of Shifts and Scales

Consider the graphs shown in the introduction to this section in figure 11.11. The data for the number of motors returned as a function of inspection expenditures looks to be a basic squaring function, but shifted and scaled. Here's another look at the graph.

It looks like the graph has been shifted to the right 60 units and up 64 units. Thus, we could start by comparing the data to the graph of $y = f(x - 60) + 64 = (x - 60)^2 + 64$. When we do this, we find that the graph starts in the right place, but climbs too quickly. We might be tempted to simply multiple this whole thing by a constant less than one in

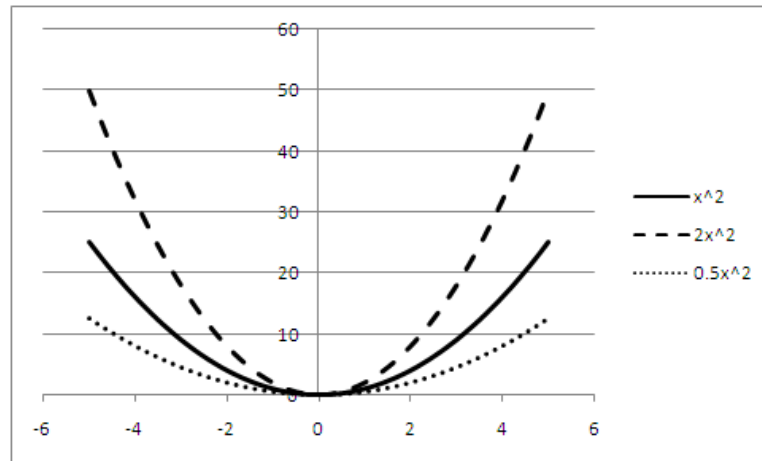


Figure 11.15: Graphs of $y = x^2$ (solid line), $y = 2x^2$ (dashed line) and $y = 0.5x^2$ (dotted line).

order to squash the graph, but this would also multiply the vertical shift by the constant, changing the starting place. We must complete the shifts and scaling in the proper order. We need to construct the fit by first shifting right, then scaling, then shifting. So, we are looking for a function of the form $y = af(x - 60) + 64 = a(x - 60)^2 + 64$.

How much should we squash the graph? In other words, how big is a ? The best approach here is to try a few data points. It looks like the point $(80, 68)$ is on the graph. Plugging these values in for x and y we get the following:

$$\begin{aligned}
 68 &= a(80 - 60)^2 + 64 \\
 68 &= a(20)^2 + 64 \\
 68 - 64 &= a(20)^2 \\
 4 &= a(20)^2 \\
 \frac{4}{20^2} &= a \\
 a &= 0.01
 \end{aligned}$$

Thus, the equation of the function that seems to match the data is $y = 0.01(x - 60)^2 + 64$, where y represents the number of motors returned, and x represents the amount of money (in thousands) spent on inspection expenditures in a given month. We should check this against a few more data points, to be certain that the function is the correct one. Since the point $(75, 66)$ also appears to be on the function, we evaluate our candidate function at this x value to see if they match. At $x = 75$ our function is equal to $y = 0.01(75 - 60)^2 + 64 = 66.25$ which is very close to the value given by the data. We don't expect a perfect fit, because the data is not taken from an abstract function, but actually came from a real situation, so there will likely be some error in the best-fit function.

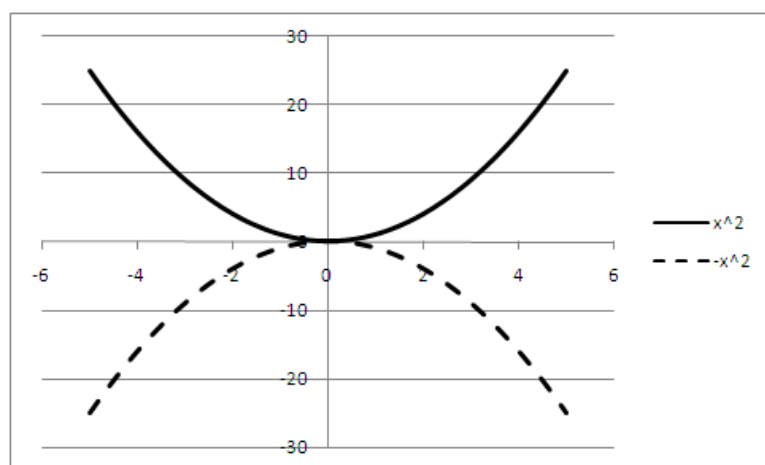


Figure 11.16: Graphs of $y = x^2$ (solid line) and $y = -x^2$ (dashed line).

11.2.3 Exploration 11B: Shifting and Scaling the Basic Models

Download and open the file C11 Exploration3. The file contains several macros, so when you open the file, you may need to click on the options button next to the security warning. Then select the option labeled “enable the content”. (This is part of the security of the computer; many viruses and computer worms are hidden in macros.) When you get the file open, you should see a screen like the one shown below:

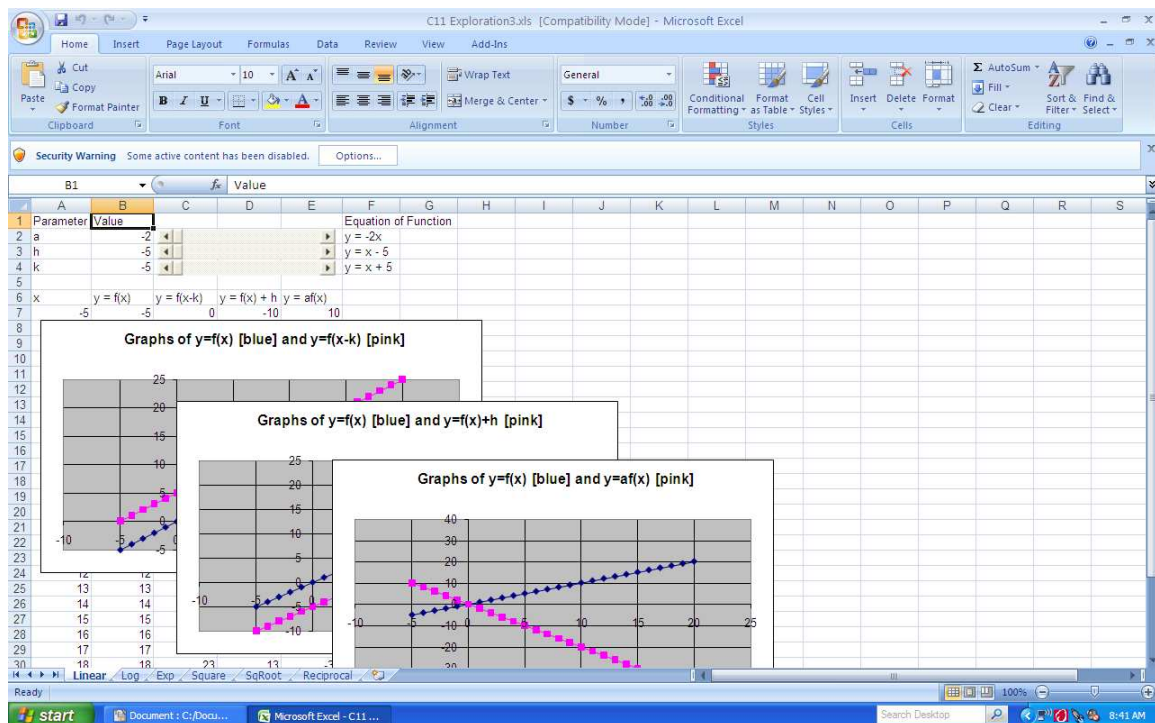


Figure 11.17: Excel file for exploring the various shifts and scalings with the basic functions.

There are six worksheets in the workbook, one for each of the basic functions we have been discussing. On each worksheet there are three slider bars and three graphs. Each graph shows the graph of the basic function itself (in blue) and one other graph (in pink). As you change the slider values, make note of how the graph of the pink function changes and how the different equations shown next to the slider bars change. To see some of the graphs, you may need to right click on them and select “Bring to Front” since Excel layers its graphs on top of each other in order to save “screen real estate”. Use the worksheets and the sliders to help fill in the details about each of the functions below.

Linear Function, $f(x) = x$		
Modification	Sketch	Description
$y = af(x)$		
$y = f(x - h)$		
$y = f(x) + k$		

Logarithmic Function, $f(x) = \ln(x)$		
Modification	Sketch	Description
$y = af(x)$		
$y = f(x - h)$		
$y = f(x) + k$		

Exponential Function, $f(x) = e^x$		
Modification	Sketch	Description
$y = af(x)$		
$y = f(x - h)$		
$y = f(x) + k$		

Squaring Function, $f(x) = x^2$		
Modification	Sketch	Description
$y = af(x)$		
$y = f(x - h)$		
$y = f(x) + k$		

Square Root Function, $f(x) = \sqrt{x}$		
Modification	Sketch	Description
$y = af(x)$		
$y = f(x - h)$		
$y = f(x) + k$		

Reciprocal Function, $f(x) = \frac{1}{x}$		
Modification	Sketch	Description
$y = af(x)$		
$y = f(x - h)$		
$y = f(x) + k$		

11.3 Homework

Mechanics and Techniques Problems

11.1. This problem deals with what happens to equations of functions and graphs of functions if you apply several different transformations, one after the other. Take $y = f(x) = x^2$ and write out each of the following functions. For each step, explain what happens to the graph in terms of how the particular change affects the appearance of the previous graph in the sequence.

Function	Written out	What happens to graph
$y = f(x) = x^2$		
$y = f(x - h)$		
$y = af(x - h)$		
$y = af(x - h) + k$		

11.2. How would the results in problem 1 be different if we changed the order to $y = a[f(x - h) + k]$?

Function	Written out	What happens to graph
$y = f(x) = x^2$		
$y = f(x - h)$		
$y = f(x - h) + k$		
$y = a[f(x - h) + k]$		

11.3. Now repeat 1 with a basic exponential function.

Function	Written out	What happens to graph
$y = f(x) = \exp(x)$		
$y = f(x - h)$		
$y = af(x - h)$		
$y = af(x - h) + k$		

11.4. Now repeat 1 with a basic logarithmic function.

Function	Written out	What happens to graph
$y = f(x) = \ln(x)$		
$y = f(x - h)$		
$y = af(x - h)$		
$y = af(x - h) + k$		

11.5. For each of the five graphs below

1. Select the best basic function to fit the data,
2. Select appropriate shifts (direction) and scaling (stretch or compress), and
3. Write down a possible equation for the graph.

11.6. Consider the data shown below in both table and graphical format.

x	y
0	2.05
5	2.69
10	3.55
15	4.23
20	4.35
24	5.08

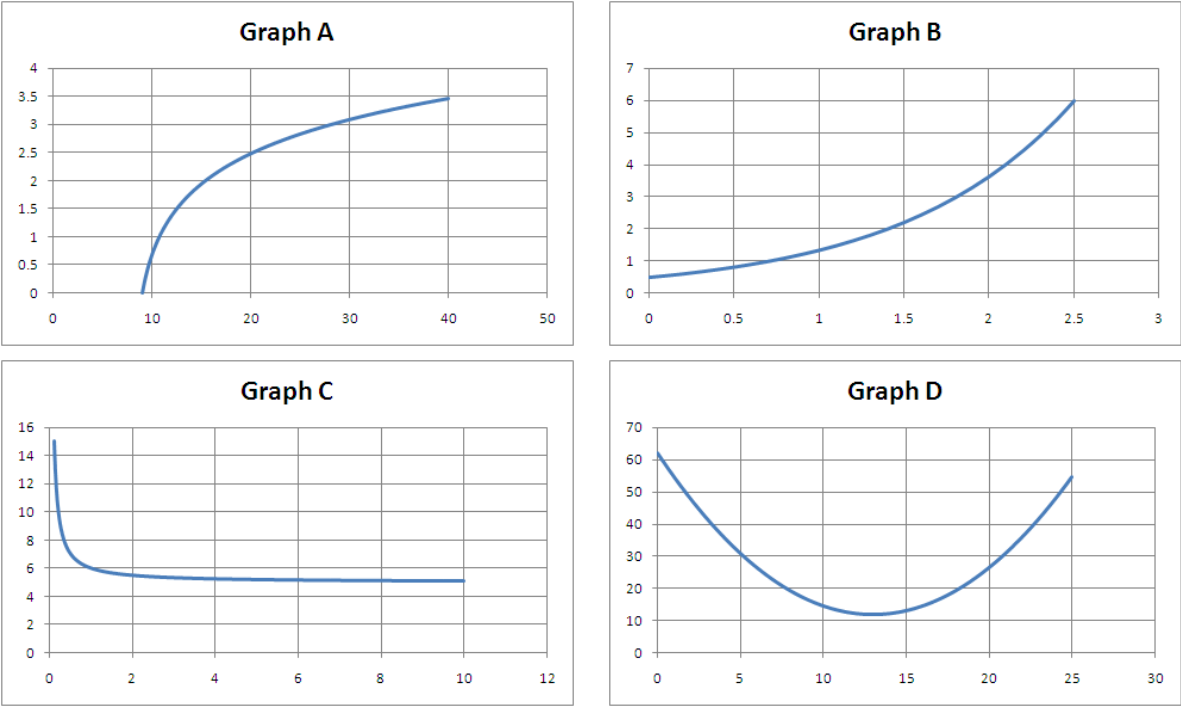


Figure 11.18: Graphs for problem 5.

1. Create a scatterplot of the data, and determine which order polynomial function (2 through 6) fits the data best. Record the results of your investigation in a table like the one shown below.

Order	Equation	R^2
2		
3		
4		
5		
6		

2. Use the parameters from your quadratic trendline (the order 2 polynomial) to manually calculate the S_e for that model. You may want to set up a spreadsheet like the one in figure 11.20.

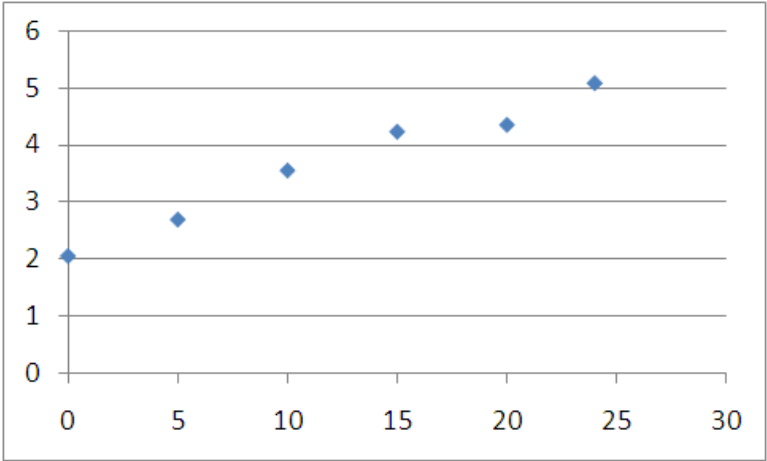


Figure 11.19: Graph of the data from problem 6.

	A	B	C	D	E	F	G	H	I
1	x	y	yhat(fitted)	Residual	Deviation		Parameters	a	
2		0	2.05					b	
3		5	2.69					c	
4		10	3.55						
5		15	4.23						
6		20	4.35						
7		24	5.08						
8									
9		ybar =>							
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									
32									
33									
34									
35									
36									
37									
38									
39									
40									
41									
42									
43									
44									
45									
46									
47									
48									
49									
50									
51									
52									
53									
54									
55									
56									
57									
58									
59									
60									
61									
62									
63									
64									
65									
66									
67									
68									
69									
70									
71									
72									
73									
74									
75									
76									
77									
78									
79									
80									
81									
82									
83									
84									
85									
86									
87									
88									
89									
90									
91									
92									
93									
94									
95									
96									
97									
98									
99									
100									

Figure 11.20: Set up to calculate R^2 and S_e in problem 6.

Application and Reasoning Problems

Coming soon

CHAPTER 12

Modeling with Nonlinear Data¹

In the last chapter, we learned a lot about different types of functions that can be used to model data when the data does not represent a proportional relationship. In this chapter, we're going to put this knowledge to use making and interpreting regression models of such non-proportional data. To do this, we need to go through a few steps.

First we transform the data using some of these functions. There are only four transformations that we need; combining them in different ways can produce all of the models we have talked about. Next we perform the regression, using these transformed variables, and some of the original variables, if needed. Unfortunately, we'll need to compute the summary measures (R^2 and S_E) by hand for some of the nonlinear models. Finally, we have to make sense of the models we get by putting them into a useful form and determining what the parameters in the model actually mean.

- Section 12.1 shows how to use transformations of data to “trick” linear regression tools into finding the best nonlinear model fits that you can make.
- Section ?? explains how to interpret the output of the regression for nonlinear models and how to determine the goodness of fit.

¹©2017 Kris H. Green and W. Allen Emerson

As a result of this chapter, students will learn

- ✓ Which transformations of the data will linearize the data
- ✓ That some summary measures are not accurate when using nonlinear regression
- ✓ That transformations of data can help to minimize non-constant variance in data
- ✓ What the parameters in each of the nonlinear models actually mean

As a result of this chapter, students will be able to

- ✓ Transform variables for use in nonlinear modeling
- ✓ Accurately compute R^2 and S_E for nonlinear models containing $\log(\text{response})$
- ✓ Transform the regression equations of nonlinear models into standard form
- ✓ Calculate the effects of changes in the explanatory variable on the response variable using “parameter analysis”

12.1 Non-proportional Regression Models

To perform nonlinear regression, we have to “trick” the computer. All the regression routines in the world are essentially built on the idea of using linear regression. This means that we must find a way to “linearize” the data when it is non-proportional. Consider the data shown in figure 12.1. It represents the cost of electricity based on the number of units of electricity produced in a given month. The relationship is obviously in the shape of a logarithmic function.

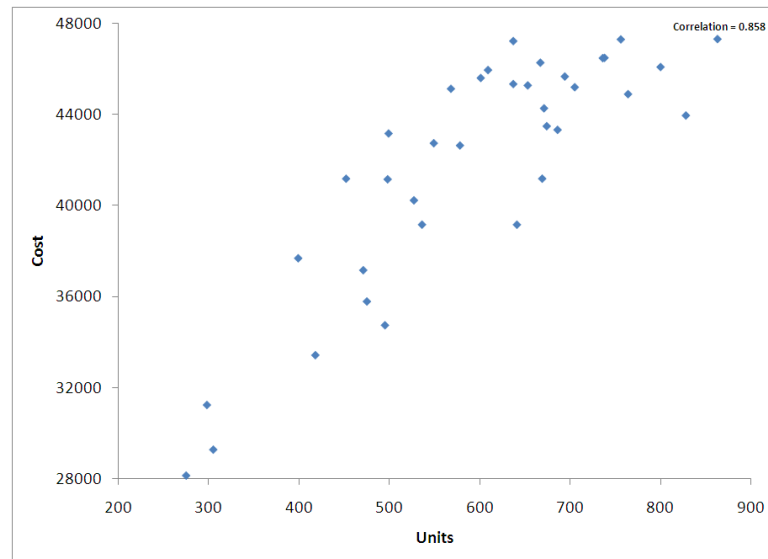


Figure 12.1: Graph of electricity cost vs. units of electricity.

Since the relationship indicates a logarithmic relationship, we examine, in figure 12.2, a graph of Cost vs. $\text{Log}(\text{Unit})$. Notice that this graph is “straighter”, indicating that we could use linear regression to predict Cost as a function of $\text{Log}(\text{Units})$. Thus, we can “trick” the computer into using linear regression on nonlinear data if we first “straighten out” the data in an appropriate way. An explanation of the straightening out process can be found in example 5.

Another way to say this is that the relationship is linear, but it is linear in $\text{Log}(x)$ rather than linear in x itself. Thus, we are looking for an equation of the form $y = A + B \log(x)$ rather than an equation of the form $y = A + Bx$. Notice that we are free to transform either the x or the y data or both. These different combinations allow us to construct many different models of nonlinear data.

We can also perform nonlinear analyses on data with more than one independent variable. In most cases, though, the only appropriate model for such data is a multivariable power model, called a multiplicative model. Such models are used mainly in production and economic examples. A famous example is the Cobb-Douglas production model which predicts the quantity of production as a function of both the capital investment at the company and the labor investment. See the examples for more information.

In the rest of this section, we’ll talk about how to select and complete the appropriate

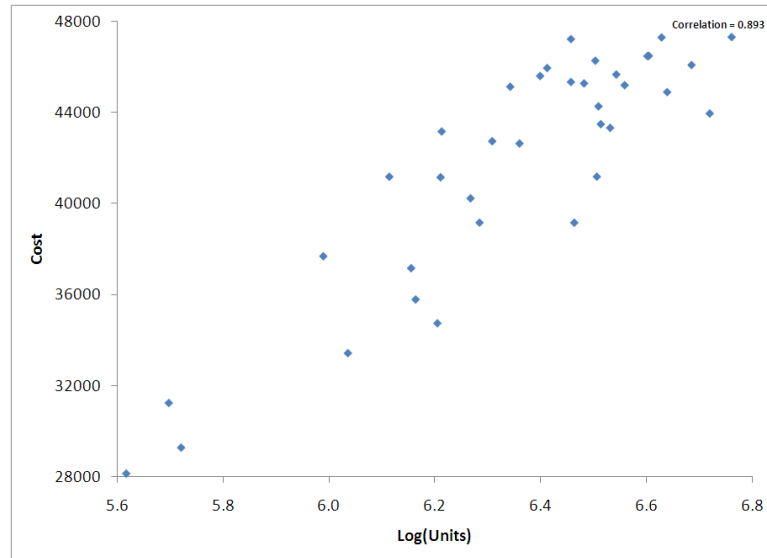


Figure 12.2: Graph of the electricity cost vs. the logarithm of electricity units used. Notice that this relationship is more linear than the one in figure 12.1 (the correlation is higher.) In a sense, we have “straightened” the data by taking the logarithm of the explanatory variable.

transformations, how to use these in regression routines, and how to compute R^2 and S_e in certain cases.

12.1.1 Definitions and Formulas

Multiplicative Model Basically this is a power function model for multivariable data. Also referred to as a “constant elasticity” model. (Elasticity is described in the next section.) A multiplicative model with two independent variables takes the form

$$y = AX_1^B x_2^C$$

where A , B , and C are all constants (parameters).

Cobb-Douglas This is a model for total production based on the levels of labor investment, capital investment, and other investments that influence productivity. If K = capital investment, L = labor investment and P = production, the Cobb-Douglas model look like

$$P = AK^B L^C$$

Notice that it is a multiplicative model as discussed above. There are some important cases in the Cobb-Douglas model depending on the values of the two powers, B and C . In general, these constants are both less than 1. The model reflects the idea that if you have a lot of labor investment (lots of workers) but not enough capital (equipment

for the workers to use) then productivity is hampered. If you have a lot of capital (equipment for production) but not the labor to use it, then production also suffers.

Non-constant Variance This is a problem that often occurs in real data. The basic issue is that the residuals seem to “fan out”. Thus, as the independent variable increases, the variability of the data around the proposed model increases systematically. (It is also possible for the variation to decrease systematically; this is less common, however.) Although the underlying pattern may be linear, non-constant variance is also “fixed” by an appropriate transformation of the variables.

12.1.2 Worked Examples

Example 12.1. One independent variable example (X transform)

The electricity data shown above (see figures 11.A.1 and 11.A.2 and the data file **C12 Power**) seems to be linear in either square root(units) or log(units) rather than linear in units. This means that we can construct a model for the cost of the electricity that is linear in either square root(units) or log(units). This model will look like $\text{Cost} = A + Bx$.

However, the x in this case will be either square root(units) or log(units) rather than units. To construct the models, we start by creating new variables in the data called “sqrt(units)” and “log(units)”. For example, StatPro allows you to do this automatically through the “Data Utilities/ Transform Variables” function, and R Commander allows you to do this through the “Data/Manage variables in active data set... “ menu. Once we have these new variables, we then go through the normal regression routines, using “Cost” as the response variable and either “Sqrt(units)” or “Log(units)” as the explanatory variable. The result of the regression routine when using sqrt(units) is shown below.

Results of multiple regression for Cost

Summary measures

Multiple R	0.8786
R-Square	0.7719
Adj R-Square	0.7652
StErr of Est	2540.5818

ANOVA table

Source	df	SS	MS	F	p-value
Explained	1	742724176	742724176	115.0698	0.0000
Unexplained	34	219454912	6454556		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	6772.5645	3290.6382	2.0581	0.0473	85.1875	13459.94
Sqrt(Units)	1448.7365	135.0544	10.7271	0.0000	1174.2730	1723.19

This leads us to the first nonlinear model for this data:

$$\text{Cost} = 6,772.56 + 1,448.74 * \text{Sqrt}(\text{units}).$$

This model is linear in square root(units). We can perform the same technique using the log(units) variable. The output from the regression routine is shown below and leads us to the model equation:

$$\text{Cost} = -63,993.30 + 16,653.55 * \text{Log}(\text{Units}).$$

This model is logarithmic in units; it is also said to be linear in log(units). This idea that the model is linear in a transformed variable is how we “trick” the computer into creating non-proportional models by performing linear regression. Notice that the logarithmic model is slightly better (it has a lower standard error) but the constant term is negative, making interpretation of this model more difficult.

Results of multiple regression for Cost

Summary measures

Multiple R	0.8931
R-Square	0.7977
Adj R-Square	0.7917
StErr of Est	2392.8335

ANOVA table

Source	df	SS	MS	F	p-value
Explained	1	7.68E08	7.67E08	134.0471	0.0000
Unexplained	34	1.95E08	5.23E06		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-63993.3047	9144.3428	-6.9981	0.0000	-82576.8329	-45409.78
Log(Units)	16653.5527	1438.3953	11.5779	0.0000	13730.3838	19576.82

Example 12.2. Another one independent variable example (Y transform)

Consider again the data in C13 Power. Suppose we decide to construct a power function fit for the data. Basically, a power model is a model in which the log(response) variable is linear in the log(explanatory) variable. Thus, we seek a model of the form

$$\text{Log}(\text{Cost}) = A + B * \text{Log}(\text{Units}).$$

For this, we first create variables log(cost) and log(units). We then perform the standard linear regression, using Log(cost) as the response and log(units) as the explanatory. The result is shown below. N.B. The summary measures are completely useless for this type of

model, since they are all based on $\text{Log}(\text{Cost})$ rather than actual cost. We must compute the correct summary measures for ourselves (see the How To Guide of this section for an example and the steps.) The actual correct summary measures are $R^2 = 0.7736$ and $S_e = 2530$. These are slightly better than the results of the linear fit ($R^2 = 0.7359$, $S_e = 2733$.)

Results of multiple regression for Log(Cost)						
Summary measures						
Multiple R	0.8967					
R-Square	0.8040					
Adj R-Square	0.7983					
StErr of Est	0.0617					
ANOVA table						
Source	df	SS	MS	F	p-value	
Explained	1	0.5312	0.5312	139.4835	0.0000	
Unexplained	34	0.1295	0.0038			
Regression coefficients						
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	7.8488	0.2358	33.2797	0.0000	7.3695	8.3281
Log(Units)	0.4381	0.0371	11.8103	0.0000	0.3627	0.5135

Example 12.3. The Multiplicative model

Consider the data shown in file **C12 Production**. This data shows the total production of the US economy (in standardized units so that it is 100 in 1899) as well as the investment in capital (K , also standardized) and labor (L , also standardized). We want to construct a model for predicting the productivity as a function of the capital and labor. We basically take two approaches with such multivariable data:

Approach 1. Try a multiple linear model.

Approach 2. If the linear model doesn't work well, try a multiplicative model.

Approach 1 in action. First we try predicting P as a linear function of K and L . (This is just like multiple linear regression models that we have seen before, so we omit some details.) The resulting model and summary measures are shown below.

$$\begin{aligned}
 P &= -2 + 0.8723L + 0.1687K \\
 R^2 &= 0.9409 \\
 S_e &= 11.1293
 \end{aligned}$$

Thus, it seems that a linear model does quite well, based on this information. However, in examining the diagnostic graphs, we notice that the residuals seem to spread out. To correct this, we try logging all the variables and producing a multiplicative model.

Approach 2 in action. So, now we transform each of the variables using the logarithmic transformation. This produces three new variables - $\log(P)$, $\log(K)$ and $\log(L)$. We then

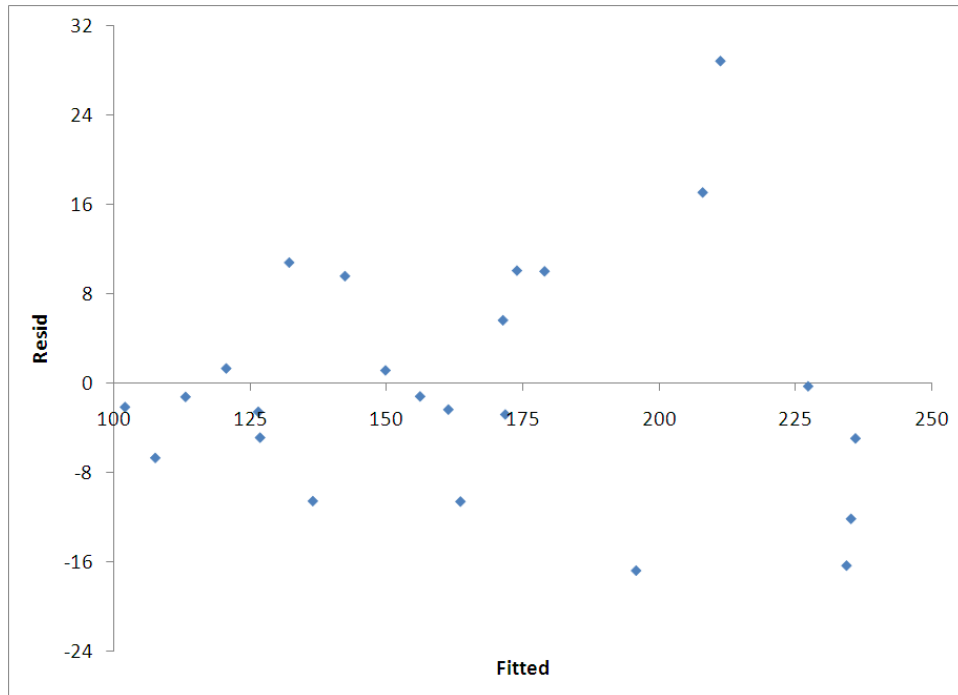


Figure 12.3: Plot of residuals versus fitted values for a linear model of predicting production vs. labor and cost.

perform a multivariable regression on $\log(P)$ as a function of both $\log(K)$ and $\log(L)$ to get the following results. Notice that we have computed the actual R^2 and S_e values using the techniques described in the computer how to for this section. Since we have logged the response variable (P) we cannot believe the regression output values for the summary measures.

$$\begin{aligned}\log(P) &= -0.0692 + 0.7689 \log(L) + 0.2471 \log(K) \\ R^2 &= 0.9386 \\ S_e &= 11.3449\end{aligned}$$

This model has about the same explanatory power as the linear model (very high, both are above 90% for R^2 .) Furthermore, we notice that the patterns in the residuals are no longer apparent. Interpreting this model will be left to the next section, but note that we can, with a little algebra, convert the model equation into the familiar form for a Cobb-Douglas production model. The result is $P = 0.9331L^{0.7689}K^{0.2471}$. Such models play an important role in many economic settings.

Example 12.4. Non-constant variance

The data in file **C12 Baseball** shows the salaries of over 300 major league baseball players along with many of their statistics for a particular season. Suppose that we want to predict the salary of a player based on the number of hits the player had during the season in order to test the assumption that better players have higher salaries.

If we do this, we see that the model is not very accurate ($R^2 = 0.34$). The reason for this is apparent in the plot of the residuals versus the fitted values (figure 12.4). One clearly sees the fan shape of these residuals, indicating that higher salaries also have higher variation from the model predictions.

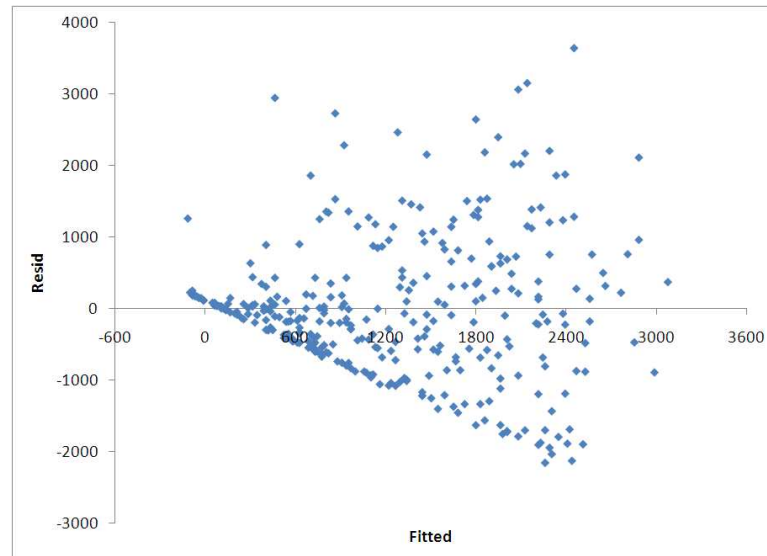


Figure 12.4: Plot of the residuals versus the fitted values when Salary is regressed against Hits.

To handle a fan that opens to the right, we typically log the response variable. Thus, we look for a model of the form $\log(y) = A + Bx$. Transforming the response variable produces the model equation $\log(\text{Salary}) = 5.1305 + 0.0151 \cdot \text{Hits}$. This model has the pattern for the residuals shown above in figure 12.5. Notice that the non-constant variance is greatly reduced. There does remain some narrowing of the pattern on the left, but this is largely due to the fact that there is a minimum salary in the data, so that there are no observations with actual salaries below a certain level.

It is also possible for the residuals to fan in the opposite pattern: spread out on the left and narrowing to the right. If this is the case with the data, we typically use the reciprocal of the response variable in the model.

Example 12.5. Straightening out data

The two graphs below show how the logarithmic function can be used to straighten out data that is non-proportional. In figure 12.6, we see data (indicated by the diamond shapes) that does not appear to be linear. These data have the coordinates (x_i, y_i) . To straighten the data out, we plot y versus the natural log of each of the x coordinates using squares to indicate these points in figure 12.7. Thus, we see how the original data points (x_i, y_i) are transformed to the data $(\ln(x_i), y_i)$ which have a less extreme curve.

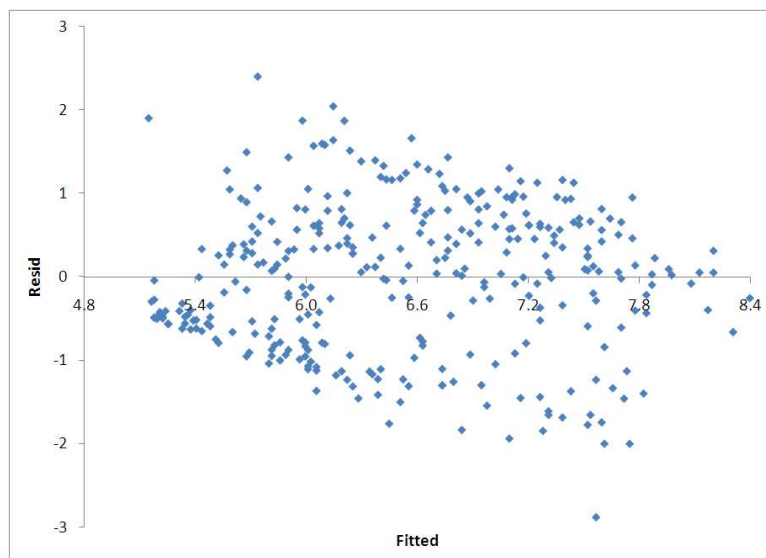


Figure 12.5: Plot of the residuals versus the fitted values when $\log(\text{Salary})$ is regressed against Hits.

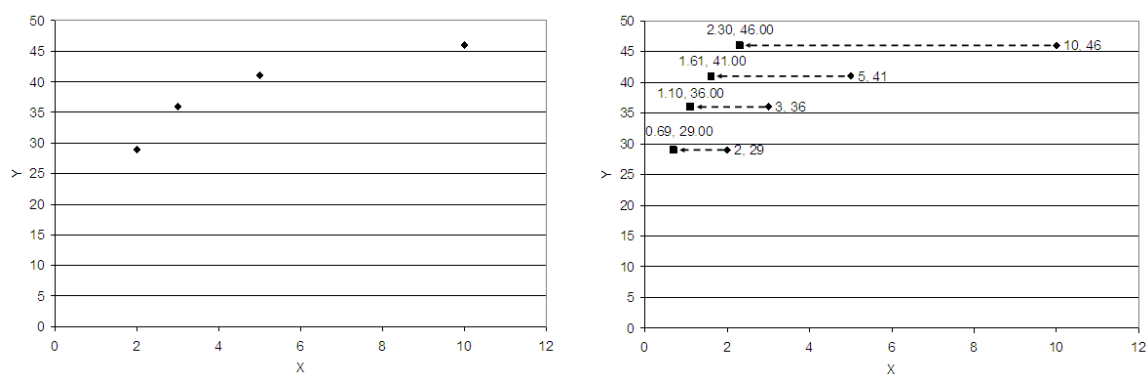


Figure 12.6: Plot of original data (left) and linearized data (right).

12.1.3 Exploration 12A: Learning and Production at Presario

(This problem is adapted from the data and example given in Data Analysis and Decision Making by Albright, Winston, and Zappe, example 11.6.)

The data from **C12 Learning** is taken from the Presario Company. This company manufactures small industrial products. The data show the length of time it took Presario to produce different batches of a new product for a customer. Clearly, the times tend to decrease as Presario gains experience with the production of this item. This indicates that the relationship between the time to complete a batch and the number of the batch is not a linear. We are going to explore this relationship.

1. First construct new variables for the logarithm of the batch number and the logarithm of the time to complete a batch.
2. Create the following scatterplots:

Dependent Variable	Independent Variable
Time	Batch
Log(Time)	Batch
Time	Log(Batch)
Log(Time)	Log(Batch)

Which of these graphs represents the most linear relationship? On what criterion (or criteria) are you judging this?

3. For each combination of variables, construct the regression model and determine the summary measures. Notice that for two of these models, the regression output will produce incorrect values for the summary measures. Which of these models is the best based on the summary measures? How does this compare with your choice of best model from the graphical approach in part 2?

12.2 Interpreting a Non-proportional Model

In the last section, we were concerned with finding the most appropriate regression model that would best fit a set of non-linear (i.e. non-proportional) data through a process of “straightening out” the data by transforming one or more of its variables. In this section, we will be concerned with how certain changes in the independent variable of such a non-proportional model bring about certain changes in its dependent variable by interpreting the model’s parameters in a way that is reminiscent of the way we study the slope parameter of a proportional model. Specifically, we will look at two ways to measure change for both the response and the explanatory variable: total change and percent change.

Total change is usually a level dependent quantity for non-proportional models. This means that we get very different amounts of total change at different levels of X values, even for the same total change in X . However, the idea of percent change incorporates this level dependency in its very definition. In fact, we have four basic combinations of the ways of measuring change. By examining these different combinations, we can develop a way of interpreting the parameters of regression models that we produce, for linear and many nonlinear models:

Total change in response variable vs. total change in explanatory variable
 Total change in response variable vs. percent change in explanatory variable
 Percent change in response variable vs. total change in explanatory variable
 Percent change in response variable vs. percent change in explanatory variable

However, it is not always easy to appreciate, and hence interpret, the parameters in the form in which they appear in the regression equations, as they appear in the first part of this chapter. This situation becomes apparent as we look at the chart of various models on page ???. It is not obvious, for example, why a model whose response variable has been logged and whose explanatory variable has not been logged is called an exponential model; likewise, it is not obvious why a model whose response variable as well as its explanatory variable has been logged is called a power model. Using the rules of exponents and logarithms, we shall rework each of these two regression models so that their coefficients become readily identifiable as the parameters in an exponential and a power function, respectively. From here, we will be able to readily interpret the effects of change in logarithmic, exponential, and power models in terms of their parameters in such a way that accounts for their level.

For example, we will find that the parameters in a logarithmic model are more easily interpreted if we look at the total change in the response variable contrasted to a 1% change in the explanatory variable. Exponential models on the other hand, are more easily interpreted by considering the percent change in the response variable contrasted to the total change in the explanatory variable. Interpreting the parameters in power functions is most easily done by examining the percent change in the response variable compared to a 1% change in the explanatory variable. For all of these models, the total or percent change in the response variable will depend directly on the values of the parameters in the model. Other non-proportional models, such as the quadratic or square root models, are not so easy to interpret in terms of their parameters and must await further developments in a later chapter.

12.2.1 Definitions and Formulas

Total Change Total change is a measure of the amount that a function changes from one data point to the other. Thus, if y is a function of the variable x we can find the value of y at two different x coordinates and then compute the total change in y . Note that the symbol “delta” which looks like a triangle is the symbol for change:

$$\Delta y = f(x_2) - f(x_1)$$

Notice that we always consider total change based on the assumption that the second x coordinate is larger than the first. In other words, we are looking at the change in y as x increases.

Rate of change This is an idea similar to the slope of a straight line, but rate of change can be applied to non-linear models. Rate of change measures the steepness of a graph at a given point (more precisely, we are talking about instantaneous rate of change). The steeper the graph is, the larger the rate of change is. If the rate of change is negative at a point, the graph is decreasing at that point. If it is positive at a point, the graph is increasing at that point. If it is zero, the graph could be at a maximum or a minimum value, or could be at a saddle point. Measuring rate of change is what the first semester of calculus is really all about. For our purposes, we want to understand the rate of change as a number. It’s useful for telling us “how much bang we get for each buck”. In other words, if we add more to the x variable (the bucks we spend) what does the rate of change say we get out (the bang). The rate of change of a function is closely related to the total change: usually we get at the rate of change through dividing the total change in Y by the total change in x . For linear functions, this number is the constant slope of the function. For nonlinear functions, the rate of change is level dependent.

Percent Change In many cases, it is easier to interpret the percent change in a quantity than to interpret the total change or the rate of change of the quantity. Percent change in a quantity is the total change divided by the original amount. Thus, if we start at the point $(x, f(x))$ and move to the point $(x + h, f(x + h))$, the total change is $f(x + h) - f(x)$, but the percent change in y is this divided by $f(x)$:

$$\frac{y_2 - y_1}{y_1} = \frac{f(x + h) - f(x)}{f(x)}$$

Notice that the percent change is a dimensionless number that represents a percent in decimal form. Thus, if the percent change of a model is 0.3 at a particular point, then this means that increasing x results in a $0.30 \rightarrow 30\%$ change in y at that point.

Units We’ve talked about this before, but it’s even more important now. Each number in a model (the constants, or parameters) will have some units associated with it. These units will help to interpret the meaning of the constant. So pay careful attention to the unit of measurement for each and every variable. Also note that the rate of change

has units; these units are always the units of the response variable divided by the units of the explanatory variable.

Elasticity Elasticity is an economic term for measuring the rate of change in a specific way. Elasticity is the actual rate of change divided by the current level. Thus, elasticity is really a measure of the percent change in the function, rather than a measure of the actual change (as the instantaneous rate of change is.) In fact, the elasticity of y with respect to x is the percentage change in y that results from a 1% increase in x .

Inverse functions Two functions, f and g , are inverses of each other if they satisfy the property that $f(g(x)) = x$ and $g(f(x)) = x$. This means that if you do something to x (like apply f to it to produce the number $f(x)$) and then do its inverse to it, you get back to the number you started with, x . In this chapter, the two functions that are important, $\ln(x)$ and $\exp(x)$, are inverses of each other.

Parameter Analysis A way of using the idea of change and percent change to interpret the coefficients (parameters) in a nonlinear regression model. Note that this is not a standard term.

Marginal Analysis This is a way of interpreting the amount of change in a function. Specifically, marginal analysis is used to answer the question “If the explanatory variable increases by one unit, by how much does the response variable change?”

Properties of Exponents You will need these properties in order to properly work with the regression output and convert it into a useable form. Sometimes you will apply these properties starting with the left side and converting it to the right side; other times you will have to go the other direction.

$$E1 \quad b^0 = 1$$

$$E2 \quad b^r b^s = b^{r+s}$$

$$E3 \quad (b^r)^s = b^{rs}$$

$$E4 \quad \frac{b^r}{b^s} = b^{r-s}$$

Properties of Logs You will need these properties in order to properly work with the regression output and convert it into a useable form. Sometimes you will apply these properties starting with the left side and converting it to the right side; other times you will have to go the other direction.

$$L1 \quad \ln(e^r) = r$$

$$L2 \quad e^{\ln(a)} = a$$

$$L3 \quad \ln(a) + \ln(b) = \ln(ab)$$

$$L4 \quad \ln(a) - \ln(b) = \ln\left(\frac{a}{b}\right)$$

$$L5 \quad r \cdot \ln(a) = \ln(a^r)$$

12.2.2 Worked Examples

Example 12.6. Converting regression output of an exponential model

The regression output for an exponential model will be of the form

$$\ln(y) = A + Bx$$

To convert this to the form “ $y = \dots$ ” we need to first exponentiate both sides of the equation in order to “undo” what has been done to y . (Remember, $\ln(y)$ and $\exp(y)$ are inverse functions, so each undoes the other.) We will go step-by-step through the process.

Algebraic Step	Explanation
$\ln(y) = A + Bx$	This is the output from the regression routine, written in equation form.
$\exp(\ln(y)) = \exp(A + Bx)$	$\exp(x)$ is the inverse of $\ln(x)$ and if we do something to one side of an equation, we must do it to both sides of the equation.
$y = \exp(A + Bx)$	Using the property that logarithms and exponentials are inverses, we know this is true.
$y = \exp(A) \cdot \exp(Bx)$	Property E2.

Thus, we are left with the functional form of the equation: $Y = e^A \cdot e^{BX}$.

To calculate (e^A) in most computer programs, use the exponentiation function, which is typically written as “=EXP(A)”. Also note that we can use property E3 to rewrite the functional form as $y = e^A (e^B)^x$. The reason for doing this is that the base of the exponent, $\exp(B)$, tells us how much things will increase. In fact, it tells us that regardless of the current level of output in the function, if x increases by 1 unit, the output will be $\exp(B)$ times that much. (Thus, if B is a number such that $\exp(B) = 2$, we know that increasing x by 1 unit results in the output, y , being multiplied by 2.)

Example 12.7. Converting regression output for power models

This is similar to converting an exponential model, only we need a few extra steps.

Algebraic Step	Explanation
$\ln(y) = A + B \ln(x)$	This is the output from the regression routine, written in equation form.
$\exp(\ln(y)) = \exp(A + B \ln(x))$	Exp(x) is the inverse of ln(x) and if we do something to one side of an equation, we must do it to both sides of the equation.
$y = \exp(A + B \ln(x))$	Property L2 (in disguise).
$y = \exp(A) \cdot \exp(B \ln(x))$	Property E2.
$y = \exp(A) \cdot \exp(\ln(x^B))$	Property L5.
$y = \exp(A) \cdot x^B$	Property L2 (in disguise).

This gives us the functional form of a power model: $y = (e^A) x^B$.

Example 12.8. Interpreting the rates of change for each model type

The examples below are taken from the data used for the introduction to this section. You can find this data in **C12 Power**. The response variable is the cost of the electricity produced based on the number of units of electricity produced that month (the explanatory variable.) For this data, we construct a number of different nonlinear models to try and explain the data based on the models. Note how each different model provides a different insight into the way the cost of electricity is dependent on the number of units of electricity that are produced.

1. Linear Models

- (a) Equation: $Y = A + Bx$
- (b) Interpretation: As X increases by 1, Y increases by B units
- (c) Example: If $\text{Cost} = 23651 + 31 \cdot \text{Units}$, for each additional unit of electricity that is produced, the cost increases by \$31. Thus, the constant B is measured in the units dollars per unit of electricity.

2. Exponential Models

- (a) Equation: $Y = Ae^{BX}$
- (b) Interpretation: As x increases by 1, y increases by a factor of $(e^B - 1)$
- (c) Example: If we have the model $\ln(\text{Cost}) = 10.1592 + 0.0008 \cdot \text{Units}$, then $\text{Cost} = 25828 \cdot e^{0.0008 \cdot \text{Units}}$, (notice: $e^{10.1592} = \exp(10.1592) = \$25,828$), for each additional unit, the cost increases by $(e^{0.0008} - 1) \approx 0.0008 = 0.08\%$. This means that if you are currently at a level of 500 units, costing \$38,531, then an additional unit will increase the cost by 0.080% of \$38,531, about \$30.82. In this case, the units of the constant are 1/units of electricity produced; this way the product of the constant B and the variable units has no units of measurement so we can exponentiate it.

3. Logarithmic Models

- (a) Equation: $y = A + B \cdot \ln(x)$
- (b) Interpretation: As x increases by 1%, y increases approximately $0.01B$
- (c) Example: If $\text{Cost} = -63993 + 16653 \cdot \ln(\text{units})$, then if the level of production (number of units) increases 1%, then the cost increases by approximately $0.01 \cdot 16653 = \$166.53$. Note that this means that the higher the production level, the greater the change required to produce the same increase in cost. At a production level of 100 units, a 1 unit increase will add about \$166.53 to the cost. However, at a production level of 500, it will take a 5 unit increase in production to increase the cost by \$166.53.

4. Power Models

- (a) Equation: $y = Ax^B$
- (b) Interpretation: As x increases by 1%, y increases approximately $B\%$
- (c) Example: If $\ln(\text{Cost}) = 7.8488 + 0.4381 \cdot \ln(\text{Units})$, then $\text{Cost} = 2563 \cdot \text{units}^{0.4381}$, since $\exp(7.8488) = 2563$. If the production level increases 1%, then the cost will increase by about 0.4381%; that is, add a percent sign after the number B to find the percent increase. At a production level of 100 units, the cost is about \$19273. If the level increases 1 unit (1%) then the cost will increase by 0.4381% of 19273 = \$84. At a production level of 500, the cost is \$39009, and a 1% increase in production (5 units) will increase the cost by about \$171.

5. Quadratic Models

- (a) Equation: $y = Ax^2 + Bx + C$
- (b) Interpretation: If A is positive, then there is a minimum point at $x = -B/2A$. If A is negative, then there is a maximum point at $x = -B/2A$
- (c) Example: Suppose we have the model: $\text{Cost} = 5793 + 98.35 \cdot \text{Units} - 0.06 \cdot \text{Units}^2$. Since the coefficient of units^2 is negative, so the model estimates there is a maximum point at a production level of $-(98.35)/2 \cdot (-0.06) = 820\text{units}$.

6. Multiplicative Models

- (a) Equation from regression output: $\ln(y) = C + B_1 \ln(x_1) + B_2 \ln(X_2)$
- (b) Equation rewritten in standard form: $Y = Ax_1^{B_1}x_2^{B_2}$. Note : $\exp(C) = A$.
- (c) Interpretation of B_1 : As x_1 increases by 1%, y increases by about $B_1\%$ from its current level (holding the other explanatory variable constant)
- (d) Interpretation of B_2 : As x_2 increases by 1%, y increases by about $B_2\%$ from its current level (holding the other explanatory variable constant)
- (e) Example: In the Cobb-Douglas model $P = 0.939037L^{0.7689}K^{0.2471}$ where P = Production, L = Labor, K = Capital, we see that as labor (L) increases by 1%, production increases by about 0.7689% from its current level. As capital increases by 1%, production increases by about 0.2471% from its current level.

If labor is currently at 200 and capital is currently at 500, then the current level of production is 256.37, so that a 1% increase in Labor (that is, 2 more units of labor are added), then production will increase by .7689% from its current level of 256.37 which is about 1.97 units. If capital increases by 1% of 500, i.e. 5, then production will increase by 0.2471% from its current level of 256.37 (increase of about 0.63 units).

We will refer to the results of this table - the rules for interpreting the parameters in each of these different types of models - as parameter analysis. To truly understand where these guidelines come from requires a little calculus. However, you can get a pretty good understanding of why these work based simply on playing with numbers in a spreadsheet. By creating a spreadsheet that calculates values of a function, total changes in the function, total changes in the explanatory variable, and percent changes in the variables, one can easily see where the rules come from and why they are only approximate. A spreadsheet for this has been constructed and is available under **C12 ParameterAnalysis**. This workbook contains a worksheet for each of the basic functional models above: linear, logarithmic, exponential, power, and quadratic. Each sheet allows you to change the parameters in the model and observe how the different ways of measuring change react.

12.2.3 Exploration 12B: What it means to be linear

One of the main ideas of a linear function is proportionality. One way to visualize this is shown in C12 StepByStep. On the first worksheet, labeled “linear”, you will see a straight line graphed. In addition, you will see three stair steps, three dotted lines (all horizontal) and two sliders at the top. The idea is that, in a linear function, if you walk a certain distance along the horizontal axis (the run), this forces you to climb up the function a certain amount (the rise).

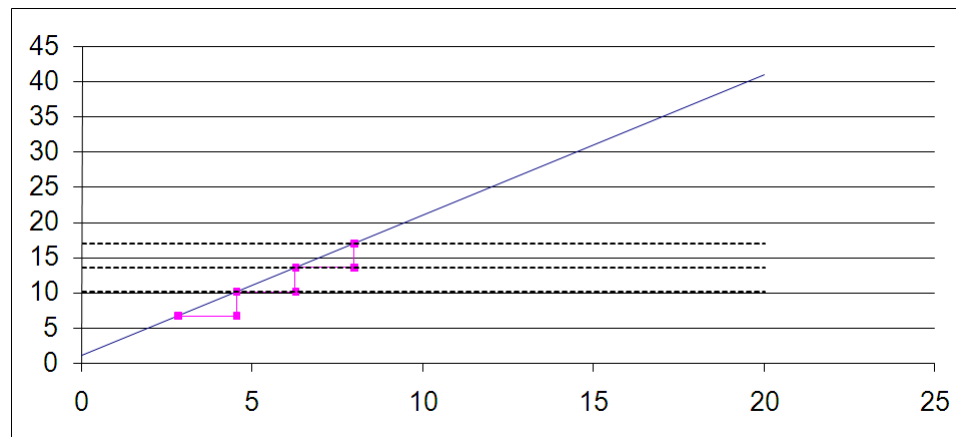


Figure 12.7: Screen shot of C12 StepByStep.

If I take three steps with the same horizontal distance each, and look at the rise that this produces, I will see something interesting; I could compute this total rise from three steps by just multiplying the rise from one step by 3. This is shown by the dotted lines; each dotted line marks the rise after a certain number of steps: the first line marks your place after one step, the second line marks your place computed by doubling the first step, and the last line marks the place you would get to by tripling the first step.

Furthermore, we can play with the sliders to change both the size of the first horizontal step and the location of the starting point for the first step. Regardless of the starting point or the size of the first step, both ways of computing the place on the line result in the same amount of change.

However, this is not the case in a non-linear function. Look at the worksheet labeled “Nonlinear”. This shows a similar set up, but with a curved graph, rather than a straight line. Here, we notice that regardless of the initial placement or the size of the first step, the two ways of computing the change are not equivalent. This is because the amount of change is level dependent in nonlinear functions.

We can summarize all of this in mathematical notation. For a linear function given by $y = f(x)$, we find that taking n steps of size Δx results in the same answer as taking one step of size Δx and multiplying this by n . Thus, for a linear function, we find the total change in y to be

$$\Delta y = f(x_1 + n\Delta x) - f(x_1) = n[f(x_1 + \Delta x) - f(x_1)].$$

12.3 Homework

Mechanics and Techniques Problems

12.1. Answer the following questions for the regression output shown below.

Results of simple regression for Log(Cost)						
Summary measures						
Multiple R	0.8529					
R-Square	0.7274					
StErr of Est	0.0728					
ANOVA table						
Source	df	SS	MS	F	p-value	
Explained	1	0.4806	0.4806	90.7367	0.0000	
Unexplained	34	0.1801	0.0053			
Regression coefficients						
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	10.1592	0.0510	199.0448	0.0000	10.0555	10.2630
Units	0.0008	0.0001	9.5256	0.0000	0.0006	0.0010

1. What is the regression equation, as taken directly from the output?
2. What kind of model does this represent (Linear, Logarithmics, Exponential, Power, Multiplicative)?
3. Convert the regression equation to standard form.

12.2. Answer the following questions for the regression output shown below.

Results of multiple regression for Cost						
Summary measures						
Multiple R	0.8931					
R-Square	0.7977					
Adj R-Square	0.7917					
StErr of Est	2392.8335					
ANOVA table						
Source	df	SS	MS	F	p-value	
Explained	1	7.68E08	7.68E08	134.0471	0.0000	
Unexplained	34	1.95E08	5.73E06			
Regression coefficients						
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-63993.3047	9144.3428	-6.9981	0.0000	-82576.8329	-45409.7765
Log(Units)	16653.5527	1438.3953	11.5779	0.0000	13730.3838	19576.7217

1. What is the regression equation, as taken directly from the output?
2. What kind of model does this represent (Linear, Logarithmics, Exponential, Power, Multiplicative)?
3. Convert the regression equation to standard form.

12.3. Answer the following questions for the regression output shown below.

Results of multiple regression for Log(Production)						
Summary measures						
Multiple R	0.9772					
R-Square	0.9550					
Adj R-Square	0.9507					
StErr of Est	0.0598					
ANOVA table						
Source	df	SS	MS	F	p-value	
Explained	2	1.5922	0.7961	222.9220	0.0000	
Unexplained	21	0.0750	0.0036			
Regression coefficients						
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-0.0692	0.4351	-0.1591	0.8751	-0.9740	0.8355
Log(Labor)	0.7689	0.1448	5.3087	0.0000	0.4677	1.0701
Log(Capital)	0.2471	0.0640	3.8634	0.0009	0.1141	0.3801

1. What is the regression equation, as taken directly from the output?
2. What kind of model does this represent (Linear, Logarithmics, Exponential, Power, Multiplicative)?
3. Convert the regression equation to standard form.

Application and Reasoning Problems

12.4. Use parameter analysis to interpret the model above for Log(Cost) as a function of Units. Your answer should be a sentence of the form “As the explanatory variable (Variable Name) changes by (1% or 1 unit), the response variable (Variable Name) changes by (amount or percent).”

12.5. Use parameter analysis to interpret the model above for Cost as a function of Log(Units). Your answer should be a sentence of the form “As the explanatory variable (Variable Name) changes by (1% or 1 unit), the response variable (Variable Name) changes by (amount or percent).”

12.6. Use parameter analysis to interpret the model above for $\text{Log}(\text{Cost})$ as a function of $\text{Log}(\text{Units})$. Your answer should be a sentence of the form “As the explanatory variable (Variable Name) changes by (1% or 1 unit), the response variable (Variable Name) changes by (amount or percent).”

CHAPTER 13

Nonlinear Multivariable Models¹

Most problems in the real world involve many variables. So far, you have encountered two types of models that have multiple independent variables: linear models and multiplicative models. These are definitely the most commonly used multivariable models since they are easier to interpret and can cover a variety of situations. But they do not cover all the possibilities. Probably the next most commonly used model is a quadratic multivariable model. This is the generalization of a parabola. This chapter will introduce you to this model in several ways.

- Section 13.1 shows how to create quadratic models using regression with interaction terms.
- Section 13.2 teaches you how to graph and visualize some of these models. This approach to graphing quadratics can then be used to graph other types of nonlinear models.

<i>As a result of this chapter, students will learn</i>	<i>As a result of this chapter, students will be able to</i>
---	--

- | | |
|--|--|
| ✓ How to interpret certain quadratic models of two variables | ✓ Create a contour plot of a function of two variables |
| ✓ The different shapes that the graph of a function of two variables can assume | ✓ Create a 3D surface plot of a function of two variables |
| ✓ How to simplify models with more than two variables when there are surrogate relationships | ✓ Use the discriminant to determine the shape of a quadratic model |
| ✓ The difference between substitute and complementary commodities | |

¹©2017 Kris H. Green and W. Allen Emerson

13.1 Models with Numerical Interaction Terms

In a previous chapter, we discussed building models using interaction terms. However, we only dealt with two of the three types of interaction terms: the interaction of two categorical variables and the interaction of categorical variable with a numerical variable. In this section, we will talk about what happens when you allow two numerical variables to interact, and what happens when you interact a variable with itself.

The second case is actually slightly easier to understand. Interacting a variable with itself produces a new variable in which each observation is the square of one of the observations of the base variable. Thus, a model built from a variable interacted with itself is a nonlinear model, specifically a square or quadratic model. This gives us another way to think about creating simple nonlinear models. Consider the data shown in the graph below, which has indication of being a parabola. The independent variable is Units (of electricity) and the dependent variable is Cost.

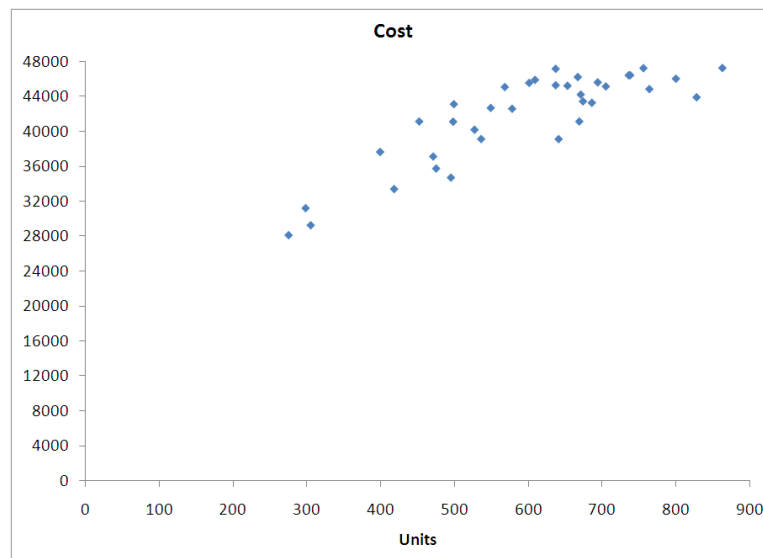


Figure 13.1: Electricity cost versus units used illustrating a nonlinear (possibly parabolic) relationship.

We can easily produce a quadratic model, and we find it has the equation

$$\text{Cost} = 5792.80 + 98.35 \cdot \text{Units} - 0.06 \cdot \text{Units} \cdot \text{Units}.$$

This model is clearly a parabola. It opens downward (as the graph shows) since the coefficient of the variable “Units · Units” is negative. (Of course, we don’t expect there to be a discount for using too much electricity, so a quadratic model is perhaps not the most appropriate here, but you get the picture.)

The other situation - interacting two different numerical variables - is much harder to visualize, since we are dealing with at least three dimensions (one for each of the base variables plus one for the dependent variable). In the next section, you will work on interpreting

such models and getting some sort of picture of what they might look like. For now, though, we concentrate on generating models of these two types, which are both quadratic models.

13.1.1 Definitions and Formulas

Interaction variable The product of two variables that constitutes a new variable and that captures, if it proves to be significant, the combined effect of the two original variables. Interaction terms can be created from any two variables. Most commonly, though, they are created from interacting either two categorical variables, or a categorical variable and a numerical variable (see chapter 10 for a discussion of such models).

Base Variable These are the original “uninteracted” variables from which the interaction terms were created.]

Quadratic model Any model made up of a combination of terms of the following forms: Constants, Constant · Variable, Constant · Variable², Constant · Var1 · Var2.

Term A term is any object added to other objects in a mathematical expression. For example, in the function shown below, there are three terms: $3x$, 2 and $5xy$.

$$f(x, y) = 3x + 2 + 5xy$$

Factor In a mathematical expression, a factor is one quantity (a variable or constant) that is multiplied with other quantities to make a term. For example, in the function above, the factors of the term $5xy$ are 5 , x , and y . The factors of the term $3x$ are 3 and x . The term “ 2 ” has only one factor, itself.

Factoring Mathematical/algebraic process of breaking terms into factored form so that several terms with similar factors can be grouped together. Often, this reveals hidden details of the model and can aid interpretation.

Self Interaction An interaction term created by multiplying or interacting a base variable with itself.

Joint Interaction An interaction term created by multiplying or interacting two different base variables.

13.1.2 Worked Examples

Example 13.1. Models built with one variable and self-interaction

Consider data on the Federal minimum wage, shown in `C13 MinWage`. This data shows the minimum wage (in dollars) at the end of each calendar year since 1950. Suppose we would like to build a model for this data in order to make projections about future labor costs for running a small company. Thus, we seek to explain the minimum wage, using the year as the independent variable.

One of the first things to note is that the years start in 1950 (when the minimum wage was established). This means that we are looking at large values for the independent variable, especially compared to the values of the minimum wage. It is helpful in situations like this to shift the independent variable to start at zero. Most software can easily transform the Year data into a new variable “Yr” representing the number of years since 1950, by subtracting 1950 from each year. (This means that “Yr = 25” is the year $1950 + 25 = 1975$.) One can also do this in Excel by simply entering the formula “=A2 - 1950” in cell C2 and copying this down the column. Graphing the minimum wage versus the year since 1950 produces a graph like the following.

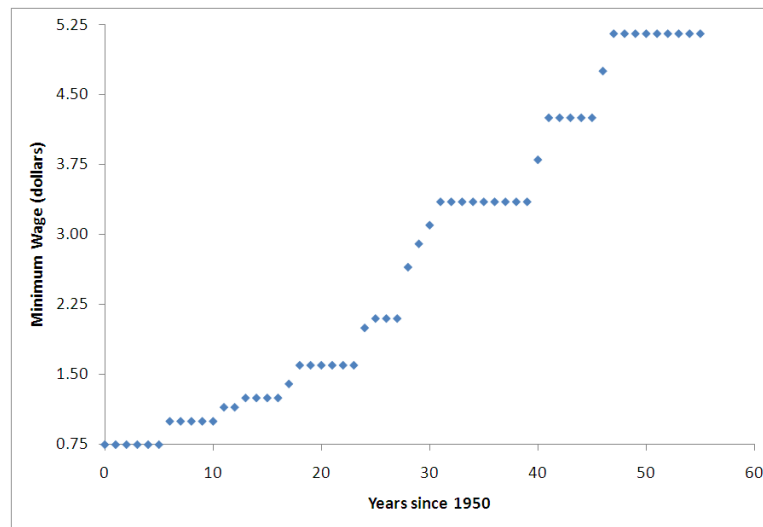


Figure 13.2: U.S. minimum wage versus years since 1950.

This graph clearly looks like part of a parabola, in spite of the high linear correlation. This means that it would be appropriate to introduce the interaction variable “Yr · Yr” and perform a multiple regression to build the model. The results of this are shown below.

The model equation is

$$\text{Minimum Wage} = 0.5196 + 0.0476 \cdot \text{Yr} + 0.0009 \cdot \text{Yr} \cdot \text{Yr}$$

We also see that the model has a coefficient of determination slightly worse than the linear model. This is due to the exact features of the graph; in particular, there are many years where the minimum wage does not change at all. The length of time the minimum wage stays constant seems to increase with time (since 1950) which stretches the graph out and makes the model slightly worse. A quadratic model, however, is clearly appropriate as can be determined from looking at the diagnostic graphs.

Results of simple regression for Price						
Summary measures						
Multiple R	0.9874					
R-Square	0.9750					
Adj R-Square	0.9740					
StErr of Est	0.2539					
ANOVA table						
Source	df	SS	MS	F	p-value	
Explained	2	133.0347	66.5174	1031.6093	0.0000	
Unexplained	53	3.4174	0.0645			
Regression coefficients						
	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	0.5196	0.0983	5.2874	0.0000	0.3225	0.7167
Yr	0.0476	0.0083	5.7618	0.0000	0.0310	0.0642
Yr*Yr	0.0009	0.0001	5.8760	0.0000	0.0006	0.0011

One thing that is not apparent from this model, however, is what it means. Using a method called “completing the square” we can rewrite the model as

$$\text{Minimum Wage} + 0.1098 = 0.0009(\text{Yr} + 26.4444)^2$$

What this version of the model shows us is that the Minimum Wage plus about \$0.11 is modeled well by a scaled horizontally shifted power function! We can use the techniques of the last chapter to make sense of this power function: for every 1% increase in the number of years since 1950, the minimum wage should increase about 2% above its present level. In 2006, which is 56 years after 1950, a 1% increase in the year would be $0.01 \cdot 56 = 0.56$ years = 6.72 months. The minimum wage predicted by the model in 2006 is about \$6.01. The interpretation of the model is that we would expect the minimum wage to increase 2% (about \$0.12) to \$6.13 roughly six to seven months into the year 2006.

Example 13.2. Modeling with two interacting variables

Consider the data shown in file C13 *Production*. These data show the total number of hours (label “MachHrs”) the manufacturing machinery at your plant ran each month. Also shown are the number of different production runs (“ProdRuns”) each month and the overhead costs (“Overhead”) incurred each month. In a previous chapter, we built the linear model shown below to explain these data.

$$\text{Overhead} = 3996.68 + 43.5364 \cdot \text{MachHrs} + 883.6179 \cdot \text{ProdRuns}$$

The model had a coefficient of determination of 0.8664 and a standard error of estimate of \$4,108.99, which was excellent compared to the standard deviation in overhead of \$10,916.81. In fact, it seemed the only problem with the model was the p-values for the constant term.

This was 0.5492, well above our 0.05 threshold for a “good” coefficient. So the question is can we improve on this without significantly complicating the model?

If we create all the possible interaction terms in the independent variables (these are $\text{MachHrs} \cdot \text{MachHrs}$, $\text{ProdRuns} \cdot \text{ProdRuns}$, and $\text{MachHrs} \cdot \text{ProdRuns}$), we could create a full regression model and then reduce it by eliminating those variables with high p-values. Unfortunately, this produces a model with all p-values well above 0.05, leaving us no idea which to eliminate first. We need a better approach. Rather than begin with all the variables and eliminate, we will use stepwise regression to build the model up, one variable at a time. The result of this stepwise regression is the model below.

$$\text{Overhead} = 35,778.20 + 0.6240 \cdot \text{MachHrs} \cdot \text{ProdRuns} + 21.2566 \cdot \text{MachHrs}$$

This model has a coefficient of determination of 0.8628 and standard error of \$4,163.77, comparable to the linear model. However, the p-values for this model, including the constant term, are zero to four decimal places! Thus, the model more accurately shows the influential variables. But is this model too complex for interpretation?

One technique you may have encountered in previous mathematics classes is called factoring. Notice that the last two terms in the model both contain the same factor, MachHrs . Let’s write the model in a different order without changing the model and then group the terms with similar factors together using parentheses, drawing that common factor out.

$$\text{Overhead} = 35,778.20 + (0.6240 \cdot \text{ProdRuns} + 21.2566) \cdot \text{MachHrs}$$

Now we notice that the model looks sort of linear. It’s like the variable is MachHrs , the y -intercept is \$35,778.20 and the “slope” is $0.6240 \cdot \text{ProdRuns} + 21.2566$. Notice that since this is not a constant slope, we cannot truly call it such, but it can be interpreted this way: For each production run during the month, the cost of running the machinery for one hour increases by \$0.6240 from its base cost of \$21.26 per hour. So even though the model is quadratic and has an interaction term, it is still simple enough to interpret.

Example 13.3. Modeling with many interacting variables

In this example, we return to the commuter rail system introduced in an earlier chapter. If you recall, Ms. Carrie Allover needed a model to predict the number of weekly riders (in thousands of people) on her rail system based on the variables Price Per Ride, Income (representing average disposable income in the community), Parking Rate (for parking downtown instead of taking the rail system) and Population (in thousands of people). Previously, we developed a multilinear model for these data:

$$\begin{aligned} \text{Weekly Riders} = & -173.1971 - 139.3649 \cdot \text{Price per Ride} + 0.7763 \cdot \text{Population} \\ & -0.0309 \cdot \text{Income} + 131.0352 \cdot \text{Parking Rate} \end{aligned}$$

This model fit the data reasonably well, but we might ask whether we can do better, since the p-value for the constant term was so high (0.4389). Let’s try a quadratic model. First, we create the interaction variables. There are four independent variables, so that gives us four variables representing self-interaction ($\text{Income} \cdot \text{Income}$, $\text{Park} \cdot \text{Park}$, $\text{Pop} \cdot \text{Pop}$, Price

· Price) and $4 \cdot 3/2 = 6$ interaction terms created from two different variables. You can see the complete list of variables in **C13 Rail System**.

Clearly the full quadratic regression model will be complicated. Fortunately, many of the p-values in the full model are well above 0.05. Rather than build our model by eliminating variables one at a time, though, let's retrace our steps and perform a stepwise regression. We'll submit "Weekly Riders" as the response variable and we will submit all of the variables (the four base variables, the four square terms and the six interaction terms) as possible explanatory variables. The software will then build the model up from nothing adding in only the relevant variables rather than having us work from the full model and eliminate variables. The result is much simpler than we might have expected.

$$\begin{aligned}\text{Weekly Riders} = & 596.491 + 0.0002 \cdot \text{Pop} \cdot \text{Pop} - 0.0864 \cdot \text{Price} \cdot \text{Pop} \\ & + 36.0244 \cdot \text{Park} \cdot \text{Park} - 0.0229 \cdot \text{Income}\end{aligned}$$

This model has a coefficient of determination of 0.9342 and standard error of 23.0119, which are not very different from the linear model we started with, but we gain one significant advantage: all the p-values are significant.

Still, our model has four independent variables involved. This makes it extremely difficult to interpret. One way to do so would be to rewrite the model slightly by factoring the terms involving Population.

$$\begin{aligned}\text{Weekly Riders} = & 596.491 + \text{Pop} \cdot (0.0002 \cdot \text{Pop} - 0.0864 \cdot \text{Price}) \\ & + 36.0244 \cdot \text{Park} \cdot \text{Park} - 0.0229 \cdot \text{Income}\end{aligned}$$

This leaves us with a model indicating that:

- For each \$1 increase in disposable income, we expect 0.0229 thousand (about 23) fewer riders each week.
- Population has a generally positive effect on ridership, but its effect is mitigated by the price per ride; for each \$1 increase in ticket price, we expect the effect of population to be decreased by 0.0864 thousand riders per thousand people in the population.

Obviously, this model is complicated. Interpreting it is still difficult. However, we can reduce this model to a quadratic model of two variables by taking advantage of some of the natural correlations in the data. Looking at the correlations (table 13.2) shows us that there are strong linear relationships between Income and Parking Rates and between Price per Ride and Parking Rates. These relationships are shown in table 13.3 below.

	Weekly Riders	Price per Ride	Population	Income	Parking Rate
Weekly Riders	1.000				
Price per Ride	-0.804	1.000			
Population	0.933	-0.728	1.000		
Income	-0.810	0.961	-0.751	1.000	
Parking Rate	-0.698	0.958	-0.645	0.970	1.000

Model	Correlation	R^2	S_e
Income = $2046.8727 + 3191.5617 \cdot \text{Park}$	0.970	0.9408	505.1306
Price = $-0.0929 + 0.5672 \cdot \text{Park}$	0.958	0.9176	0.1072

In the equation above, we substitute these relationships (replace Income with $2046.8727 + 3191.5617 \cdot \text{Park}$ and replace Price with $-0.0929 + 0.5672 \cdot \text{Park}$) and eliminate those two variables (which are surrogate variables for Parking Rate, apparently). The reduced model looks like

$$\begin{aligned} \text{Weekly Riders} = & 596.491 + 0.0002 \cdot \text{Pop} \cdot \text{Pop} - 0.0864 \cdot (-0.0929 + 0.5672 \cdot \text{Park}) \cdot \text{Pop} \\ & + 36.0244 \cdot \text{Park} \cdot \text{Park} - 0.0229(2046.8727 + 3191.5617 \cdot \text{Park}). \end{aligned}$$

Simplified, this model becomes

$$\begin{aligned} \text{Weekly Riders} = & 549.618 + 0.0002 \cdot \text{Pop} \cdot \text{Pop} + 0.00799 \cdot \text{Pop} - 0.0490 \cdot \text{Park} \cdot \text{Pop} \\ & + 36.0244 \cdot \text{Park} \cdot \text{Park} - 73.0868 \cdot \text{Park}. \end{aligned}$$

This two-variable quadratic model is simpler in many ways than the original nonlinear model. However, we will leave interpretation of this model to the next section, when we learn how to picture this model as a surface in three-dimensions.

13.1.3 Exploration 13A: Revenue and Demand Functions

File C13 Exploration A contains weekly sales and revenue information for two different companies. The first worksheet, labeled “Company 1” shows the quantities of two complementary commodities that are sold by this company. These items are X and Y. The second sheet contains data on two substitute commodities sold by “Company 2”.

1. Formulate a quadratic regression model for Company 1’s revenue as a function of the quantity of each item that is produced and sold.
2. Formulate a quadratic regression model for Company 2’s revenue as a function of the quantity of each item that is produced and sold.

You should now have two revenue functions that look something like this:

$$R(q_1, q_2) = Aq_1^2 + Bq_2^2 + Cq_1q_2 + Dq_1 + Eq_2 + F$$

Where the capital letters are constants and variables q_1 and q_2 represent the quantity of goods of each type.

3. Explain why, in the revenue formula above, you would expect F , the constant term, to be zero. Do your regression models match this prediction?

We are going to use these revenue functions to determine the demand functions for the products in each case. Recall that the demand function gives the unit price that the market will pay for something, given the supply (in this case the quantities q_1 and q_2) of the item(s) being sold. To find the demand functions, we need to write the revenue function in the form

$$R(q_1, q_2) = q_1p_1 + q_2p_2$$

In this formula, the p_1 and p_2 are the unit prices. We will assume that these are both linear functions of the two quantities.

4. What does it mean in the last sentence when it says that p_1 and p_2 are a linear function of the quantities? Give a sample function that could represent p_1 or p_2 .
5. Try to find the demand functions for each situation. You can do this by (a) factoring the regression models you have formulated above and (b) assuming that the term with the coefficient C in the revenue formula is split equally between the two demand functions.
6. Use your demand functions to fill in the tables below, showing the estimated prices customers would pay at each company for different supplies of the two goods.

Company 1			
q_1	q_2	p_1	p_2
1000	1000		
1100	1000		
1000	1100		

Company 2			
q_1	q_2	p_1	p_2
2000	2500		
2100	2500		
2000	2600		

7. Based on your demand functions (you should now have four: two for each scenario) and your data in the tables above what do you think are meant by the terms “complementary commodities” and “substitute commodities”?

13.2 Interpreting Quadratic Models in Several Variables

When dealing with multivariable models, there are, literally, an infinite number of ways to explore them, depending on what kind of graph you want, which part of the model you want to graph, whether you would prefer looking at the data in a table of numbers, or a host of other possible choices. It helps to have some basic skills and options for visualizing functions with two independent variables. As we'll see, graphing them requires three dimensions, one for each independent variable and one for the dependent variable. Thus, if you want to graph a model with more than two independent variables, you need some mighty special paper!

Obviously, one way to gain an understanding of how the function behaves is to make a table of data. You've seen such tables before for functions of several variables, you just didn't realize it. One very common example relates to the weather. You've heard of wind chill probably. This is a measure of how cold the air feels, based not only on the actual temperature, but also on the wind speed. To use such a table (like the one below) you simply locate the intersection of the wind speed (down the left column) and air temperature (across the top row) to find the wind chill. Such a process defines a function of two variables. If we let W stand for the wind chill, S for wind speed and T for air temperature, then we could write

$$W = W(S, T)$$

to represent the relationship; this emphasizes that W is a function of S and T . For example, $W(25, 10) = -29$ indicating that a 25 mph wind on a 10 degree day makes the air feel like it is actually 29 degrees below zero!

Wind Speed (mph)	Ambient Air Temperature (degrees Fahrenheit)																
	35	30	25	20	15	10	5	0	-5	-10	-15	-20	-25	-30	-35	-40	-45
5	33	27	21	16	12	7	1	-6	-11	-15	-20	-26	-31	-35	-41	-47	-54
10	21	16	9	2	-2	-9	-15	-22	-27	-31	-38	-45	-52	-58	-64	-70	-77
15	16	11	1	-6	-11	-18	-25	-33	-40	-45	-51	-60	-65	-70	-78	-85	-90
20	12	3	-4	-9	-17	-24	-32	-40	-46	-52	-60	-68	-76	-81	-88	-96	-103
25	7	0	-7	-15	-22	-29	-37	-45	-52	-58	-67	-75	-83	-89	-96	-104	-112
30	5	-2	-11	-18	-26	-33	-41	-49	-56	-63	-70	-78	-87	-94	-101	-109	-117
35	3	-4	-13	-20	-27	-35	-43	-52	-60	-67	-72	-83	-90	-98	-105	-113	-123
40	1	-4	-15	-22	-29	-36	-45	-54	-62	-69	-76	-87	-94	-101	-107	-116	-128

But, making tables of the data from a function is only one way to study its behavior. And, the table of numbers may be difficult to read and interpret. In addition, the spacing of the values in the table may hide some important features. For example, the wind chill table makes it appear that no matter what, if the wind speed increases, the air feels colder (wind chill is lower). But what if between 20 and 25 mph, it actually gets a little warmer for some reason? Our table would not show this.

So, another common tool for studying such functions is to create 3D surface plots of them. If we copy the table above into our spreadsheet and create such a plot, we get a figure

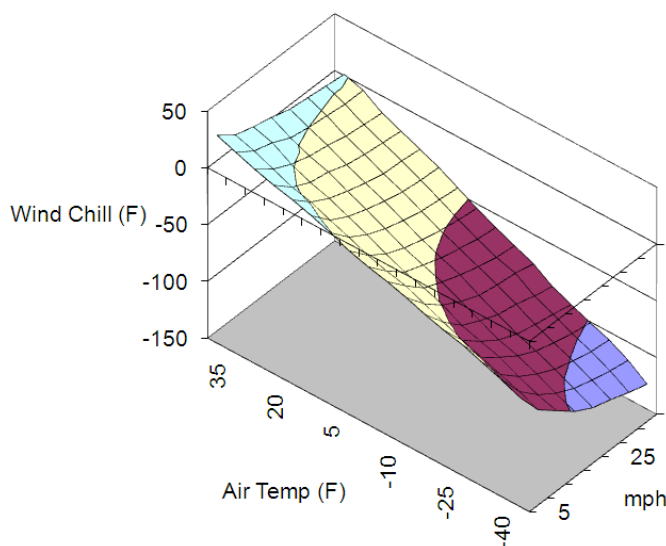


Figure 13.3: 3D plot of wind chill versus air temperature and wind speed.

like the one below. We can adjust the perspective of the graph, but otherwise, it has many of the same features as all the scatterplots we've used before.

In this section, we will use this graphical tool to help us understand the different types of quadratic models that we may get from applying the techniques of the previous section. In general, we will be dealing with models of the form

$$f(x_1, x_2) = E + A_1x_1 + A_2x_2 + B_1x_1^2 + B_2x_2^2 + Cx_1x_2$$

And will want to know what different shapes the graphs of such functions may take. Fortunately, there are only a few possibilities, and we will learn some ways of quickly classifying any function as being one of these types (either a bowl-shaped surface, a hill-shaped surface, or a saddle-shaped surface)

While it may seem restrictive to study such as specific class of functions, it turns out that there are several good reasons for it. The first is that it arises easily in modeling, as the techniques of the last section showed. The second is that if we zoom on the surface of any random function of two variables, on a small enough scale it looks like a quadratic. Thus, studying these objects gives us a lot of tools for understanding more complex objects.

13.2.1 Definitions and Formulas

Dimensions For each variable (independent or dependent) in a model, you need one dimension in order to create a graph of the model. Thus, a model like $y = f(x)$ needs two dimensions, one for y and one for x . A model like the general quadratic below needs three dimensions for its graph.

Surface Plot A graphic representation of a function of one variable (two dimensions) is a scatterplot. Creating a similar type of graph for a function of two variables requires three dimensions. Each point has three coordinates, and the height of the point above the xy -plane is the value of the function. When the points are connected together, they form a surface in three dimensions.

General Quadratic Model The general quadratic model we will use in this text is

$$f(x_1, x_2) = E + A_1x_1 + A_2x_2 + B_1x_1^2 + B_2x_2^2 + Cx_1x_2$$

In this, we assume that at least one of the B coefficients is non-zero. Other texts may refer to the model in slightly different terms, but the important things to note are that (1) this is a polynomial model (in two variables) and (2) the degree of each term (sum of the powers of each variable) is either 0, 1 or 2. For example, the terms with a B coefficient all have one variable raised to the second power and the other raised to the zeroth power, so they are degree 2. The cross term (the term with the C coefficient that involves both independent variables) has both variables raised to the first power, so its degree is $1 + 1 = 2$ as well.

Discriminant There are several mathematical objects that go by the name “discriminant”. Each is used to discriminate between several alternatives. In this case, we are referring to a quantity that can be derived from the formula for the general quadratic that helps decide whether the graph will look like a bowl, a hill or a saddle. Using the symbols above, the discriminant is the quantity

$$D = 4B_1B_2 - C^2$$

The shape of the graph (as we will see in the examples), depends on this quantity in the following ways:

1. If $D > 0$ and $B_1 > 0$, then the graph will look like a bowl.
2. If $D > 0$ and $B_1 < 0$, then the graph will look like a hill.
3. If $D < 0$, then the graph will look like a saddle.
4. If $D = 0$, then the discriminant is not helpful.

There are two other possible shapes for the graph, which occur if the coefficients in front of all instances of one variable are zero. In that case, the graph looks like either a trough (if the remaining B coefficient is positive) or a speed bump (if the coefficient is negative).

Depending on your viewpoint and the exact values of your graph, you may not be able to see it has a particular shape, though (see example 5).

13.2.2 Worked Examples

Example 13.4. Looking at a multi-linear function

Recall the model from the previous section that represented our best, linear efforts to model the overhead based on the machine hours and production runs:

$$\text{Overhead} = 3996.68 + 43.5364 \cdot \text{MachHrs} + 883.6179 \cdot \text{ProdRuns}$$

File **C13 Production2** shows a table of values for this function, calculated over a domain similar to that present in the data. Below is a 3D surface plot of these data, showing the linear structure.

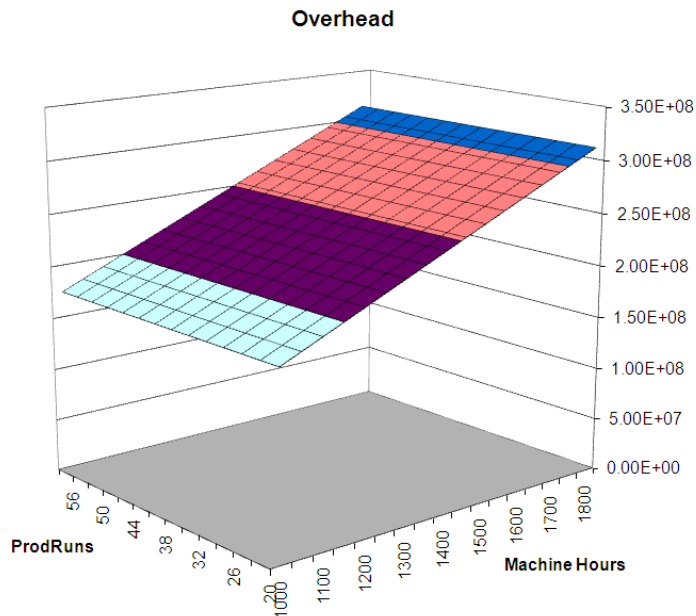


Figure 13.4: Linear two-variable model of overhead versus Production Runs and Machine Hours.

Notice that this graph appears to be a flat plane, like a piece of paper tilted at an angle. Any linear function of two variables has such a graph.

Example 13.5. Looking at a quadratic function of two variables

Here is one possible graph for a quadratic function of two variables. This is based on the quadratic model of the overhead costs found in **C13 Production2** in the worksheet labeled “Example 13B2”. It uses the model shown below.

$$\text{Overhead} = 35,778.20 + 0.6240 \cdot \text{MachHrs} \cdot \text{ProdRuns} + 21.2566 \cdot \text{MachHrs}$$

Notice that the formula in cell C5 uses mixed cell references (see the “How To Guide” for details) in order to calculate the overhead from a given number of machine hours (in column B) and a given number of production runs (row 5).

$$C5 = 35778.2 + 0.624 * \$B5 * C\$4 + 21.2566 * \$B5$$

The graph of this model is shown below.

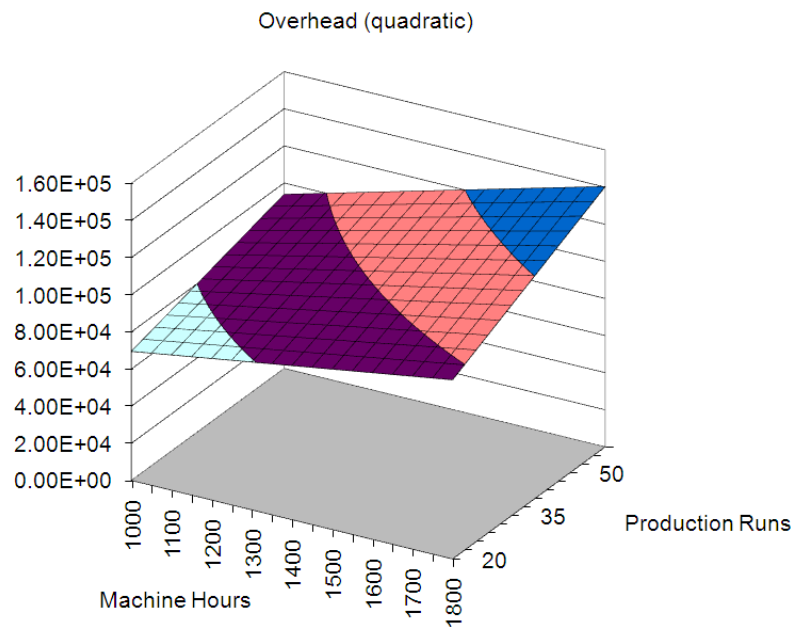


Figure 13.5: Quadratic two-variable model of overhead versus Production Runs and Machine Hours.

Notice that this graph also appears, at first glance, to be linear - like a plane. However, the contour lines on the surface between the different colored regions are curved, indicating that this is truly a nonlinear model. The reason it doesn't look quadratic is because of the particular set of values of MachHrs and ProdRuns we have used to graph the function. When we graph it over a larger region, we can clearly see the warped "saddle" shape of the surface become apparent. Of course, we could never have negative values of machine hours or production runs in a given month, so the actual data will never show this. Thus, we see that even when the data may be best represented by a nonlinear model, it may not be clear from the graph.

Also note that in the notation given in the "Definitions and Formulas" for the discriminant, we have $B_1 = B_2 = 0$ and $C = 0.6240$. This means that the discriminant, D , is -0.62402 , which is less than zero, confirming that we should see a saddle in the graph.

For the sake of completeness, we view the graph of overhead from above (graphed on the region with all independent variables positive). Such a graph is called a contour plot and shows curves (called contours) that separate regions based on their coordinate in the third dimension. Notice that all of the contours are curved, another indication that the underlying graph is nonlinear. In fact, it can be shown that these curves are hyperbolas, a type of object closely related to parabolas.

Example 13.6. Another quadratic surface

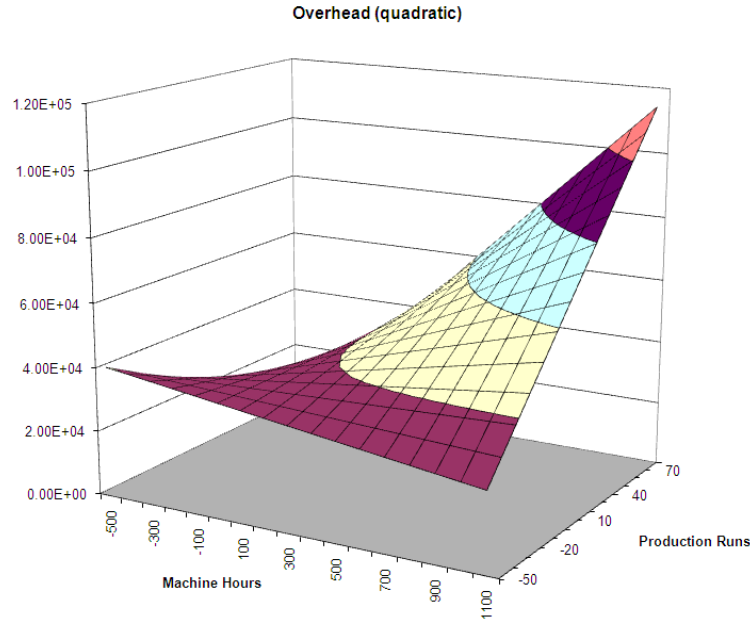


Figure 13.6: Quadratic two-variable model of overhead versus Production Runs and Machine Hours. Note that this is graphed over a different domain than in figure 13.5 emphasizing the nonlinear nature of the graph.

Let's look at a graph of the surface representing the quadratic Weekly Riders model from example 3. This model, after reducing it to two variables, became

$$\begin{aligned} \text{Weekly Riders} = & 549.618 + 0.0002 \cdot \text{Pop} * \text{Pop} + 0.00799 \cdot \text{Pop} - 0.0490 \cdot \text{Park} \cdot \text{Pop} \\ & + 36.0244 \cdot \text{Park} \cdot \text{Park} - 73.0868 \cdot \text{Park} \end{aligned}$$

When graphed over the region with Parking Rates from \$0.50 to \$2.50 and Population between 1,000 thousand people and 2,000 thousand people, we appear to see a linear model. But a calculation of D gives $D = 0.0264$ which is positive. Since the coefficients of the squared terms are both positive, this seems to indicate that we should see a bowl-shaped surface. How are we to reconcile the calculation with the graph?

This is always part of the problem in graphing and interpreting nonlinear models, especially those of several variables: such functions tend to have large domains, and tend to look very different at different locations in the domain. To emphasize this, we look at the graph on a slightly expanded domain where the shape is more evident.

Example 13.7. Multiplicative models

As a final example, we will look at a graph of one of the other multivariable, nonlinear models we have encountered, the multiplicative model. The model below is a Cobb-Douglas production model. P represents the total production of the economy, L represents the units of labor available and K represents the units of capital invested. We met such models in the

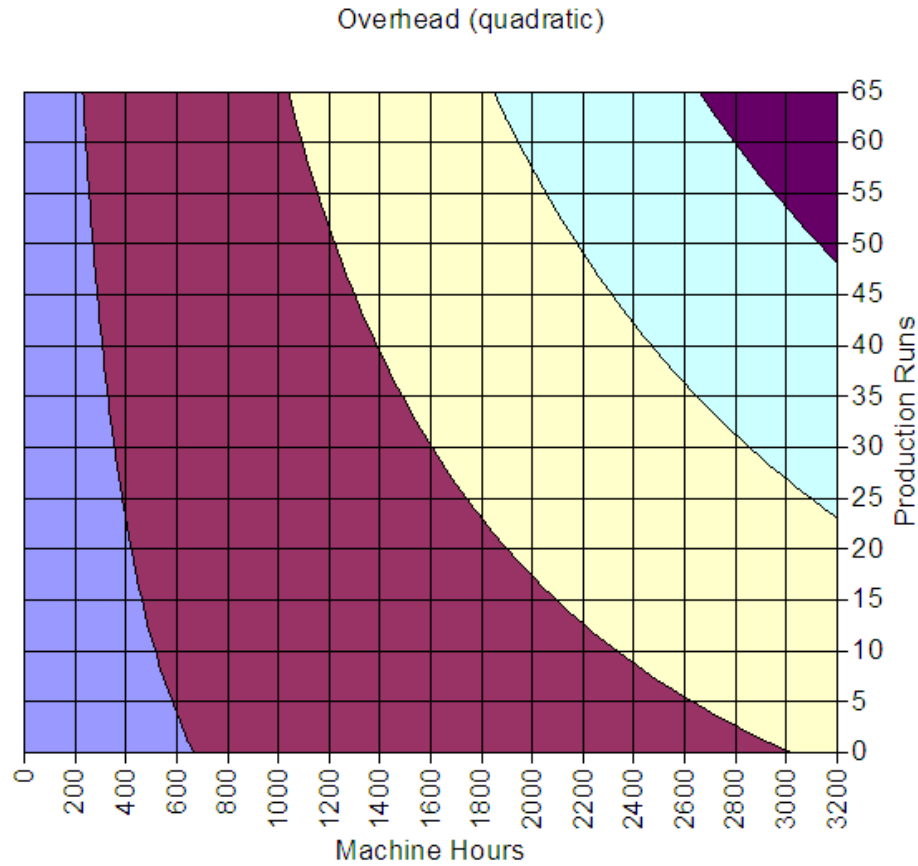


Figure 13.7: Contour view of the quadratic model of overhead. Note that the contours (or level curves) are not straight lines, as in a linear model, but are curved.

last chapter and applied parameter analysis to their interpretation. But what do they look like?

$$P = 0.939037L^{0.7689}K^{0.2471}$$

As you can see from the graph below, when we plot the production for reasonable values of the labor and capital (both positive) the contours look like those of a saddle-shaped surface, but the graph does not look like a saddle. The graph shows that if either of the inputs is zero (capital or labor) the production is zero. It also shows that if you increase either input (or both) you continue to get more output.

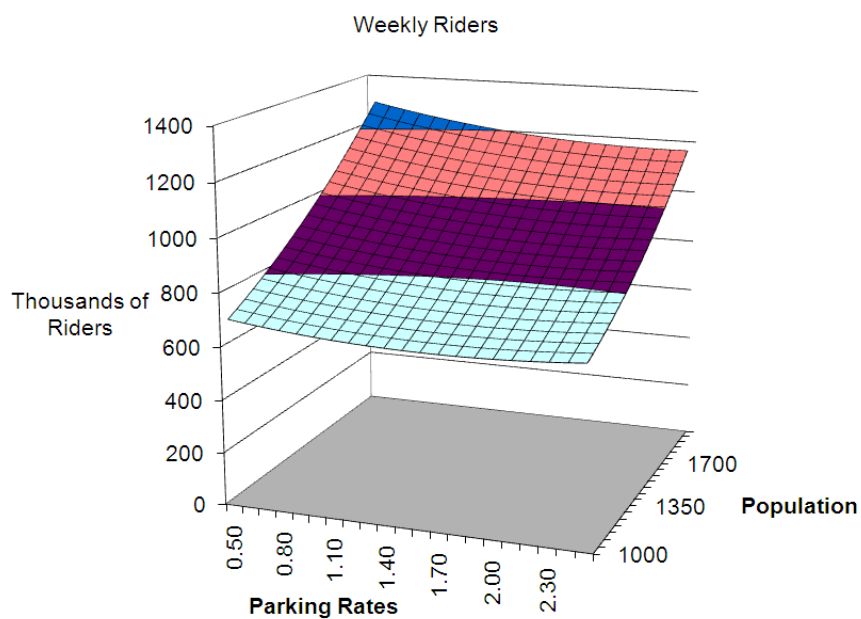


Figure 13.8: Quadratic model of weekly riders versus population and parking rates.

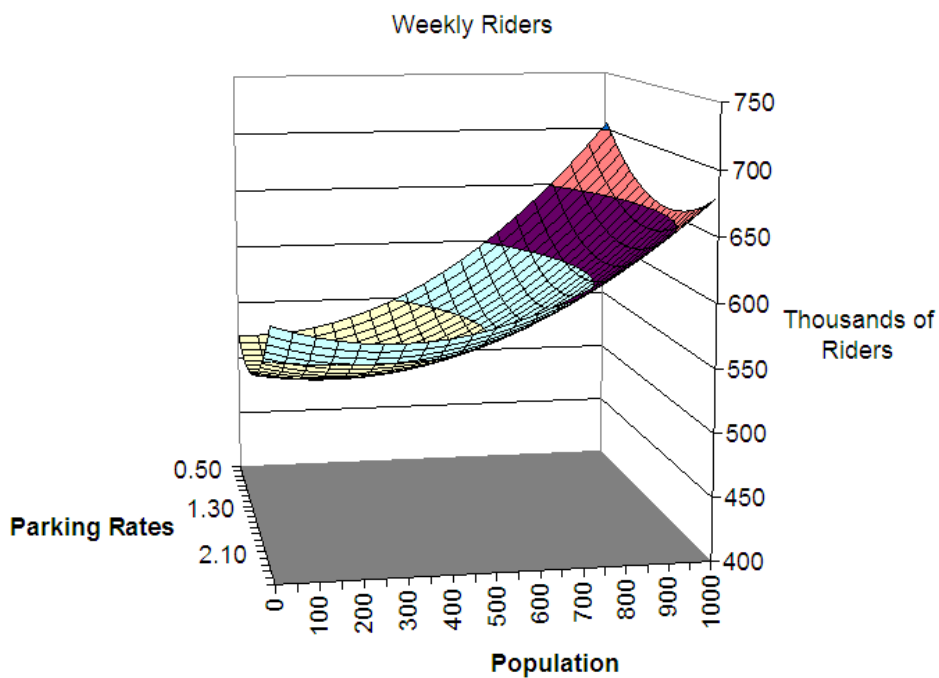


Figure 13.9: Different view of the graph in 13.8 showing the bowl-shape.

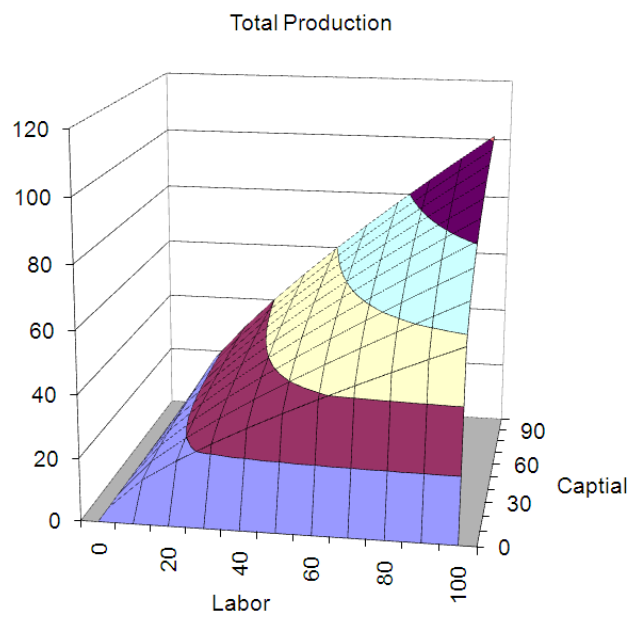


Figure 13.10: 3D plot of a Cobb-Douglas model, illustrating the nonlinear nature of the model.

13.2.3 Exploration 13B: Exploring Quadratic Models

In this exploration, you will get a chance to connect the different shapes of the quadratic graphs to the values of the coefficients and see some realistic examples where these different shaped graphs might occur. Consider the revenue generated from selling two different products. Since revenue is the quantity sold (q_1 will be the quantity of item 1 sold; likewise for item 2) times the unit price of the item (p_1 will be the unit price of item 1) we can reasonably assume that the revenue function looks something like this:

$$R(q_1, q_2) = q_1 p_1 + q_2 p_2$$

Depending on the particular goods, we might have the prices of each item related to the quantity of both items sold. Two common situations in which this occurs are when the items are either substitute commodities, which means that people buy one or the other, but not both, or when they items are complementary commodities, where people who buy one item tend to buy the other. For example, a car company might sell one model of SUV and one model of sedan; most people buy one or the other. Thus, sedans and SUVs tend to be substitute commodities. On the other hand, since all cars need tires, we expect increased car sales to result in increased tire sales; cars and tires are complementary commodities.

We could get these relationships for the prices from the demand functions for the two items. For now, we'll assume that the demands are linear in the prices so that:

$$p_1 = c_1 + a_1 q_1 + b_1 q_2 \quad \text{and} \quad p_2 = c_2 + a_2 q_1 + b_2 q_2$$

In these expressions, the coefficients a , b , and c are all constants. The exact values of these constants depend on the relationship between the two commodities being sold.

Open the file **C13 Revenue Exploration** to explore how these coefficients influence the shape of the graph and the decisions that you might make in order to achieve the best possible revenue. When you open the file, depending on your computer's security settings, you may need to click on the "Enable Macros" button in order to make the exploration active. If all is working properly, you should have two slider bars in the upper right corner and moving these around should change the shape of the graph; if it doesn't see the "How To Guide" below for details on adjusting the computer's security settings.

It is important to note that there are, potentially, six constants in the expression that you could change. We have rigged the exploration file, though, so that you can control just two of these with the slider bars, and the other four will change in a particular way. This makes it easier for you to see what is happening on the graph and allows you to focus your attention on the important features. The coefficients that you can change with the sliders are in cells C3 and D4: these represent the quantities a_1 and b_1 in the expressions above for the demand. You will also notice that the discriminant is calculated for you, in cell G1, to help you make some sense of what you are seeing.

Part A. First, move the sliders around to get a feel for how they interact and produce different shapes of the surface. Then concentrate on specific values of the coefficients that produce the different shapes. Finally, for one example of each shape, explain what the values of the coefficients mean in terms of the relationship between the two goods under investigation.

Interpretation of the Graphs

Now, focus on one of your graphs. The method we will use to interpret the graph is referred to as the “method of sections”. The idea is that we fix the value of one of the independent variables; for example, we could let $q_2 = 500$. Now we imagine moving across the surface of the graph, always keeping q_2 fixed, but letting the other variable, q_1 increase. The interpretation follows by thinking about what happens to the dependent variable as the free variable increases at a fixed value of the other variable (the “sectioning variable”). For example, if you push the two sliders all the way to the right, so that cells J1 and J2 show the value of 1000, you have a graph that looks like a hill. Now, imagine setting $q_2 = 500$ and exploring the surface along this path by letting q_1 increase from 0 to 300. You might describe this exploration in the following way:

Along the section $q_2 = 500$, the total revenue seems to be increasing until the point where q_1 is about 200. Up to that point, the revenue is increasing, but at a decreasing rate (the hill is concave down). After $q_1 = 200$, the revenue begins to decrease as q_1 increases.

Similar statements can be made along any section (fixed value of one of the variables). This is very much like our interpretations of multivariable models that we have used before. The main differences are that (1) this is a graphical method and (2) we are referring to this as “sectioning in q_2 ” rather than “controlling for q_2 ” as we did in the algebraic versions.

Part B. Now, for each of the graphs you focused on in part A, describe several sections of the graph. Be sure to section the graph in both of the variables. You may want to change the viewing angle for the 3D graphs to help you visualize the surface better for some sectionings (See the How To guide for this).

13.3 Homework

Mechanics and Techniques Problems

13.1. Answer each of the following questions, given the function of two variables: $f(x, y) = 8xy - 3x^2 + 2y^2$.

1. Find the value of the function when $x = 2$ and $y = 1$.
2. Determine a value of y so that when $x = 10$, the function is equal to 124. You may use algebra, Goal Seek or some other method to find the answer, but explain your solution method.
3. Create a graph of the function of one variable $g(x)$ where $g(x) = f(x, 3)$.

13.2. Using the discriminant identify the shape of the 3D surface plot of each function below. Describe the shape as being either: a bowl, a hill, a saddle, or impossible to tell.

1. $f(x, y) = 2x^2 - 3xy + y^2 + 4x - 5$
2. $g(x, y) = 3x^2 - 2xy + y^2 + 4y - 5$
3. $h(x, y) = -3x^2 + 2xy - y^2 + 4y - 5x + 1$
4. $k(x, y) = -0.3x^2 + 0.2xy - 0.1y^2 + 4y - 5x + 1$

13.3. Get Bent, Inc. sells assembled and unassembled recumbent bicycles. The estimated quantities demanded each year for the assembled and unassembled bikes are x and y units when the corresponding unit prices (in dollars) are

$$\begin{aligned} p &= 2000 - \frac{1}{5}x - \frac{1}{10}y \\ q &= 1600 - \frac{1}{10}x - \frac{1}{4}y \end{aligned}$$

1. Find the annual total revenue function, $R(x, y)$.
2. Find the approximate domain of the revenue function. That is, find the set of values of x and y such that the unit prices are all positive.
3. Create a 3D surface plot of the revenue function for all points (x, y) in the domain.
4. Create a 3D contour plot of the revenue function for all points (x, y) in the domain.

13.4. The revenue function below was developed as a model for the revenue data “Shaken and Stirred” collected regarding its sales of gin (x) and vodka (y). The sales quantities of each are measured in liters. The company would like to know if the revenue function supports the notion that their products are complementary commodities.

$$R(x, y) =$$

1. Factor the expression to put it into the form below. Assume that the mixed term (the xy term) splits equally into the two demand functions.
2. From your factored revenue function, identify the demand functions for gin (x) and vodka (y) sold by Shaken and Stirred.
3. Analyze your demand functions and explain whether the products are complementary commodities or substitute commodities.

13.5. The contour diagram below shows the total revenue from selling two different products.

1. Give at least four sets of production pairs (q_1, q_2) such that the revenue is positive.
2. Give at least four sets of production pairs (q_1, q_2) such that the revenue is greater than 200,000.

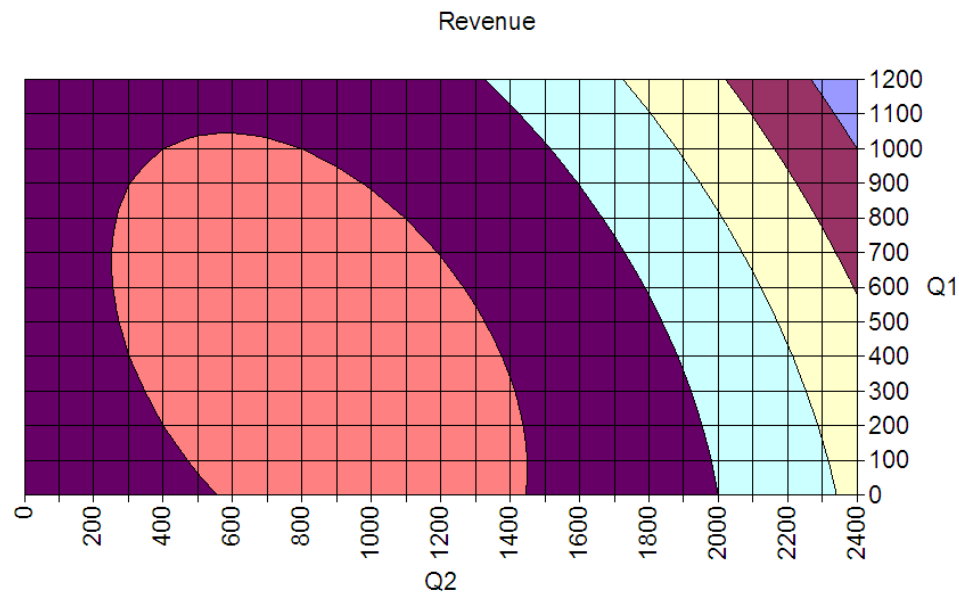


Figure 13.11: Revenue versus quantity of two products being sold, problem 5.

Application and Reasoning Problems

13.6. The graphs below show contour plots of the demand function for one product out of a pair of products sold by the same company. In each graph, the demand function plots the unit price when x and y units of the two products are demanded. Which company is selling two complementary commodities? Which is selling two substitute commodities? Explain your answer.

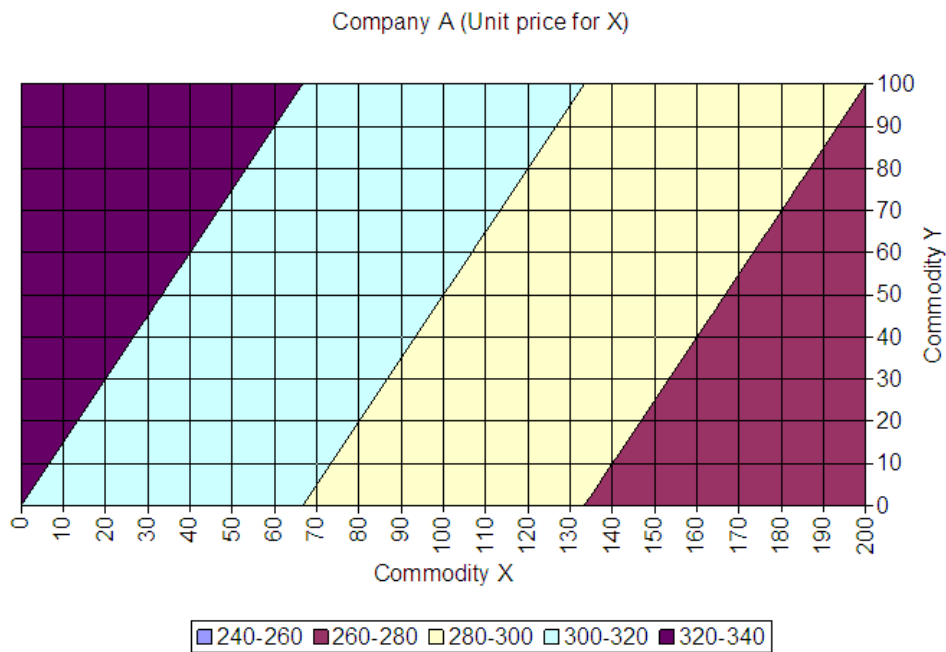


Figure 13.12: Contour plot of demand function for Company A in problem 6.

13.7. Metro Area Trucking has been gathering data regarding a different approach to predicting maintenance costs of its trucking fleet. There is a considerable growing body of research suggesting that uneven tire tread wear is related to maintenance costs for a variety of reasons including worn front end parts, worn or weak suspension, and even the vibrations of a roughly running engine. The surface of the roadway has been shown to affect uneven tire wear, which might relate to maintenance costs even apart from tire wear, and uneven tire wear is a direct contributor to high gasoline costs. Metro has developed an index for measuring uneven tire wear. Every three months the treads of the four tires of a van are each measured in three places by a digital gauge to the nearest 64th of an inch. The standard deviation of the three measurements taken on each tire is calculated and then scaled from 1 to 100 in whole numbers for easy reading. This is called the tire's tread index. The more uneven a tire is, the larger its standard deviation, and the higher its tread index. The largest index measured from the four tires on the van is recorded. The idea is that this index, which is a measure of the driving conditions to which the truck is subjected, interacted with the

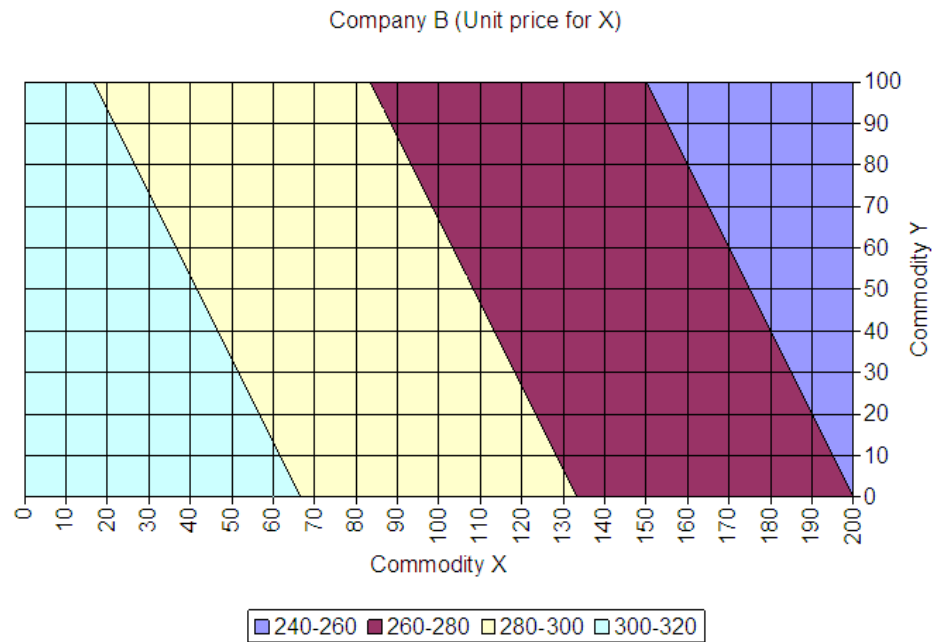


Figure 13.13: Contour plot of demand function for Company B in problem 6.

number of miles the truck is driven, might very well be a good predictor of maintenance cost.

1. From the data in **C13 Truck Data**, build a model with interaction terms (self and joint)
2. Discuss the goodness of fit of your model
3. Interpret the model.

13.8. Consider the following model to explain the number of tickets sold each week in a large metro public transportation system:

$$\begin{aligned} \text{Riders} = & 1486.7960 + 0.0681 \cdot \text{Income} - 29.3 \cdot \text{TicketPrice} - 2.3324 \cdot \text{GasPrice} \cdot \text{GasPrice} \\ & + 1.4625 \cdot \text{TicketPrice} \cdot \text{Income} + 13.8049 \cdot \text{TicketPrice} \cdot \text{TicketPrice} \end{aligned}$$

In this model, the variable “Income” represents average weekly disposable income for a family of four in the greater metropolitan area (in dollars), “TicketPrice” represents the price of a ticket on the transit system (in dollars), and “Gas Price” is the median price for a gallon of regular unleaded gas (in dollars).

But the model, with three variables, is too complicated for explaining to the city council at the upcoming meeting. You have noticed that, within the time span that this model was based upon, you found that

$$\text{Income} = 260.00 - 3.1 \cdot \text{TicketPrice}$$

Use this information to find a simpler way to express the model and interpret the simplified model both algebraically and graphically.

13.9. For a fixed amount of principal, A (in dollars), the monthly payment (\$) for a loan of t years at a fixed APR of r is given by the formula below.

$$P = f(A; r, t) = \frac{Ar}{12 \left[1 - \left(1 + \frac{r}{12} \right)^{-12t} \right]}$$

1. Create a 3D surface plot for the monthly payment of such an amortized loan for a reasonable domain of t and r . Use $A = \$100,000$ as the principal for the loan.
2. Using your graph, what happens to the monthly payments as the interest rate r increases, but the term of the loan (t) stays fixed? Does it depend on the value of t , or is the effect independent of t ? Explain.
3. Using your graph, what happens to the monthly payments as the term of the loan t increases, but the interest rate (r) stays fixed? Does it depend on the value of r , or is the effect independent of r ? Explain.

13.10. Home mortgage rates are designed so that the amount of principal and amount of interest in each payment varies over the life of the loan, but the monthly payment remains fixed. For a loan of A dollars and a term of t years, the total amount of principal paid by the end of month i of the loan is given by the formula below.

$$B = f(A, t; r, i) = A \left[\frac{\left(1 + \frac{r}{12} \right)^i - 1}{\left(1 + \frac{r}{12} \right)^{12t} - 1} \right]$$

1. Suppose you borrow \$100,000 for a home on a 30-year loan at 6.25% APR. How much will you have left to pay after 1 year (12 months)? After 5 years (60 months)? After 15 years (180 months)?
2. Suppose you borrow \$125,000 for a home on a 30-year loan. Create a 3D plot showing the amount of principal remaining after month i at an interest rate of r . Use values of r between 2% and 10%, in intervals of 0.25%. Make sure your graph covers the entire period of the loan.
3. From your graph, what can you infer about the amount of principal in each monthly payment when you are at the beginning of the load repayment? At the end?

Part V

Analyzing Data Using Calculus Models

In this unit, we will explore how calculus tools can help us understand the models with have built from data. In particular, we will focus on the notion of rate of change, which is a more general approach to thinking about slope. As we will see, this notion is powerful, and can help us determine a lot about our models.

For starters, the rate of change will help us generalize the notion of slope from a linear function. Most models do not have a fixed slope; instead, the slope changes depending on where along the model you are currently exploring. Once we understand a little about rate of change, we can use this to find places where our model has a maximum value or a minimum value, which is useful for decision making purposes. For example, if our model represents the profit from selling a quantity of items we produce (for some reason, everyone refers to items as widgets when they don't have a good name for them), then finding the maximum point on the model will help us know how many widgets to make in order to achieve the most profit possible, based on our assumptions about the market that were used to build our profit model.

Rate of change is a concept you are probably familiar with already. Slope is the linear version of it. In calculus, we study rate of change under several different names. Usually, we refer to it as the derivative. Sometimes it is referred to as the instantaneous rate of change. This is a notion that makes some sense. Consider driving in your car. If you drive 100 miles in 2 hours, you averaged 50 miles per hour, which is the average rate of change of your distance from your starting point. But it is highly unlikely that at every instant during the two hours you were going exactly 50 mph. At some point, you were probably stopped at a light; at some point you sped up to pass a slower car. Your average rate of change was 50 mph, but at each moment, you have a tool that tells you the instantaneous rate of change (derivative) of your distance from home: the speedometer of your car. If you graphed the distance from your starting point as a function of time, you would probably not see a straight line (which would give you a constant rate of change). It would be twisty and curvy, always increasing in distance from the starting point, but at many different speeds. The speedometer, though, always gives you the rate of change at that instant on the curve.

The other half of calculus is about something called the integral. In chapter 17 you will get a brief introduction to this concept, and you will see how it is related to finding areas under a curve. It turns out, though, that there is a remarkable mathematical theorem called the Fundamental Theorem of Calculus that relates this way of computing areas to the derivative! Thus, these two operations, finding slopes and finding areas, are inverses of each other. Throughout the unit, you will be exploring deep ideas in calculus, but we'll focus on key concepts and examples and will constantly be applying this to the business setting, so don't get too worried; we will not by any means, deal with all the complexity that can possibly exist in studying calculus.

Optimization and Analysis of Models¹

This chapter is designed to help you take your knowledge of building models to the next level - applying them to solve problems involving questions about optimization. In general, optimization is the process of trying to make something as efficient as possible, or as large as possible or as cheap as possible. It's the study of minimizing or maximizing a quantity, like profit, as a function of some other quantity, like production. In order to optimize a quantity, though, we need a few things. The first is a skill you already have - the ability to create a model equation that represents how the quantity to be optimized varies as a function of some other quantity. For example, we might produce a model equation describing how the profits of a company depend on the number of items they produce, since the more you produce: (a) the more you can sell, generating more revenue but (b) the more it costs, in labor and materials. The other tool that you need is a knowledge of marginal analysis, which measures how a change in the independent variable will cause a change in the dependent variable in a model. We will focus our study on the marginal analysis and optimization of polynomial models, although this is only the tip of the iceberg.

- Section 14.1 introduces the concept of a **derivative** and shows you how to compute them for power functions.
- In section 14.2 the derivative is used to find the maximum and/or minimum values that a model takes on.

¹©2017 Kris H. Green and W. Allen Emerson

As a result of this chapter, students will learn

- ✓ What marginal analysis is
- ✓ How to interpret the results of marginal analysis
- ✓ What the derivative of a power function is
- ✓ What the derivative of a polynomial function is

As a result of this chapter, students will be able to

- ✓ Compute the derivative of a power function
- ✓ Compute the derivative of a polynomial
- ✓ Maximize or minimize a polynomial, using both algebra and software tools

14.1 Calculus with Powers and Polynomials

We have spent some time discussing the basic families of functions. These functions can be used to model the behavior of various real-world business situations. For example, suppose we have data based on the total cost of paying back a loan (for a fixed principal and fixed payback period). We can use this data to develop a function, call it $C(r)$, which represents this cost as a function of different interest rates on the loan. Suppose interest rates are increasing. How will this affect the cost of paying back the loan?

This question really centers on how the function C changes as the interest rate r increases. To answer this question, we will turn to our knowledge of families of functions. In particular, we will use what we know about the parameter A in the general formula for a linear function, $y = A + Bx$.

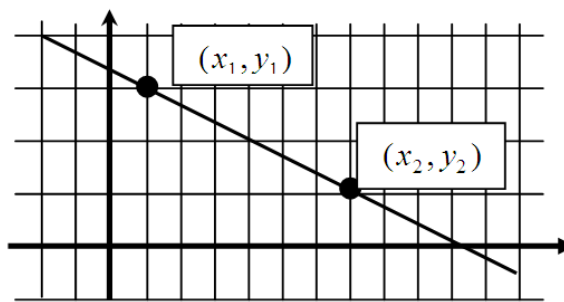


Figure 14.1: Slope between two points.

Look at the graph of the linear function shown in figure 14.1. Also shown on the graph are two points. These points are labeled with the coordinates (x_1, y_1) and (x_2, y_2) . What is the total change in the linear function between the two points?

Between these points, there is a change of $y_2 - y_1$. This is just the vertical separation between the two points. Now, how quickly is the function changing at the first point? This is not a question of total change, but of the rate of change of the function. Another way of asking this question is “If I make a small change in x from x_1 to x_2 , how much will the function change?” To answer this question, we look at the slope of the line. As you may recall, the slope of a line can be calculated from the formula

$$\text{slope} = A = \frac{y_2 - y_1}{x_2 - x_1}.$$

For the function above, we see that the two points have coordinates $(1, 3)$ and $(7, 1)$. Thus, the slope of the line is $(1 - 3)/(7 - 1) = -2/6 = -1/3$. The negative tells us that the function (in this case a straight line) is decreasing. This means that, as we move from left to right, the value y of the function gets smaller. There are several nice things about straight lines that we can see from this example. First, unlike nonlinear functions, the slope of a straight line is exactly the same at every single value of x . This means that the slope of the function at the first point is $-1/3$ and slope at the second point is also $-1/3$ and the slope at $x = 249$ is also $-1/3$. Second, it is easy to calculate the slope of a straight line. We

simply look at the change in the values of the function (the y values) and divide this by the change in the x values between the two points. This will not hold for any other family of functions.

To find the slope of a nonlinear function, we take advantage of a property of smooth functions. As illustrated in the graphs in figure 14.2, if we have the graph of a nonlinear function, and we zoom in on the graph, it begins to look linear.

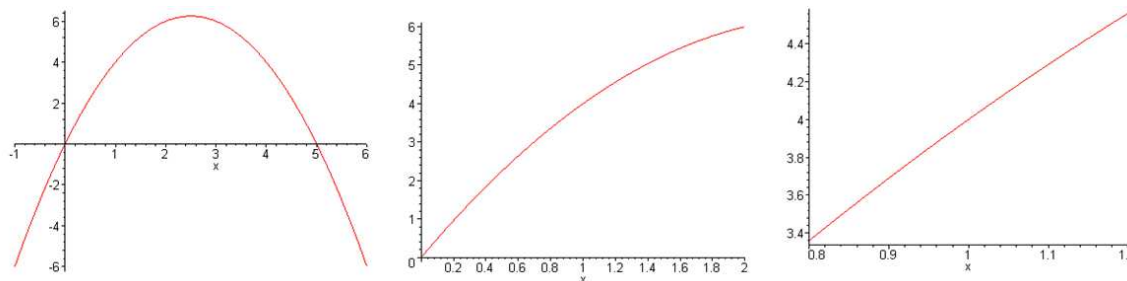


Figure 14.2: Series of graphs showing how the function changes as we zoom in on $x = 1$.

For some functions, we need to zoom in more, and for others we zoom in less to see this linear-like appearance. In order to calculate the slope, we will use this feature, called local linearity, to determine the slope of a function at any point. Specifically, if we pick two points on the function, and draw a line between them, we will call the slope of this line the average rate of change of the function. If we call these two points $(x_1, f(x_1))$ and $(x_2, f(x_2))$, then the average rate of change between the points is

$$\text{average rate of change} = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Notice that the graph in figure 14.3 shows how the average rate of change can be quite different from the actual rate of change (called the instantaneous rate of change or derivative).

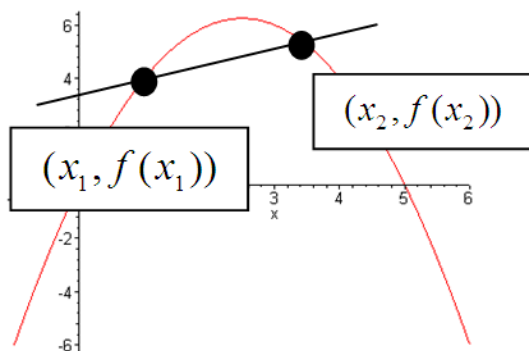


Figure 14.3: Average slope between two points.

However, if we move the second point closer to the first, we can get a more accurate approximation to the instantaneous rate of change of the function near the first point. If the

two points are close enough, the average rate of change will be a very good approximation to the instantaneous rate of change. This fact will help us in many cases where we only have data, instead of an actual function.

14.1.1 Definitions and Formulas

Quotient A quotient is simply the result of dividing one quantity by another quantity.

Average slope The average slope between two points on a function is what you get when you start with a function (f), evaluate it at two points (say x_1 and x_2) and then take the difference of these values, $f(x_2) - f(x_1)$ and divide it by the distance between the two x -values ($x_2 - x_1$). Thus,

$$\text{average slope} = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Note that the order is important! If you start with x_2 first in the numerator, you must also start with x_2 in the denominator. The graph below shows the basic idea and illustrates why it's called average slope and not the actual slope. The dashed line between the two points represents the average slope of the function (the curved line) between those two points. In between the two points, though, notice that there are places where the curve has a more negative slope than the average slope and places where the slope is even positive!

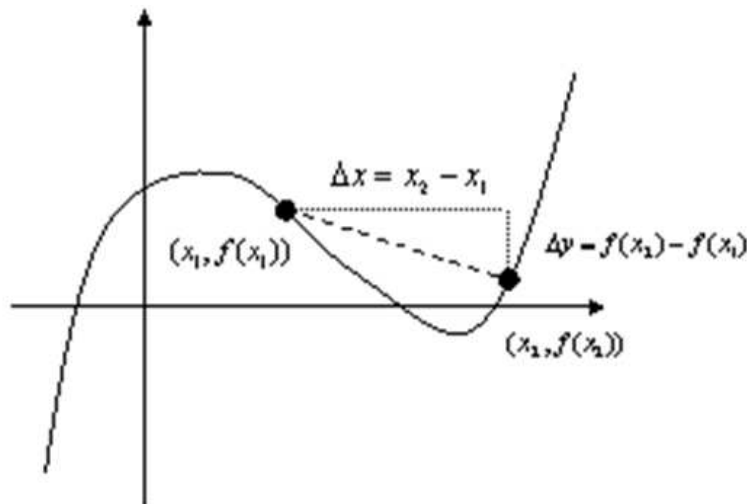


Figure 14.4: Average slope between two points.

Difference quotient The difference quotient is another way of writing the average slope. Instead of looking at the average slope between x_1 and x_2 , we look at the average slope between x_1 and $x_1 + h$, where we think of h as a small number. So $x_1 + h$ is another way of writing x_2 . This form of x_2 allows us to focus on how the function changes at

x_1 . Using $x_1 + h$ in place of x_2 changes the denominator of the average slope formula. Instead of $x_2 - x_1$, we have $(x_1 + h) - x_1 = h$. So, the average slope formula takes on a new name and a new look:

$$\text{Difference quotient} = \frac{f(x_1 + h) - f(x_1)}{h}$$

Consider the line passing through the point $(x_1, f(x_1))$ and having the same slope as the difference quotient, with a fixed value of h , say 1. If we look at this line for smaller and smaller values of h (say 0.1, 0.01, 0.001, etc.) we see that the line eventually becomes “parallel” with the function at the point $(x_1, f(x_1))$. This visual process of watching the line become parallel can be carried out mathematically through a limit.

Marginal Analysis This is a financial/business term for the process of finding the instantaneous rate of change of a function at a point. Essentially, this is a difference quotient, and it is useful for answering the question “If my independent variable increases by 1 unit, how much will my dependent variable increase (or decrease)?” Another way to think of this is:

How much bang do I get for each additional buck that I spend?

Marginal Cost Basically, when the word “marginal” is followed by a term like “cost”, it means that you are looking at the instantaneous rate of change of the cost function, which is just its derivative.

Marginal Profit Instantaneous rate of change of the profit function.

Marginal Revenue Instantaneous rate of change of the revenue function.

Derivative function The derivative function is a function derived from the slopes of another function. Basically, at each point $(x, f(x))$ the function has a slope, usually denoted by $f'(x)$. If we collect all these slopes into a new function, so that plugging in a value of the independent variable, x , results in the slope of f at that point, then we have the derivative function. The derivative of a function at a point is also denoted by the notation $\frac{\partial f}{\partial x}$ which indicates that we are interested in the slope of f in the x -direction. Thus, a positive number for the derivative means that as x increases (we always move to the right) the value of f is increasing. Likewise, a negative value of the derivative indicates that the function is decreasing at that point. Officially, the derivative of a function at a point is computed by taking the difference quotient and letting h go to zero. This is noted mathematically by the “limit of the difference quotient”:

$$f'(x) = \lim_{h \rightarrow 0} \left[\frac{f(x + h) - f(x)}{h} \right]$$

Second derivative Since the first derivative of a function is (usually) a function itself, we can take its derivative. We refer to the derivative of the derivative of a function as

the second derivative. It is denoted by f'' or $\frac{\partial^2 f}{\partial x^2}$. Since the derivative tells how fast the function is changing, the second derivative tells us how fast the first derivative is changing. Thus, it measures the rate of change of the slope, which is called concavity. In a graph, concavity is easy to see: it refers to the direction and steepness of the way the function bends. If it bends up (looks like a cup) then the concavity is positive. If it bends down (looks like a frown) then the concavity is negative. If the function is almost flat, then the concavity is close to zero.

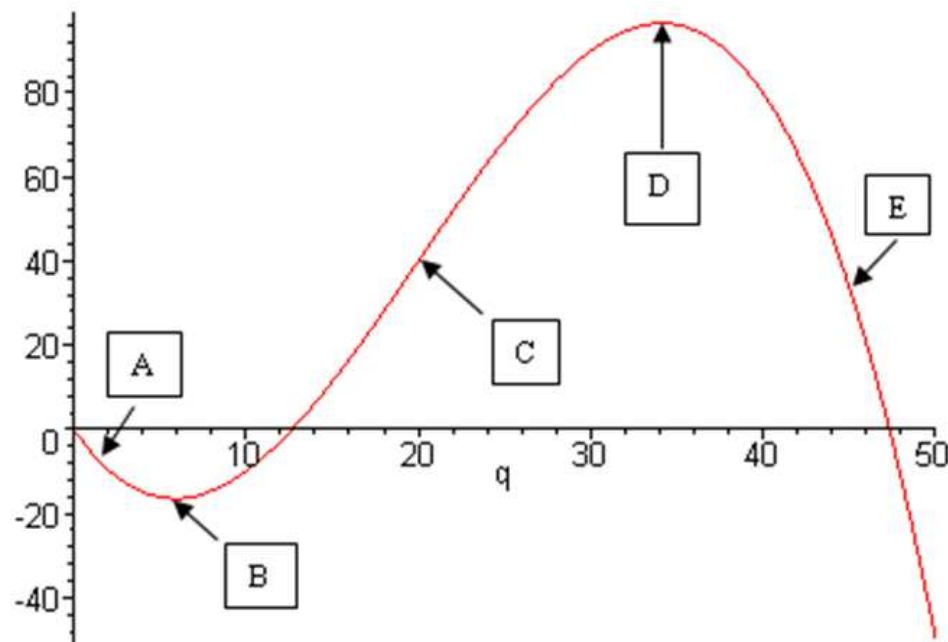


Figure 14.5: Graph and explanation showing the connections between f , f' , and f'' .

In this graph, there are five points marked A - E. The function and its derivatives are described at each of these points below.

- A. Here the function is negative, the slope is negative (it is a decreasing graph) and the second derivative (concavity) is zero, since the graph is basically flat. Thus, $f(A) < 0$, $f'(A) < 0$, $f''(A) = 0$.
- B. Here we have the function negative (it is below the x -axis, the line $y = 0$). The slope is zero, since the graph is horizontal at this point. The concavity is positive since the graph is curving upward. Thus, $f(B) < 0$, $f'(B) = 0$, $f''(A) > 0$.
- C. Here the function is positive, the slope is positive (it is an increasing graph) and the second derivative (concavity) is zero, since the graph is basically flat. Thus, $f(C) > 0$, $f'(C) > 0$, $f''(C) = 0$.
- D. Here we have the function positive (above the x -axis, $y = 0$), the slope is zero, since the graph is horizontal at this point, and the concavity is negative since the graph is curving downward. Thus, $f(D) > 0$, $f'(D) = 0$, $f''(D) < 0$.

- E. Here the function is positive, the slope is negative (it is a decreasing graph) and the second derivative (concavity) is zero, since the graph is basically flat. Thus, $f(E) > 0$, $f'(E) < 0$, $f''(E) = 0$. In addition, since the function much steeper at point E than at point A, we know that the slope at E is more negative. Thus, we can also say that $f'(E) < f'(A)$.

14.1.2 Worked Examples

Example 14.1. Marginal Analysis with Difference Quotients

In example 1 we developed a model for the cost of electricity as a function of the number of units of electricity produced. Later in that chapter we used parameter analysis to explore how the function behaved. This analysis was all in terms of percent changes, which is somewhat limiting. In this example, we are going to use marginal analysis through the difference quotient to interpret how much each unit of electricity affects the total cost of producing the electricity. (In later examples, we will refine this process using a shortcut method called the derivative.) The cost model we will use is the square root model given by

$$\text{Cost} = 6,772.56 + 1,448.74 \cdot \text{Sqrt}(\text{Units}).$$

Suppose that we are currently producing 500 units of electricity. How much would it cost to produce one more unit of electricity? We can put this into a spreadsheet to compute it fairly easily. The results are shown below, and were obtained by setting up a formula for the difference quotient of the function, with a variable for h so that we can let h get very small. This lets us see what the instantaneous rate of change of the cost function is.

A	6772.56				
B	1448.74				
X	500				
H	X+H	F(X)	F(X+H)	DF=F(X+H)-F(X)	DF/H
10	510	39167.37	39489.72	322.3443693	32.23444
1	501	39167.37	39199.75	32.37862999	32.37863
0.1	500.1	39167.37	39170.61	3.239319164	32.39319
0.01	500.01	39167.37	39167.7	0.323946492	32.39465
0.001	500.001	39167.37	39167.4	0.032394795	32.3948

From this, it seems that when current production is at 500 units, each additional unit of electricity will cost approximately \$32.39. In contrast, if are currently producing 1,000 units of electricity, the marginal cost is about \$22.91 per unit.

A	6772.56				
B	1448.74				
X	1000				
H	X+H	F(X)	F(X+H)	DF=F(X+H)-F(X)	DF/H
10	1010	52585.74	52814.24	228.4960877	22.84961
1	1001	52585.74	52608.64	22.9008669	22.90087
0.1	1000.1	52585.74	52588.03	2.290601805	22.90602
0.01	1000.01	52585.74	52585.97	0.229065334	22.90653
0.001	1000.001	52585.74	52585.76	0.022906585	22.90658

Example 14.2. Finding the Derivative of a Power Function

While it is possible to use basic algebra and the definition of the derivative (as a limit of the difference quotient) for marginal analysis this process can be tedious and will be difficult for some of the basic functions. Instead, we're going to use trendlines to experiment to find a shortcut for the derivative of a power function. We begin with the power function $f(x) = x^2$. Here's an outline of what we'll do:

1. Set up a spreadsheet that has places to enter the parameters of the function (A and B).
2. Add in a place for us to enter a value for h , the number we need in the difference quotient.
3. Create columns for x and $f(x)$ and compute the values for the column under $f(x)$ from the values listed under the x column.
4. Next we add columns to compute $x + h$ and $f(x + h)$.
5. We add a column to compute the difference quotient from the data we've already set up.
6. Since we have lots of x values (running down the table,) we now have a bunch of points of the form $(x, \text{difference quotient of } f \text{ at } x)$. If we make a scatterplot of these points, we can fit a trendline to these data and determine the equation of the difference quotient in the process. Thus, we are close to experimentally determining an equation for the derivative function.
7. Up till now we have kept h fixed. We can then simulate the limit of the difference quotient by making h a smaller and smaller number, until we think we see what the "real" equation would be with h equal to zero. (Note that we can't actually set h to be zero, since we would be dividing by zero, which gives an error!)

The screen shots below will show you what our spread sheet looks like at the end of this process. To go through this procedure, open the file **C14 SquareDerivative**. Starting with the power function

$$f(x) = Ax^B = x^2.$$

and setting h initially to 0.1, and listing x values from 0 to 10 in steps of 0.5, we get the table of data (using steps 1-6 above) shown in figure 14.6.

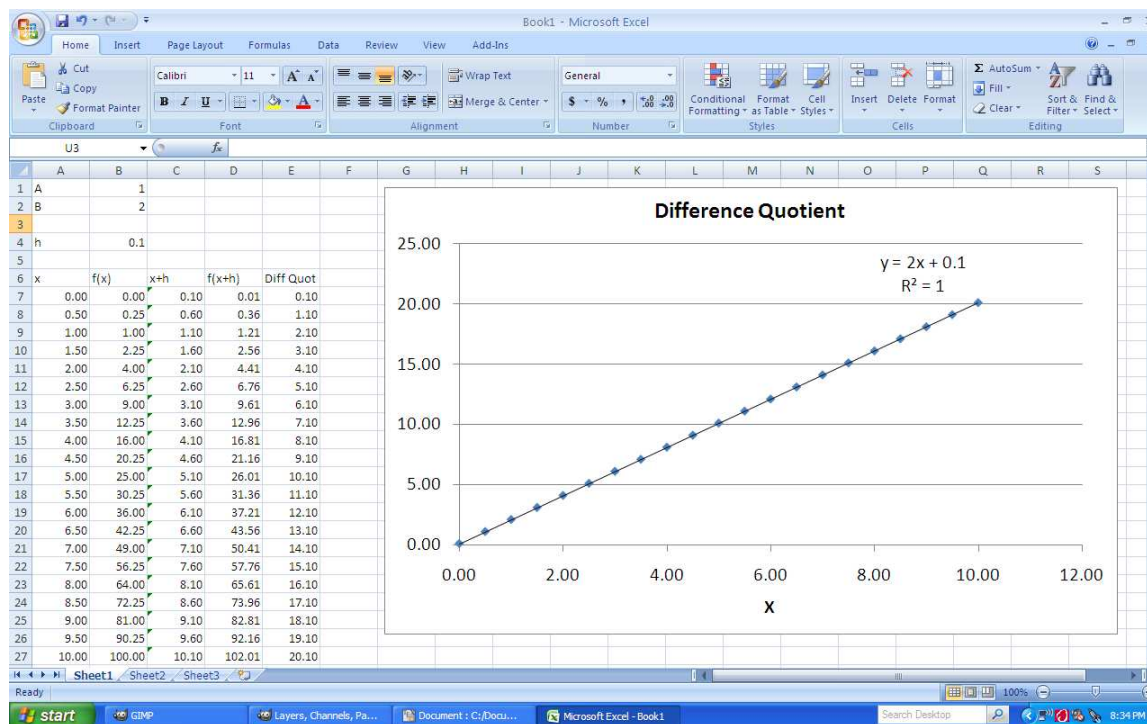


Figure 14.6: Difference quotient worksheet.

Now, what kind of trendline does the difference quotient make? It looks like a straight line, so let's add a linear trendline to the graph. From this we get a fairly accurate equation:

$$y = 2x + 0.1 \text{ with } R^2 = 1.$$

But, we have a (mathematically speaking) pretty large value of h . Let's vary h and collect the results of the trendline into a table like the one below. Notice that as h gets smaller, the y -intercept of the trendline decreases. Since the derivative is the limit as h goes to zero of this difference quotient, we can reasonably conjecture that as h shrinks down to zero, so does the y -intercept, leading us to the following simple rule:

The derivative of the function $y = x^2$ is the function $y = 2x$.

h	y	R^2
0.1	$y = 2x + 0.1$	1
0.01	$y = 2x + 0.01$	1
0.001	$y = 2x + 0.001$	1
0.0001	$y = 2x + 0.0001$	1
0.00001	$y = 2x + 0.00001$	1
0.000001	$y = 2x + 0.000001$	1
0.0000001	$y = 2x + 0.0000001$	1

However, this only gives us the derivative of one single power function. What about all the other ones? How can we determine their derivatives without going through this fairly lengthy process every time? We've actually almost got the answer, since our spreadsheet is set up to allow us to change the parameters in the power function and find rules for those as well. This is what the exploration in this section is all about - finding the rules for ALL of the power functions. It turns out to be relatively simple.

Example 14.3. Marginal Analysis with Derivatives

Suppose we know that our costs for producing q thousand goods are $C(q) = q^2$, where C is measured in millions of dollars. If we are currently producing 10,000 goods, how will our costs increase if we add an additional 1,000 goods to the production?

For this situation, we are currently producing $q = 10$ thousand items and want to know what happens to the cost if we produce $q = 11$ thousand items. This is an increase of 1 (in our units of q) so it is a question about the marginal cost. Since the marginal cost is really just the derivative (slope) of the cost function, we can use the last example to help us out. In that example, we used spreadsheets, difference quotients, and regression to learn that the derivative of x^2 is $2x$. Thus, the derivative of the cost function, denoted C' , is $C'(q) = 2q$ and the marginal cost of producing 10,000 goods is $C'(10) = 2(10) = 20$. The units of the derivative are (units of function/units of independent variable) so the complete answer is:

If the cost of producing q goods is $C(q) = q^2$, where C is measured in millions of dollars and q is measured in thousands of goods, then the marginal cost of producing 10,000 goods is 20 million dollars per thousand goods.

This means that if we want to increase production to 11,000 goods, we can expect an increase in the costs of about 20 million dollars. If we wanted to produce 12,000 goods the cost would increase by approximately $(\$20 \text{ million per thousand goods}) \cdot (2 \text{ thousand goods}) = 40 \text{ million dollars}$ (since 12,000 is 2,000, or 2 thousands, greater than 10,000.) If, on the other hand, we decrease the production to 9,500 goods, then the cost will change by about $(\$20 \text{ million per thousand goods}) \cdot (-0.5 \text{ thousand goods}) = -10 \text{ million dollars}$. (The negative sign simply means that the cost decreases if we decrease production.)

14.1.3 Exploration 14A: Finding the Derivative of a General Power Function

Using the file C14 PowerDerivative try to determine the shortcut derivative rules for general power functions. Phrase each of your rules as a sentence like the one in italics in example 2. The tables below may help you to organize your results in order to make sense of them. For each, you should probably use $h = 1\text{E-}6$ or smaller.

- A. Start with changing A and see what that does. Complete the following table to help you record your observations and make conjectures about the general form of the derivative of the function $f(x) = Ax^2$.

A	B	$F(x)$	$F'(x)$

Your sentence describing the shortcut rule:

- B. Now set $A = 1$ and see if you can find the derivative rule for $f(x) = x^B$. Start with integer powers of B to find the pattern, then test your pattern for non-integer values of B . You may need to delete the row containing $x = 0.0$ from the data table in order to use the appropriate trendline.

A	B	$F(x)$	$F'(x)$

Your sentence describing the shortcut rule:

- C. Finally, try to combine your rules above to find the general shortcut rule for the derivative of the function $f(x) = Ax^B$.

A	B	$F(x)$	$F'(x)$

Your sentence describing the shortcut rule:

- D. For the ultimate challenge, try to find out what the derivative rule for polynomials is. Start with a simple one, like $f(x) = x^3 + x^2 + 1$ and see if you can figure out what happens. (Hint: Polynomials are really just sums of power functions with non-negative integer powers.)

14.2 Extreme Calculus!

Now that we've learned a little about marginal analysis, we can apply this knowledge to help answer questions that are really important. For example, suppose we would like to minimize the cost of producing our product, working on the theory that this will save us money. How would we go about this process of optimizing the cost?

First of all, we need to know what causes the cost of production to vary. Typically, the simplest quantity that determines production cost is, you guessed it, the number of items that we produce. After all, each one of them uses a certain amount of materials that aren't free; each one of them requires labor; production probably involves machines which use electricity and so forth. So, we could start by getting together data that shows the total cost each month (or week or whatever) along with the total cost of production that month (or week or whatever). We can then use our model-building skills to determine an equation that represents the cost of production as a function of the number of items produced.

Now, how can this help us find the amount of production that will result in the lowest overall cost? We actually have several tools available. We could create a table of values from the function and look for the lowest cost. That could be difficult, though, since our table will only show some of the possible values: it may be that we skip over the best spot if we're not very careful. We could also graph the function, but then scale is an issue; we may have to keep redrawing the graph on larger and larger scales to see where this minimum occurs. The most commonly used approach, though, is based on marginal analysis.

Think about it this way. We could imagine "walking" along the function in the direction of increasing production. As we do this, the slope along which we climb is determined by the rate of change of the function - marginal analysis. If the marginal cost is negative, we are going downhill; this means that by increasing the production we can decrease the costs a little. If the slope is very large and negative, then we are far from the minimum cost. As we get closer to the minimum of the cost, this slope will level out. In fact, if we go too far, we could wind up increasing the costs - like climbing out of a hole. That means that we need to go back in order to decrease the cost.

This idea of walking along the function is a little hard to implement on a computer. It's much easier to think about what the function must look like near the minimum cost. We know that on one side of the minimum, the slope is negative, because we are decreasing the cost as we increase production. On the other side of the minimum (we've gone too far!) the slope is positive. Now the slope is the marginal cost. This is a number associated with each value of production. If it is negative on one side of the minimum, and positive on the other side of the minimum, then we can conclude (assuming a mathematical property called continuity) that at the minimum, the slope (marginal cost) is exactly zero. This basic idea can be used to solve any optimization problem - simply set the marginal whatever to zero and solve the resulting equation.

14.2.1 Definitions and Formulas

Critical point Any point on the graph of function f where the derivative is zero is a critical point. Thus, we can find all the critical points by solving the equation $f'(x) = 0$. Often, this will be a nonlinear equation and will require some algebra to solve.

Extrema An extrema is some “extreme” point on a function: either a maximum or a minimum.

Local Maximum A local maximum is a point on the graph of a function that is higher than all the points that are close by it. Thus, the point looks like the top of a hill. Point D in the graph at the end of the definitions from the last section is an example of a local maximum. See the graph below.

Local Minimum A local minimum is a point on the graph of a function that is lower than all the points that are close by it. Thus, the point looks like the bottom of a valley. Point B in the graph at the end of the definitions from the last section is an example of a local minimum. See the graph below.

Global Maximum A global maximum is the highest point on a function anywhere not just when compared to points near it. Most functions have lots of hills and valleys; only the highest peak in the “mountain range of the function” would be the global maximum. See the graph below.

Global Minimum A global minimum is the lowest point on a function anywhere not just when compared to points near it. Most functions have lots of hills and valleys; only the lowest valley in the “mountain range of the function” would be the global maximum. See the graph below.

Optimization This is the process of finding and classifying all the extrema for a function and then using this to solve some problem. For example, we may have a function that describes our profits from manufacturing a quantity q of a good. Optimization would help us answer the question “How many of this good should we make in order to get the highest profit?”

Second Derivative Test Solving the equation $f'(x) = 0$ only finds extreme points. You then need to classify the points as maxima or minima (the plurals of maximum and minimum, respectively). One way to do this is by graphing the function. The other way is by evaluating the second derivative of the function at the critical point. If the second derivative is negative, you have a maximum (the graph is concave down, as at Point D in figure 14.5, pge 339.) If the second derivative is positive, you have a minimum (the graph is concave down, as at Point B in 14.5.) If the second derivative is zero, then you don’t have a maximum or a minimum, necessarily.

14.2.2 Worked Examples

Example 14.4. Using optimization to sketch polynomials

This example assumes that you have learned (from the last section) the following derivative rule:

The Sum Rule for Derivatives: The derivative of the sum of two functions is the sum of the derivatives of the two functions. In other words, $(f + g)' = f' + g'$.

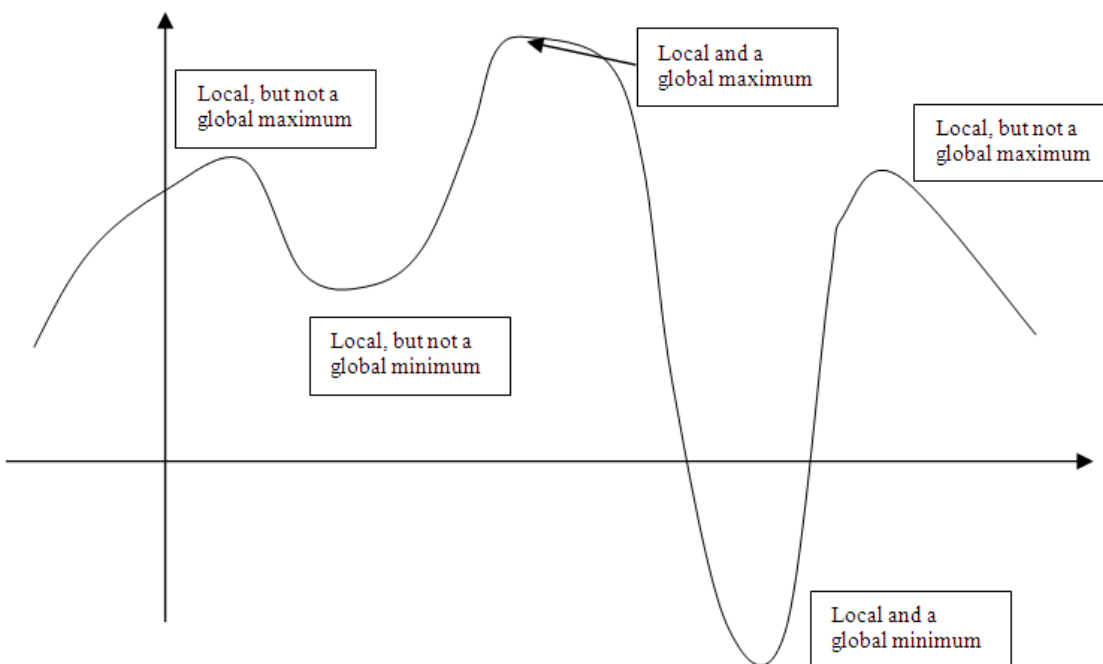


Figure 14.7: Example of a function with several local maxima and minima.

Since a polynomial is just a sum of power functions, we can use this rule to determine the derivative of a polynomial: It's just the sum of the derivatives of the individual power functions that make up the polynomial. Thus, the derivative of the polynomial $f(x) = 3x^4 - 5x^3 + 2x - 7$ is just $f'(x) = 12x^3 - 15x^2 + 2$. (The derivative of $3x^4$ is $12x^3$. The derivative of $-5x^3$ is $-15x^2$. The derivative of $2x$ is $2(1)x^{1-1} = 2x^0 = 2$, and the derivative of a constant is zero.) We can use this to learn about the properties of polynomials and what they look like.

For example, suppose we have a general fifth degree polynomial. Thus, the function can be written (generally) as $g(x) = a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, where the a 's represent constants. What would the derivative of this function be? Well, we apply the power rule to each term and get: $g'(x) = 5a_5x^4 + 4a_4x^3 + 3a_3x^2 + 2a_2x + a_1$. This is a fourth degree polynomial, as expected. How can this help us visualize the graph of g ?

For starters, notice that if we try to find all the critical points of g we will have to solve the equation $g'(x) = 0$. This is a fourth degree polynomial equation and can have, at most, four solutions. Thus, there are at most four critical points in the graph of g . If we were to locate these critical points, we could begin to sketch the graph. Let's take the specific polynomial $h(x) = 6x^5 + 15x^4 - 130x^3 - 210x^2 + 720x + 300$. Its derivative is $h'(x) = 30x^4 + 60x^3 - 390x^2 - 420x + 720$. To find the critical points, we set this derivative equal to zero and solve the equation. Since this equation can be factored as

$$0 = 30(x^4 + 2x^3 - 13x^2 - 14x + 24) = 30(x - 1)(x + 2)(x - 3)(x + 4)$$

we see that the derivative is zero at the points where $x = 1, -2, 3, -4$. There are four

critical points. By plugging them into the function, we can find the y -coordinate of these points, and then graph them. Finally, we notice that since the leading term is a fifth degree power function with a positive coefficient, the function is increasing to the right. Since it is an odd-degree polynomial, it must do the opposite on the left, so it decreases to the left. In the end, we can sketch the graph quite accurately.

Example 14.5. Maximizing Profits with Derivatives

Suppose that the cost of producing q goods is

$$C(q) = 0.01q^3 - 0.6q^2 + 13q$$

and we sell these goods for \$7 apiece. How many of our product should we make (and sell) in order to maximize our profit?

The revenue function will be $R(q) = 7q$. This comes from the fact that revenue is simply the number of products sold times the selling price per product. The profit function (remember: profit = revenue - cost) will then be

$$P(q) = 7q - 0.01q^3 + 0.6q^2 - 13q.$$

The marginal profit is given by the derivative of the profit (which can be computed using the rules we have developed so far). We find that

$$P'(q) = 7 - 0.03q^2 + 1.2q - 13 = -0.03q^2 + 1.2q - 6.$$

We set this function equal to zero and do some algebra (anmely, the quadratic formula) to find that when $q = 5.86$ and $q = 34.14$ the profit function has a critical point. We can also get these results by entering the following data in our spreadsheet.

	A	B
1	q	1
2	P'(q)	=-0.03*B1 ² +1.2*B1-6

In Excel, we can then use the goal seek procedure with the following information. “Set Value” to B2, “To value” 0, and “By changing” B1. Note that this will only locate the first extreme point, $q = 5.86$. To be sure you do not miss the other points, it is good to first graph the function and visually locate some values that are close the extreme points. Then enter one of these values in cell B1. From the graph of $P(q)$ we find that there is an extreme point near $q = 30$. If we enter 30 in cell B1 and then repeat the goal seek procedure described above (with B2, 0, and B1 in the “Goal Seek” dialog box) the computer will locate the other extreme point. In R, we can use the `uniroot` to accomplish the same analysis.

Now, which of these two points is a maximum and which is a minimum? To answer this, we’ll apply the second derivative test. This is simple; we just find the second derivative of the profit function and evaluate its sign at each critical point. The second derivative of the profit function is just the derivative of the first derivative. So, we find

$$P''(q) = -0.06q + 1.2.$$

We then compute easily that $P''(5.86) = 0.8484$, which is positive. This indicates that the point $(q = 5.86, P = -16.57)$ is a local minimum - not a good place to be! At the other point we find that $P''(34.14) = -0.8484$, so the point $(q = 34.14, P = 96.57)$ is a local maximum for the profit function. That's where we want our production and sales!

This tells us that if we sell our product at \$7 each and incur a cost given by the function above then we can achieve a maximum profit of \$96.57 dollars by making and selling 34.14 units of our product.

Example 14.6. Minimizing Average Cost

Suppose that we have a fixed cost of \$2000 each month. This cost includes electricity, rent, and equipment. In addition, if it costs us \$12 per good manufactured (including materials and labor), we have a total monthly cost of

$$C(q) = 12q + 2000.$$

Suppose that instead of minimizing the total cost, we now we want to minimize the average cost function. The average cost function, $\bar{C}(q)$, is basically the cost function divided by the quantity produced (i.e., average cost = total cost of making q goods divided by q .) Thus, the average cost function for this scenario is

$$\bar{C}(q) = 12 + \frac{2000}{q}.$$

This is not a polynomial (the $1/q$ term is really q^{-1} , which not a positive integer power) but we can use the sum rule and product rule to get its derivative:

$$\bar{C}'(q) = 0 + 2000 \frac{d}{dq} (q^{-1}) = 2000(-1)q^{-2} = \frac{-2000}{q^2}.$$

Now, this function is not like our other examples: there is no minimum! We cannot solve the equation $\bar{C}'(q) = 0$ because no value of q will solve this. However, we notice that as q increases, the derivative of the average cost decreases (the derivative is always negative.) This means that making more of our product will always reduce the average cost.

If, instead, we had a slightly more realistic cost function (taking secondary costs into effect) like

$$C(q) = 0.05q^2 + 12q + 2000,$$

then we can optimize the function. Following the same steps as before, we get the average cost function as

$$\bar{C}(q) = 0.05q + 12 + \frac{2000}{q}$$

and we find its derivative as

$$\bar{C}'(q) = 0.05 - \frac{2000}{q^2}.$$

Setting this derivative to zero, we get

$$0 = 0.05 - \frac{2000}{q^2} \rightarrow 0.05 = \frac{2000}{q^2} \rightarrow q^2 = \frac{2000}{0.05} \rightarrow q = \sqrt{\frac{2000}{0.05}} = 200.$$

Thus, to minimize the average cost of producing the goods, we should make 200 goods. This is especially useful, since the cost function itself is only minimized for a negative number of goods! (Try it. You should get the derivative of the cost function as $C'(q) = 0.1q + 12$ which is minimized at $q = -12/0.1 = -120$ goods.)

14.2.3 Exploration 14B: Simple Regression Formulas

We have made extensive use of simple regression so far in this book. But how does simple regression work? How does the computer know how to compute the slope and y-intercept of the line that will minimize the total squared error in our approximation to the data? Wait a minute. That phrase “minimize the squared error” sounds important. It sounds like we can use calculus to find the answer.

First, let’s do this with an example. Consider the following data points. We want to find the best fit (least-squares) regression line for these data.

x	1	2	3	4	5
y	5	7	8	9.5	12

What we want to do is to minimize the total squared error between the data and the regression line. If the line has the regression equation $y = A + Bx$, fill in the rows of the spreadsheet (file **C14 Regression**) with the appropriate calculation for each data point. (For now, just guess a value of the slope and y-intercept. Place these values as parameters on the spreadsheet.) Now, add up all the squared errors to get the total error, $E(A, B)$. This is a function of two variables, and we could treat it with calculus directly, but we’ll simplify everything slightly by noting that the regression line always passes through the point (\bar{x}, \bar{y}) which means that $\bar{y} = A + B\bar{x}$. Rearranging this, we get $A = \bar{y} - B\bar{x}$. Now, let’s put all this into the spreadsheet. You should have a sheet that looks a lot like the one below.

Now that we have the formulas entered, we can minimize the error function using the solver routine in Excel or the `uniroot` in R. Click on the cell containing the error value. Then click on “Tools/ Solver” and enter the values as shown in the screen shot below. It should very quickly find the value of the slope (B) that minimizes the total squared error. Now run simple regression on the data (Y = response, X = explanatory) to see what the regression routine gives as the best values for the parameters.

If we do this in general, using calculus and algebra, we can find some interesting facts. The total error function will look like (remember, we have eliminated the A variable with the relationship above)

$$E(B) = \sum_{i=1}^n (y_i - (A + Bx_i))^2 = \sum_{i=1}^n (y_i - A - Bx_i)^2 = \sum_{i=1}^n (y_i - Bx_i - (\bar{y} - B\bar{x}))^2$$

We can rearrange this last expression to be a little friendlier:

$$E(B) = \sum_{i=1}^n [(y_i - \bar{y}) - B(x_i - \bar{x})]^2 = \sum_{i=1}^n [(y_i - \bar{y})^2 - 2B(x_i - \bar{x})(y_i - \bar{y}) + B^2(x_i - \bar{x})^2]$$

This can be rearranged a little to get an expression that really looks like a second degree polynomial in B (with ugly coefficients - but they’re just numbers!)

$$E(B) = B^2 \sum_{i=1}^n (x_i - \bar{x})^2 - 2B \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n (y_i - \bar{y})^2$$

The derivative of this is just

$$E'(B) = 2B \sum (x_i - \bar{x})^2 - 2 \sum (x_i - \bar{x})(y_i - \bar{y})$$

Setting this right hand side of this last expression equal to zero and solving for the parameter B we see that the error is minimized when

$$B = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$

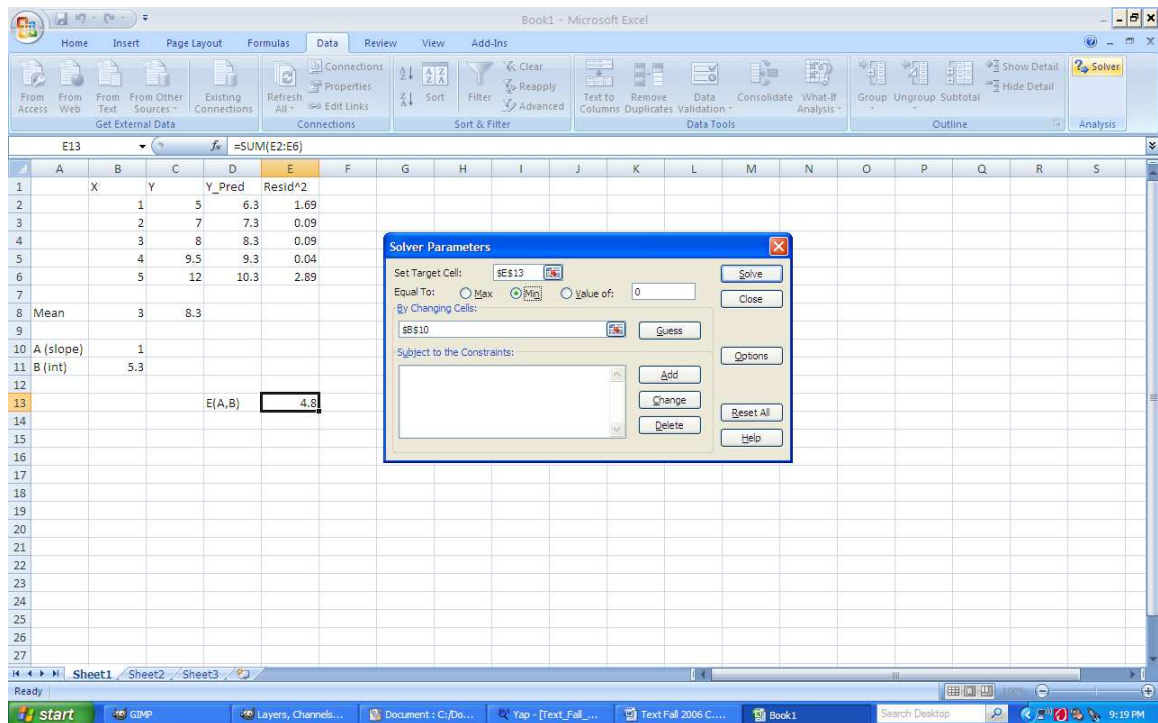


Figure 14.8: Screen shot for minimizing the total squared error.

14.3 Homework

Mechanics and Techniques Problems

14.1. The function $\bar{c}(q) = 0.1q + 3 + \frac{2}{q}$ represents the average cost for producing q of a product. (Assuming that $q > 0$.) Find the minimum average cost and the number of goods that should be produced in order to achieve this minimum.

14.2. The function $\bar{c}(q) = \frac{10,484.69}{q} - 2.250 + 0.000328q$ gives the average cost for producing q goods.

1. Find a formula for the total cost of producing q goods by multiplying the average cost function by the number of goods produced.
2. Find the minimum total cost and the number of goods that should be produced in order to achieve this minimum total cost.

14.3. Given the points $(1, 12)$, $(2, 7)$, $(3, 5)$ and $(4, 6)$, assume that a linear function fits these points. Assume that the linear function passes through the point (\bar{x}, \bar{y}) so that the y -intercept, A , is given by $A = \bar{y} - B\bar{x}$ where B is the slope of the least-squares regression line.

1. Write down the exact error function, $E(B)$, as a function of slope for the total squared error between the data points and the regression line.
2. Minimize your total squared error function to find the slope of the least-squares regression line. Show all steps and explain all work.

Application and Reasoning Problems

14.4. We are given the following information regarding a product:

$$\text{Demand function: } p = 400 - 2q$$

$$\text{Average Cost: } \bar{c} = 0.2q + 4 + \frac{400}{q}$$

1. The demand function gives the price people are willing to pay for the product, based on its availability (measured by q , the production). Use this to find the revenue function for this product.
2. Find the total cost function.

3. Find the profit function (profit = revenue - cost).
4. Use your profit function to determine the maximum profit and the number of goods to produce in order to achieve this maximum profit.
5. Based on your optimization of the profit, what is the price (look at the demand function!) at which the maximum profit occurs.
6. Suppose that the government imposes a tax of \$22 per unit on the product. What is the new maximum profit, number of goods needed to achieve maximum, and price?

14.5. Consider the profit graphs of each of the two companies shown in figure 14.9 from two different perspectives: the managers of the company (who want to keep their jobs) and the shareholders in the company (who want to make more money). For each graph, consider everything: the value of the function plotted on the graph, the rate of change of that function and the concavity (rate of change of the rate of change). Answer the following questions:

1. What might the managers say about this profit scenario in order to justify that they have been doing a good job leading the company and should keep their jobs?
2. What might the shareholders say to challenge the way the managers have run the company?

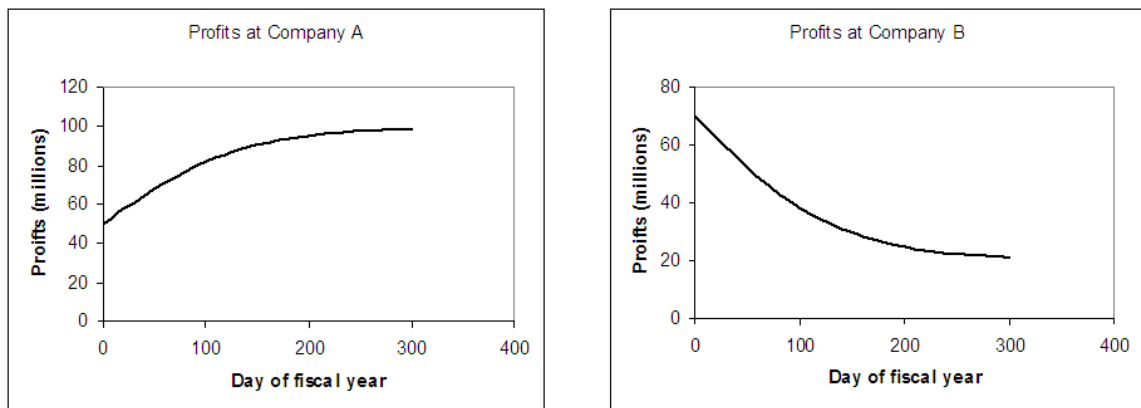


Figure 14.9: Profits over a year at two companies. Which is doing better?

14.6. Re-examine the situation in problem 5, only this time, imagine that the graphs given show the rate of change in the profits (millions of dollars per day) rather than the profits themselves.

14.7. Data file `C14 MacroSoft Profits` contains data on weekly profits over each of the past 52 weeks. The profits are in thousands of dollars. Also shown are the corresponding number of units of software sold each week. At yesterday's board meeting, the operations manager claimed that the data shows profits are increasing as we produce more units of software. This means that the company can produce as much software as they want and continue to make profits. The CEO never believes news this good. Analyze the data, build some models for the profits, and analyze this claim.

Deeper Exploration of Logs and Exponentials¹

Not all of the models that we can use to describe real world data are based on power functions or polynomials. In fact, we saw in earlier chapters that there are many situations where exponential or logarithmic models may be needed. We also developed a way of interpreting the coefficients of these models using parameter analysis. However, parameter analysis does not give us the power needed to locate maxima and minima for such models. Only calculus tools, specifically the derivative, can do this. In this chapter, you will work with the derivatives of exponential and logarithmic functions, and you will further apply these tools to analyze models of the business world. When you have finished this chapter, you will know how to deal with many of the basic functions found in the real world. The symbolic analysis portion of this chapter will show you how, using multiplication, division and composition of models, we can build many more types of models and analyze them using calculus.

- Section 15.1 shows how to find the derivative of logarithmic and exponential models.
- Section 15.2 demonstrates another use of logarithms and exponentials for compound interest and amortization.

<p><i>As a result of this chapter, students will learn</i></p>	<p><i>As a result of this chapter, students will be able to</i></p>
--	---

- | | |
|--|---|
| <ul style="list-style-type: none"> ✓ How to use the calculus tool of derivatives to analyze models involving logarithms ✓ How to use derivatives to analyze models involving exponentials ✓ How compound interest works, including continuously compounded interest | <ul style="list-style-type: none"> ✓ Take derivatives of exponential functions ✓ Take derivatives of logarithmic functions ✓ Compute compound interest |
|--|---|

¹©2017 Kris H. Green and W. Allen Emerson

15.1 Logarithms and their derivatives

As we have seen, there are many times when the model you develop will need to go beyond the power or polynomial models. For a multitude of reasons, the exponential and logarithmic models are the next most common models:

1. Exponentials are easy to interpret based on percent changes; thus, they can easily represent mathematically the process of accruing interest for loans or other accounting-related phenomena.
2. Logarithms are useful for dealing with some of the potential problems in modeling data, specifically the problem of non-constant variance.
3. Logarithms can be useful for simplifying many other models for analysis, since logarithms (remember the properties listed in section 12.2.1) can be used to convert many expressions involving multiplication and division into addition and subtraction problems.

These reasons alone are sufficient to justify learning how to properly use derivatives to analyze such functions. Before we get to technical, though, it's worth looking at the functions themselves and trying to figure out what we expect to happen. If we look at a graph of an exponential function, we notice immediately that the slope is always increasing. The slope is always positive, and the curve is always concave up. Thus, we expect the derivative to (a) always be positive and (b) increase as x increases. While these observations seem to tell us a lot, we have to remember that we are only looking at a small portion of the complete graph of the function, so it is possible that somewhere far from where we are looking this behavior will change. Once we have the derivative in hand, however, we can find out if this happens. (You'll have a chance to work with this in one of the problems at the end of the chapter.)

This is all in stark contrast to logarithmic functions. The graph of a logarithmic function shows more complex behavior. While it is true that the graphs seems to be always increasing, notice that the slope is decreasing as we move to the right. Thus, the logarithmic function seems to be concave down everywhere, even though it is increasing. Is it possible that somewhere far down the line the graph actually starts to decrease? We must also bear in mind that whatever we learn about one of the functions can be applied to the other, since logarithms and exponentials are inverses of each other.

The following section is devoted to learning about the derivatives of logarithmic functions. The development of this will mimic the path we took in chapter 14 to develop the derivative formulas for the power and polynomial models. Along the way we will encounter some other rules for taking derivatives: the chain rule, product rule and quotient rule. These will give us the ability to differentiate (take the derivative of) functions that are made of combinations of basic functions like logarithms and power functions. The next section will explore the exponential function and its applications to one of the most frequently used economics and business scenarios: compound interest.

15.1.1 Definitions and Formulas

Composition of Functions This is one way of making a new function from two old functions. Essentially, we take one function and “plug it into” the other function. For example, if we compose $f(x) = 2x^3$ and $g(x) = 4x - 5$ we get either $h(x) = (f \circ g)(x) = f(g(x)) = 2(4x - 5)^3$ or we get $k(x) = (g \circ f)(x) = 4(2x^3) - 5$ depending on the order of the composition. In general, the two orders are not the same.

Chain rule We’ll be using this rule a lot. The symbolic analysis section will explain it in more detail, but the basic idea is that if you have a function composed with another function and you need the derivative of the combined object, you use the chain rule to “chain together” derivatives of each function. For example, if we start with the functions $f(x)$ and $g(x)$ above and compose them into $h(x)$ the new function h is no longer a simple power function or polynomial (although we could multiply it out into a polynomial.) But since it is composed of these simpler functions, we can still take its derivative. In fact, the chain rule says that

$$\frac{d}{dx}f(g(x)) = \frac{df}{dg} \cdot \frac{dg}{dx}.$$

Thus $h'(x) = [df/dg][dg/dx] = [2 \cdot 3g(x)^2] \cdot [4] = 24(4x - 5)^2$. A derivation and proof of the chain rule are somewhat technical; for now, think of this as a way of chaining together the derivatives so the objects which look like (but aren’t really) fractions will cancel out. In the above illustration of the chain rule, the first “fraction” has the numerator we want (df) and the second “fraction” has the denominator we want (dx). Each of these “fractions” has a dg term that “cancels out” to give the derivative we want: df/dx .

Product rule The product rule allows us to take derivatives of functions that are products of simpler functions. It says that

$$\frac{d}{dx}[f(x) \cdot g(x)] = g(x) \cdot \frac{df}{dx} + f(x) \cdot \frac{dg}{dx}.$$

The proof of this rule will be given in the symbolic analysis section, and will make use of the derivative of a logarithm and the chain rule.

Quotient rule The product rule allows us to take derivatives of functions that are products of simpler functions. It says that

$$\frac{d}{dx} \left[\frac{f(x)}{g(x)} \right] = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}$$

The proof of this rule will be given in the symbolic analysis section, and will make use of the derivative of a logarithm and the chain rule.

15.1.2 Worked Examples

Example 15.1. Derivative formula for logarithmic models

In example 1 we developed a model for the cost of electricity as a function of the number of units of electricity produced. Marginal analysis can help us to make more specific sense of this model by helping us to interpret how much each unit of electricity affects the total cost of producing the electricity. The model had the form $f(x) = A + B \cdot \ln(x)$:

$$\text{Cost} = -63,993.30 + 16,653.55 \cdot \text{Log}(\text{Units})$$

Suppose that we are currently producing 500 units of electricity. How much would it cost to produce one more unit of electricity? We can put this into a spreadsheet to compute it fairly easily. The results are shown below, and were obtained by setting up a formula for the difference quotient of the function, with a variable for h so that we can let h get very small. This lets us see what the instantaneous rate of change of the cost function is (this data is reproduced in the first worksheet of C15 LogDerivative).

A	-63,993.30				
B	16,653.55				
X	500				
H	X+H	F(X)	F(X+H)	DF = F(X+H)-F(X)	DF/H
10	510	39501.99	39831.77	329.7840438	32.9784
1	501	39501.99	39535.26	33.27383724	33.27384
0.1	500.1	39501.99	39505.32	3.330376973	33.30377
0.01	500.01	39501.99	39502.32	0.333067669	33.30677
0.001	500.001	39501.99	39502.02	0.033307067	33.30707

From this, it seems that when current production is at 500 units, each additional unit of electricity will cost approximately \$33.31. In contrast, if are currently producing 1,000 units of electricity, the marginal cost is about \$16.65/unit.

A	-63,993.30				
B	16,653.55				
X	1000				
H	X+H	F(X)	F(X+H)	DF = F(X+H)-F(X)	DF/H
10	1010	51045.35	51211.06	165.7083324	16.57083
1	1001	51045.35	51061.99	16.64522877	16.64523
0.1	1000.1	51045.35	51047.01	1.665271738	16.65272
0.01	1000.01	51045.35	51045.51	0.166534667	16.65347
0.001	1000.001	51045.35	51045.36	0.016653542	16.65354

Now, what can this tell us about the derivative formula of a logarithmic function? Quite a lot, actually. Notice that as the production level increased (from 500 to 1000 units) the derivative (approximated by the column labeled “DF/H”) decreased. Thus, we expect the derivative of a logarithmic function to be a decreasing function. This makes perfect sense when looking at the graph of a logarithmic function, since the graph “flattens out” the farther you move along the x -axis. We can repeat the same method of analysis from chapter ?? to build a table of values for $[\ln(x)]'$. If we plot these values, we get a graph much like the one below (see the second worksheet of Ch15 LogDerivative).

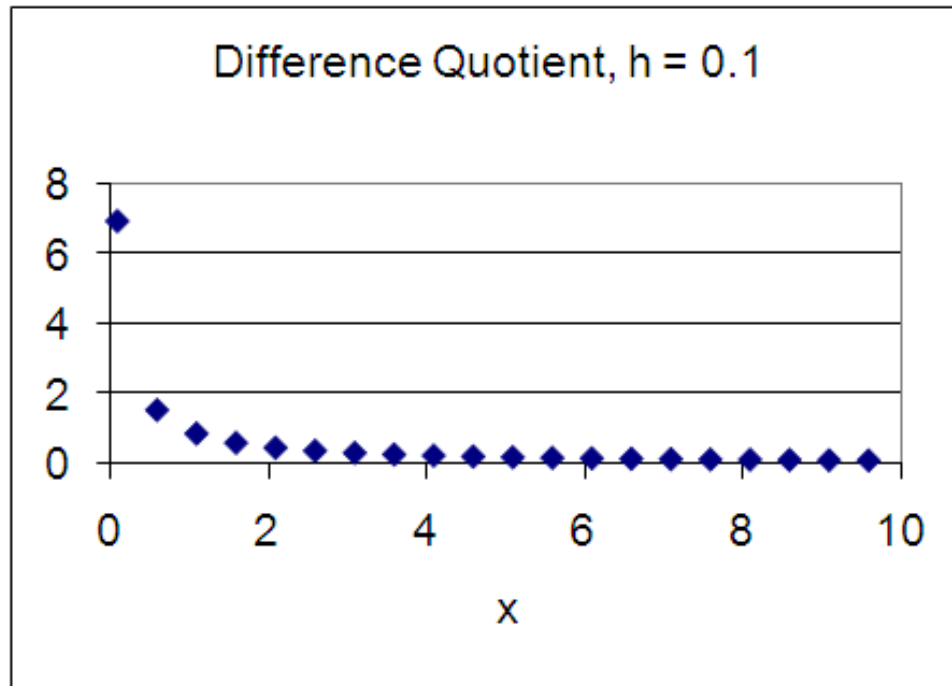


Figure 15.1: Difference quotient of a basic logarithmic function.

Notice that the difference quotient appears to be very similar to the inverse function, $f(x) = x^{-1}$. This is a power function, so we can superimpose a trend line on this data using a power function. If we do, we find remarkable agreement, even with $h = 0.1$. Reducing h will, however, quickly achieve a nearly perfect fit for the inverse function to the difference quotient. While we have not truly proven this, we can assert with some confidence that

$$\frac{d}{dx} \ln(x) = \frac{1}{x}.$$

Now, we can use this along with what we already know about derivatives to determine the derivative of a more complete logarithmic model:

$$\frac{d}{dx} (A + B \ln(x)) = \frac{d}{dx} (A) + \frac{d}{dx} (B \ln(x)) = 0 + B \frac{d}{dx} \ln(x) = \frac{B}{x}.$$

Thus, we expect that the derivative of the logarithmic function above (with $A = -63,993.30$ and $B = 16,653.55$) to be equal to $B/x = 16,653.55/x$. So when the production level is 500,

the derivative should be $16,653.55/500 = 33.3071$, which is extremely close to the number we estimated using the difference quotient above. If the production level is 1000, we expect the derivative to be $16,653.55/1000 = 16.65355$, which is again very close to the estimates determined earlier.

Example 15.2. Derivative of a logarithmic function

Find the derivative of the function $f(x) = 3 - 2\ln(5x)$ with respect to the variable SxS .

$$\begin{aligned}
 f'(x) &= \frac{d}{dx}(3 - 2\ln(x)) \\
 &= \frac{d}{dx}(2) + \frac{d}{dx}(-2\ln(x)) && \text{Using the sum rule for derivatives} \\
 &= 0 - 2\frac{d}{dx}\ln(5x) && \text{Derivative of a constant is zero AND deriva-} \\
 & && \text{tive of a constant times a function} \\
 &= -2 \cdot \frac{1}{5x} \cdot \frac{d}{dx}(5x) && \text{Using the chain rule} \\
 &= -2 \cdot \frac{1}{5x} \cdot 5 && \text{Computing the derivative of the linear func-} \\
 & && \text{tion} \\
 &= -\frac{2}{x} && \text{Simplifying the derivative}
 \end{aligned}$$

Example 15.3. A more complex derivative

Now for the hardest example yet. Find the derivative of the compound function below:

$$h(x) = \frac{(3 + 2x + x^2)(5 + x)^4}{2 + 3x + 7x^2}.$$

There are several different paths we could take through this problem. We'll do it here by using the logarithmic derivative (one could use the chain, product and quotient rules all at once also). To do this, we take the natural logarithm of both sides and simplify the resulting mess that appears on the right hand side.

$$\begin{aligned}
 \ln(h(x)) &= \ln\left[\frac{(3 + 2x + x^2)(5 + x)^4}{2 + 3x + 7x^2}\right] \\
 &= \ln(3 + 2x + x^2) + \ln(5 + x)^4 - \ln(2 + 3x + 7x^2) \\
 &= \ln(3 + 2x + x^2) + 4\ln(5 + x) - \ln(2 + 3x + 7x^2)
 \end{aligned}$$

Taking the derivative is now a matter of using the chain rule, piece by piece. For example, we know that the derivative of the left hand side with respect to the variable x is just $h'(x)/h(x)$, where $h'(x)$ is the derivative we really want. Now we need to take the derivative of the right hand side; we'll do it in three parts, one for each term on the right hand side.

$$\begin{aligned}
 \frac{d}{dx}\ln(3 + 2x + x^2) &= \frac{1}{3 + 2x + x^2} \cdot \frac{d}{dx}(3 + 2x + x^2) = \frac{2 + 2x}{3 + 2x + x^2} \\
 \frac{d}{dx}[4\ln(5 + x)] &= 4 \cdot \frac{d}{dx}\ln(5 + x) = 4 \cdot \frac{1}{5 + x} \cdot \frac{d}{dx}(5 + x) = \frac{4}{5 + x} \\
 \frac{d}{dx}\ln(2 + 3x + 7x^2) &= \frac{1}{2 + 3x + 7x^2} \cdot \frac{d}{dx}(2 + 3x + 7x^2) = \frac{3 + 14x}{2 + 3x + 7x^2}
 \end{aligned}$$

Now we can put this all together to get

$$\frac{1}{h(x)} \frac{dh}{dx} = \frac{2+2x}{3+2x+x^2} + \frac{4}{5+x} - \frac{3+14x}{2+3x+7x^2}$$

Cross multiplying by $h(x)$ then gives us the derivative of h with respect to x

$$\frac{dh}{dx} = \left[\frac{2+2x}{3+2x+x^2} + \frac{4}{5+x} - \frac{3+14x}{2+3x+7x^2} \right] \cdot \frac{(3+2x+x^2)(5+x)^4}{2+3x+7x^2}$$

After a great deal of work, this can simplify to

$$\frac{dh}{dx} = \frac{(2+2x)(5+x)^4}{2+3x+7x^2} + \frac{4(3+2x+x^2)(5+x)^3}{2+3x+7x^2} + \frac{(3+14x)(3+2x+x^2)(5+x)^4}{(2+3x+7x^2)^2}$$

If we get a common denominator, we can further simplify this, but it doesn't really help.

15.1.3 Exploration 15A: Logs and distributions of data

Part 1. Open the data file **C15 WaitTimes**. This first worksheet (labeled “part 1”) contains a list of 400 service times at Beef ’n Buns. Generate a histogram of the data to match the histogram below. Notice that the distribution of service times is significantly right-skewed.

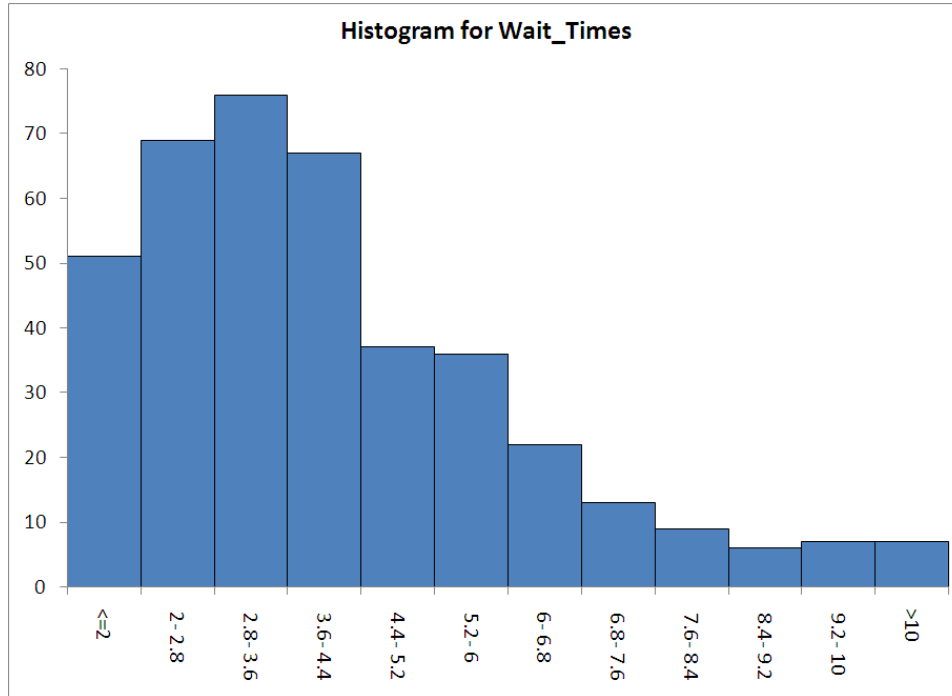


Figure 15.2: Histogram of wait times at Beef n’ Buns, showing the right-skewed distribution.

One of the assumptions about linear regression involves the distribution of the data. If we were to try and create a regression model to predict service times, we would find this model to have significant error, due to the data’s skewness. There is, however, an easy way to normalize the data in order to produce a better model. Create a column of wait times that has been transformed by taking the natural logarithm. Create a histogram of these logged wait times. What do you see? Under what circumstances might this be a useful tool for model building?

Part 2. The second worksheet in the file illustrates another property of logarithms. In fact, it is this property that makes the process in part 1 work. This sheet shows a graph of the natural logarithm, along with vertical and horizontal lines passing through the data points. From looking at the graph, which has points that are equally spaced in the x direction, can you explain why logarithmic functions are sometimes described as “compressing data”? Your task is to first change the x coordinates of the points (in column B) so that the change in y between successive points is the same - exactly the same. Then, use the other information in the data table and what you know about the property of logarithms to explain why this particular spacing of x values solves the problem. What other x values would work?

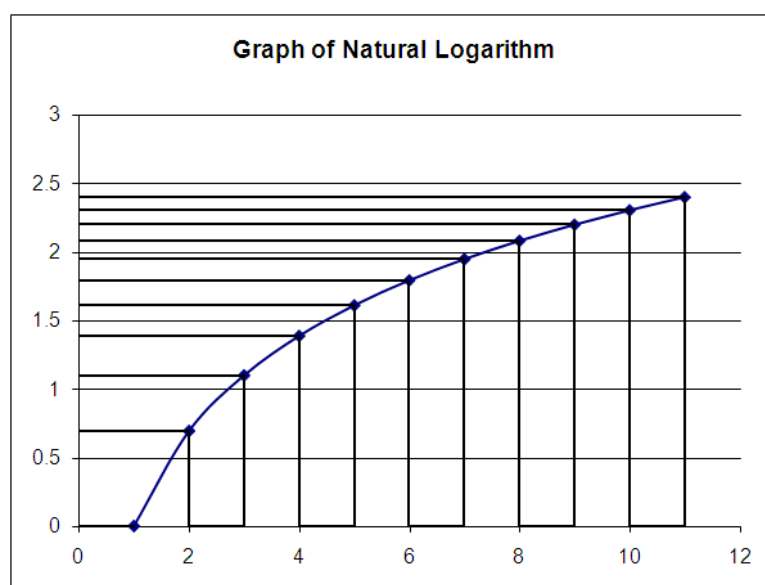


Figure 15.3: Graph of the natural logarithmic function, showing their basic properties.

15.2 Compound interest and derivatives of exponentials

Compound interest is one of the foundations of modern finance. The basic idea is that your investment will earn interest on the amount invested (the principal) as well as the interest itself. There are two primary versions of compound interest that we will explore in this section. The first is the easiest to make sense of, the case where there are a fixed number of times each year when the interest is computed and then added to the account. The other version is harder to understand intuitively because it involves interest being computed an infinite number of times. While it may seem that this would give you an infinite amount of money, since the interest rate for each period is infinitesimally small (it is the annual percentage rate divided by the number of compounding periods, so it is extremely small) the total amount reaches a fixed limit related to the number e .

Once we understand the basics of compound interest it can be applied to many other economic and financial concepts, such as present value and future value of an investment. The present value of an investment is the amount you would need to invest today in order to achieve a fixed level at the end of the investment period. This situation is most easily understood through the modern day phenomenon of the lottery. Most lotteries offer the winner two choices of payment: a lump sum now or small payments made over a longer period of time, say 20 years. If the winner “won” \$1 million, she would, for example, have to choose between monthly payments of \$50,000 each year for twenty years (a total of \$1 million) or a lump sum payment of \$548,811.64 right now. Ignoring all taxes, of course, which substantially change the problem under consideration, the reason the lump sum payment is so much less than the actual winnings is that you are getting it now. If you were to invest it at 3% for 20 years, you would have about \$1 million at the end, the same amount as the lottery winnings. Since the lottery company would have access to the money in the 20-year payment version, they would be earning interest on the \$1 million over that entire 20 years. But if they have to pay you all right now, they lose that interest. Thus, the present value of the \$1 million lottery winning is about \$550,000, assuming a 3% interest rate annually. We will further explore the idea of present value in the problems for this section.

15.2.1 Definitions and Formulas

Principal The amount of money initially invested or borrowed; it is the basis for computing the interest for the investment or loan.

Simple interest Simple interest is a way of computing the value of an investment based on giving interest one time only: at the very end of the investment period.

Compound interest Compound interest involves breaking the lifetime of the loan or investment into many periods. During one period, simple interest is used to compute the value of the loan or investment. During the next period, the interest for that is based not on the original principal, but on the current value of the loan including all interest from previous periods. Thus, with compound interest, you earn interest on your interest.

Continuously compounded interest This is a form of compound interest that uses, essentially, an infinite number of infinitesimally short investment periods for computing the interest. When this is done, we find that the exponential function with base e is a natural way to express the investment value.

15.2.2 Worked Examples

Example 15.4. Compound interest formulas

Suppose we were to invest an amount of principal, P , in an account that earns an interest rate r each year (this is the APR, or Annual Percentage Rate). This means that at the end of the first year, you will earn rP additional money. Thus, after one year, your account, A , has the value

$$A(1) = P + rP = P(1 + r).$$

If you were to leave the money in the account for a second year, you would earn interest not only on the principal, but also on the interest you earned the first year:

$$A(2) = P(1 + r) + P(1 + r)r = P(1 + r)(1 + r) = P(1 + r)^2.$$

What if you let the money earn interest for a third year? You would have a total of

$$A(3) = P(1 + r)^2 + P(1 + r)^2r = P(1 + r)^2(1 + r) = P(1 + r)^3.$$

With a little work, we can show that, in general, after n years at an APR of r your principal P will earn a total of

$$A(n) = P(1 + r)^n.$$

Now, suppose that our interest is not computed annually, but is computed every month, based on the APR. This means that the actual monthly interest rate is $r/12$ and that in a single year we have 12 compounding periods. Similar logic to the previous case will tell us that after t years of compounding the interest monthly at this rate we will have

$$A(t) = P \left(1 + \frac{r}{12} \right)^{12t}$$

dollars in the account. Similarly, if we let the money be compounded n times each year, we will have an interest rate of r/n each period and a total of nt compounding periods after t years. This gives us an amount of

$$A(t) = P \left(1 + \frac{r}{n} \right)^{nt}.$$

This is, obviously, an exponential function, but with a base of $(1 + r/n)$ rather than the natural base of e . However, they are related. Consider what happens if we invest \$1 at 100% APR for one year under different compounding periods, as shown in the table below.

Schedule	Number of Periods	Total Amount
Annual	1	2
Monthly	12	2.61303529
Weekly	52	2.692596954
Daily	365	2.714567482
Hourly	8760	2.718126692
Each minute	525600	2.718279243
Each second	31536000	2.718281781
Every tenth of a second	315360000	2.71828187
Every hundredth of a second	3153600000	2.718281661

Notice that the amount of money does continue to grow, but not at the same rate. In fact, it seems that the amount of money we are earning is approaching a fixed amount. Mathematically, it has been proven that this is the case and that the number this approaches is the number e :

The number e is the amount of money earned in an account after investing \$1 for one year at 100% interest, compounded continuously.

Mathematicians write this fact using the limit notation:

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e.$$

We can now use this fact to generate a formula for continuously compounded interest. First, we introduce a new variable m so that $n = r \cdot m$. Then we have an equivalent expression for the interest given by

$$\lim_{n \rightarrow \infty} \left(1 + \frac{r}{n}\right)^n = \lim_{m \rightarrow \infty} \left(1 + \frac{1}{m}\right)^{mrt} = \left[\lim_{m \rightarrow \infty} \left(1 + \frac{1}{m}\right)^m \right]^{rt} = e^{rt}.$$

Thus, our formula for the amount in an account with n compounding periods changes to the following formula if we compound it continuously:

$$A(t) = Pe^{rt}.$$

Example 15.5. Derivatives of exponential functions

Now that we know about the derivatives of logarithmic functions, we can easily use the idea of a logarithmic derivative to determine the derivative of an exponential function. One of the most common exponential functions to occur in the business world relates to the future value of an investment. To get to this, though, we'll need to develop the idea of compound interest.

So, although it took us a little while to get there, and we skipped a few steps, we see that the exponential function is closely tied to the idea of compound interest. We can now

ask the following. Suppose you have invested a fixed amount of money P at a fixed rate of interest r . How quickly (in time) is your money growing in value?

The question “how quickly” immediately reminds us of the idea of rates of change, so we know we are really talking about the derivative of the amount of money in the account. So, what is the derivative of the amount? We’ll use our knowledge of logarithmic derivatives to help. We really want to know the derivative of $A(t)$, but we don’t know the derivative of an exponential. However, the exponential function and the logarithmic function are inverses of each other, so the formula for the amount can be rewritten as

$$\ln(A(t)) = \ln(Pe^{rt}) = \ln(P) + \ln(e^{rt}) = \ln(P) + rt \ln(e) = \ln(P) + rt$$

where we have used the rules for manipulating logarithms and the fact that $\ln(e) = 1$. Now, we can take the derivative of each side of this equation, using the chain rule:

$$\frac{d}{dt}(\text{left hand side}) = \frac{d}{dt} \ln(A(t)) = \frac{1}{A(t)} \frac{dA}{dt}.$$

Now, the derivative of the right hand side is easy, since it’s really a linear function (note that $\ln(P)$ is a constant; it doesn’t depend on the variable t with respect to which we are taking the derivative):

$$\frac{d}{dt}(\text{right hand side}) = \frac{d}{dt} (\ln(P) + rt) = \frac{d}{dt} \ln(P) + \frac{d}{dt} (rt) = 0 + r = r.$$

We can now put all this together, since we have done the same thing to both sides of the equation (namely, take the derivative with respect to t), so they are still equal to each other.

$$\frac{1}{A(t)} \frac{dA}{dt} = r \Rightarrow \frac{dA}{dt} = rA(t) = rPe^{rt}.$$

So, the true rate of increase of your account value is an amount of $r \cdot \exp(rt)$ dollars per year. If you let it sit for $t = 10$ years at a rate of 2.5% your money will be increasing at a rate of $A'(10) = 0.025 \cdot P \cdot \exp(0.025 \cdot 10) = 0.025 \cdot P \cdot \exp(0.25) = 0.032P$ dollars per year. If you had invested \$1000 initially, this would come to a growth rate of about \$32/year.

Example 15.6. Derivative of an exponential function

Find the relative rate of change of the function $g(r, t, P) = Pe^{rt}$ with respect to the variable r . The relative rate of change is just the rate of change divided by the function itself, so we have the relative rate of change as $(1/g) \cdot$ (derivative of g with respect to r).

$$\frac{1}{g} \frac{\partial g}{\partial r} = \frac{1}{g} \frac{\partial}{\partial r} (Pe^{rt})$$

$$= \frac{1}{g} \cdot P \cdot \frac{\partial}{\partial r} (rt)$$

$$= \frac{1}{g} \cdot P \cdot e^{rt} \cdot r$$

$$= \frac{1}{g} \cdot r \cdot g$$

$$= r$$

Definition of relative rate of change, using partial derivative notation since there are several variables in the function

Derivative of a constant times a function

Derivative of an exponential AND chain rule

Derivative of a linear function

Simplification

This means that the relative rate of change of the formula for continuously compounded interest is just equal to the interest rate itself. To understand what this means, think about the units of the rate of change with respect to r : units of dollars divided by units of interest. When we divide this by the amount (dollars) we get the relative rate of change, which is measured in $1/(\text{interest rate})$. This is a relative amount, so it is like a percentage. Thus, each actual 1% increase in the interest rate (from 1% to 2% or from 5.25% to 6.25%) will increase the value of our account for a fixed amount of principal invested for a fixed period of time by r %.

Example 15.7. Application of Marginal Analysis to Business Decisions

The analysis team at Koduck has determined the following information about your current production level:

$$\begin{aligned}\text{Marginal cost (MC)} &= \$2.25/\text{unit} \\ \text{Marginal Revenue (MR)} &= -\$1.15/\text{unit}\end{aligned}$$

What does this mean for Koduck?

For starters, we note that a negative value for marginal revenue means that if you increase production by 1 unit, your overall revenue (price * number sold) will actually drop. (This could be because you have already flooded the market; after all, how many pictures of water fowl can you sell in a given city?) The fact that the marginal cost is positive means that it will cost you more to make one more unit of product. Thus, it seems that increasing current production levels would not be wise: The total cost would rise and the revenue would drop, leading to lower profits. No one wants that. In fact, we should probably decrease production in order to increase profits! If we decrease production by 5 units, say, then we can expect the revenue to increase:

$$\text{Change in Revenue} = \text{MR} * \text{change in production} = (-\$1.15/\text{unit}) * (-5 \text{ units}) = \$5.75.$$

At the same time, this would result in a decrease in cost:

$$\text{Change in Cost} = \text{MC} * \text{change in production} = (\$2.25/\text{unit}) * (-5 \text{ units}) = -\$11.25.$$

This results in a total change in the profit of \$17! It is a fact (which we will explore later) that the maximum possible profit (= revenue - cost) must happen when the marginal cost and the marginal revenue are equal. Since we can increase profits by lowering production, we must be producing more units than necessary to achieve the maximum profit.

15.2.3 Exploration 15B: Loan Amortization

In practice, the types of interest discussed in this chapter (simple, compound, and continuously compounded) are only parts of larger schemes for determining interest. One common application of simple interest is in loan amortization. The idea is that you take out a loan for a specified amount of principal, at a particular APR, for a set period of time. This time period is broken into smaller time periods (for example, a fifteen year loan for a house might be broken into monthly payment periods) and during each period you pay back some principal and some interest. However, while the total amount of each payment is generally held constant, the amount of that payment devoted to interest and principal repayment are not. In this exploration, you will construct a spreadsheet to explore the way a loan is repaid.

Suppose we take out a \$130,000 loan for a property. If the loan is at 6% interest (APR) and we pay it back monthly over a fifteen year period (180 payments) how much will we need to pay per month? Start by entering the basic information on the loan, as shown below in cells A1:B4. In cell B4, put your guess for the amount you would expect to pay. Try to be reasonable, keeping in mind that none of the interest schemes above will actually give you the amount, since the amount of interest to be paid at any one time is determined by the remaining principal on the loan. In cell E1, enter a formula to calculate the monthly interest rate (it is the APR divided by the Number of Periods in a year). Now, set up the loan amortization table headers as shown. Under “Period” enter the numbers 1, 2, 3, etc. up to 180 at 12 monthly periods per year, this will carry our loan through 15 years.

	A	B	C	D	E	F
1	Principal	\$130,000		Periodic Int	0.005	
2	APR	6.00%				
3	Num Periods	12				
4	Est Payments					
5						
6						
7						
8						
9						
10	Period	Interest	Principal	Cumulative Interest	Cumulative Principal	Remaining Principal

Figure 15.4: Setup for computing a loan amortization

Now, the interest for a particular period is easy to compute: it's just simple interest on the remaining principal balance. So, for the first period, all we need to do is multiply the periodic interest rate by the original loan amount. Once we have this, the amount of principal in the first payment is the total monthly payment minus the interest that period. The cumulative interest is just a place to track the running total on the interest we have paid and the cumulative principal tracks the total we have paid on the original loan amount; the remaining principal is the original loan minus the cumulative principal. Your formulas for the first period will probably be slightly different than the formulas for the other periods, but once you have the formulas entered, you can copy them down the table. Since the goal is to pay off the loan in 15 years (180 monthly payments) try changing the “Est Pay” amount

until you find a monthly payment that leads to a balance of zero remaining principal in period 180 (cell F190).

Once you have played with this a little, you can use “Goal Seek” (in Excel) or `uniroot` in R to compute the actual monthly payment required to pay the loan off by a certain period of time. Try constructing a table listing different monthly payments based on changing one of the loan parameters (like the interest rate). Pay particular attention to the cumulative interest paid on the loan.

N.B.: There is a way to compute, just from the loan information, the monthly payment required. This formula, however, requires a lot of computational work, and we can get the same information by playing with it in our spreadsheet. Details of the formula will be discussed in the Symbolic Manipulation supplement for those interested. There are also automatic formulas in most software for computing loan amortizations. If you are interested, look up the functions PMT, IPMT and PPMT.

15.3 Homework

Mechanics and Techniques Problems

15.1. For each of the following functions, compute the first derivative of the function with respect to the independent variable.

1. $f(x) = 3 \ln(x) + 5$
2. $h(t) = -2e^{3t}$
3. $g(s) = 5s + 3s \ln(s^2 - 4)$
4. $p(y) = 5ye^{-y^2}$
5. The logistic function $f(x) = \frac{A}{1+e^{-Bx}}$, where A and B are constants.

15.2. Find the local maxima and minima of the function given in 1d) above. Use this information to help sketch a picture of what the function looks like when plotted as $p(y)$ versus y .

15.3. The present value of an investment is the amount of money you would need to invest at a particular interest rate r for a specified period of time t in order for the investment to rise to a total value of V .

1. Assuming that there are n compounding periods per year, determine a formula for the present value of an investment.
2. Assuming that the interest is compounded continuously, determine a formula for the present value of an investment.
3. Using your formulas in a) and b) fill in the table showing the present value of a 10-year investment that has a value of \$1 million. Your table should compute this for the following range of interest rates: 1%, 2.5%, 5% and should show the results for annual compounding, monthly compounding, daily compounding and continuous compounding.

Interest Rate	Annual compounding	Monthly compounding	Daily compounding	Continuous compounding
1%				
2.5%				
5%				

Application and Reasoning Problems

15.4. Suppose that you are a manufacturer of widgets. At your current level of production, you have determined that each one unit increase in the production level will decrease the revenue by \$0.28. Each unit of increase in the production level leads to a drop in costs of \$0.34. Each day, your plant is improving efficiency, so each day the production level is expected to increase by 32 units. At what rate is the profit changing? Would you continue to increase the production? Why?

15.5. Prove that the exponential function of the form $y = Ae^{Bx}$ is an always increasing function of x (assuming that B is positive and A is positive). In other words, show that this function never reaches a maximum and then starts to decrease. Such functions are referred to as monotonically increasing.

15.6. Prove that the logarithmic function $y = A + B \ln(x)$ is a monotonically increasing function.

Optimization in Several Variables with Constraints¹

In a previous chapter, you explored the idea of slope (rate of change, also known as the derivative) and applied it to locating maxima and minima of a function of one variable (the process was referred to as optimization). However, we know that most functions that model real world data are composed of several variables, so we need slightly different techniques for this. If you recall the one-variable case, we only needed to set that derivative to zero to find the local maxima and minima. When there are n independent variables, there are n different partial derivatives. We can find the location of the maxima and minima by find the points at which all n of these derivatives are zero at the same time (simultaneously). This involves a great deal of algebra, and is not always possible to do without resorting to numerical methods that only find approximate locations.

To make matters worse, we also find that rarely are we optimizing a function by itself. Consider, for example, revenue for selling a certain number of products. The more you sell, the more you earn, so there is no maximum revenue; we can make as many as we want and still earn more revenue. But in the real world, we have to account for the cost of the objects we are selling, which includes raw materials, labor and equipment to produce them, marketing, distribution, and other costs. These extra conditions, known as constraints, make finding an optimum solution much more difficult. In this chapter, we will focus on defining such constraints and phrasing them mathematically. We will then see how to set up a spreadsheet to solve the optimization problem under these constraints.

- Section 16.1 shows you how to use the tools we have already built up to identify and model the constraints that limit your options in trying to find an optimal solution.
- In section 16.2, you will learn how to build the constraints into Excel and to use SOLVER to find the optimal solution.

¹©2017 Kris H. Green and W. Allen Emerson

As a result of this chapter, students will learn

- ✓ What constraint functions typically look like
- ✓ About sensitivity analysis

As a result of this chapter, students will be able to

- ✓ Formulate constraints for optimizing a function
- ✓ Formulate a constrained optimization problem for the “Solver” package in Excel or the `lpSolve` in R

16.1 Constraints on Optimization

So far, we have analyzed data by building models of the data and then interpreted those models. We have worked with models as equations that take one or more variables as input and have even worked with nonlinear functions. But analyzing the data and building the model is only part of the process. It is important that our model be useful for answering questions about the underlying situation and that we be able to use our model to make decisions. One of the most common uses of a model is in optimization, where we seek to make some quantity (such as profit or cost) either as large as possible (for profit) or as small as possible (for cost). In an earlier chapter, we did this with functions of a single variable, making use of a concept from calculus: the derivative. We found that when the derivative of a function is zero, the function is at a critical point, and that critical points are the only candidates for being optimum values of the function.

But this process ignores two things. The first is that most functions or models have several independent variables. Consider, for example, the commuter rail system examples we have used before. In that case, we built a model with a total of four variables. Our one-variable optimization process won't work here. The second thing we have ignored is that we are seldom free to choose just any values of the independent variables in order to achieve our optimum results. We are often constrained by resources. These resource constraints could involve time, money, personnel or just about anything that could limit our ability to reach certain values or combinations of values for the independent variables.

To correctly deal with the first problem, multiple-variable functions, we need to use partial derivatives (one for each independent variable) and solve several equations simultaneously. The idea is similar to the one variable case, but we now need all of the partial derivatives to be zero at the same exact point (set of values of the independent variables). We will not be looking into this here, because most of the common multivariable functions, linear and multiplicative, do not have critical points, and so we find no optimum solutions. Instead, we'll focus on the second aspect of optimization, applying constraints.

To begin, we must learn how to formulate the constraints. Typically, these will take the form of inequalities, rather than equations. After all, if the most you have to spend on production is \$100,000 and you can achieve a slightly higher profit by using on \$95,000, why not do it? So, rather than forcing our constraints to be equations, where quantities are equal to each other, we will use inequalities, where some quantity is either less than, greater than, less than or equal to, or greater than or equal to, some other quantity. We'll also see that most optimization problems involve multiple constraint conditions. For example, one constraint may involve time, one might involve cost of raw materials, one might involve equipment, and one might involve distribution.

16.1.1 Definitions and Formulas

Constraint Anything (such as time, money or budget, personnel or other resources) that limits your options regarding possible values of the variables in your model

Inequality An expression similar to an equation, except the quantities are related by either be less than ($<$), greater than ($>$), less than or equal to or greater than or equal to

(\leq or \geq) each other, rather than requiring them to be exactly equal. For example, the inequality

$$10 * (\text{Labor Hours}) + 2 * (\text{Items made}) \leq \$1000$$

Could be used to express the situation where we pay \$10/hour for labor and each item we make uses \$2 of raw goods, electricity and equipment time. If we must keep daily costs below \$1000, then the total cost spent of labor ($= \$10/\text{hour} * \text{number of hours used}$) plus the total cost in materials ($= \$2/\text{item} * \text{number of items made}$) must be less than or equal to the budget, in this case \$1000.

Objective function This is the quantity (or quantities) that you are trying to optimize. It is sometimes referred to as a target or target cell in spreadsheets.

Optimization variables These are the variables you can change (sometimes called the changing variables) in order to achieve your optimum solution.

Maximize In some optimization problems, you seek to make the objective function as large as possible. Such problems are maximization problems.

Minimize In some optimization problems, you seek to make the objective function as small as possible. Such problems are minimization problems.

Explicit constraint Explicit constraints describe those items that clearly given to you as goals during your optimization process. For example, limitations on resources (materials, labor) and limitations on demand are often stated explicitly.

Implicit constraint Implicit constraints refer to those quantities that you must recognize are also constraints on your optimization process. For example, in optimizing company profits by producing different quantities of different goods, the number of units of each good to produce might need to be an integer (unless it's measured in pounds). The quantity produced must also be non-negative. You must learn to recognize these types of constraints as well.

16.1.2 Worked Examples

Example 16.1. Understanding the optimization problem

We'll start with a typical sort of optimization problem. Suppose you want to maximize the profit from selling a variety of products, each of which requires different amounts of different materials that have different costs. Each of these products has a different demand function and sells for a different amount of money. As is common, you are required to remain under budget for both materials and labor.

Suppose our factory makes three products: kitchen tables, kitchen chairs and juice carts. We would like to maximize our profits from selling these three items, given the following information about the price, cost and production time for making each item (each value in the table is given per item produced):

Item	Assembly time	Finishing time	Materials cost	Selling price
Chairs	1 hr	1 hr	\$5	\$18
Tables	1 hr.	4 hr	\$15	\$54
Carts	2 hr	2 hr	\$10	\$36

In addition, at our factory we have the following weekly constraints:

	Maximum hours available	Cost per hour
Assembly	250 hr	\$4
Finishing	350 hr	\$7

We now need to take this raw information and frame it into a solvable optimization problem.

First, what is it we wish to optimize? That would be the profit from selling these items. Thus, the objective function is profit. In this case, the profit is given by the following:

Objective Function (Maximize)

$$\begin{aligned}
 \text{Profit} &= \text{Revenue from selling items} - \text{Cost for producing items} \\
 &= (\text{Revenue from chairs} + \text{Revenue from tables} + \text{Revenue from carts}) \\
 &\quad - (\text{Cost of making chairs} + \text{Cost of making tables} + \text{Cost of making carts}) \\
 &= \text{Profit from chairs} + \text{Profit from tables} + \text{Profit from carts}
 \end{aligned}$$

Note that one of our assumptions will be that we sell all of the chairs, tables and carts we make, and that all of these items sell for the listed price.

What are the input variables? In this problem, the only things we can change are the number of each item produced. Thus, we have three input variables: the number of chairs produced, the number of tables produced, and the number of juice carts produced. Once we know these numbers, we can compute the total cost for production, the total time for production and the profits we will earn.

What constraints do face? From the given information, it seems that we must keep the total amount of time to produce our chairs, tables, and carts within the amounts listed in the second table. Thus, we have two explicit constraints: the total time for assembly must be less than 250 hrs, and the total time for finishing must be less than 350 hrs.

Constraint #1 (Assembly time \leq 250 hr)

Assembly time = chair assembly time + table assembly time + cart assembly time \leq 250

Constraint #2 (Finishing time \leq 350 hr)

Finishing time = time to finish chairs + time to finish tables + time to finish carts \leq 350

At the same time, we have several implicit constraints. These are not stated directly, but are relevant to making sure we have a problem with a sensible solution. Thus, we need to be sure that we produce either zero or a positive number of each product. This gives us the following constraints:

Constraint #3 (Number of chairs produced ≥ 0)

Constraint #4 (Number of tables produced ≥ 0)

Constraint #5 (Number of carts produced ≥ 0)

We have three more implicit constraints: the number of each product that we make and sell must be an integer. After all, we cannot produce and sell one-third of a cart to make one-third of the revenue.

Constraint #6 (Number of chairs produced must be an integer)

Constraint #7 (Number of tables produced must be an integer)

Constraint #8 (Number of carts produced must be an integer)

Thus, even this relatively simple problem leads to eight constraints on the three input variables to maximize a single objective function.

Example 16.2. Mathematizing the optimization problem

Now that we have examined the problem from example 1 and determined the constraints, we can convert everything into mathematical notation.

Assign names to the variables

The first thing you must do is assign names to the variables. In this case, we will choose the obvious names:

C = number of chairs to be produced and sold

T = number of tables to be produced and sold

J = number of juice carts to be produced and sold

Converting the objective function

To write the objective function as a mathematical expression we need to understand what goes into it. The final form of the profit is probably easiest to work with:

$$\text{Profit} = \text{Profit from chairs} + \text{Profit from tables} + \text{Profit from juice carts}$$

To determine the profit from selling chairs, we look at four things: the revenue from selling C chairs, the materials cost for making C chairs, the assembly cost for making C chairs and the finishing cost for making C chairs.

$$\begin{aligned} \text{Revenue from chairs} &= (\text{selling price of each chair}) * (\text{number of chairs}) = 18C \\ \text{Materials cost for chairs} &= (\text{materials cost for one chair}) * (\text{number of chairs}) = 5C \\ \text{Assembly cost for chairs} &= (\text{hours per chair}) * (\text{cost per hour}) * (\text{number of chairs}) \\ &= (1\text{hours/chair}) * (\$4\text{per hour}) * (C\text{chairs}) = 4C \\ \text{Finishing cost for chairs} &= (\text{hours per chair}) * (\text{cost per hour}) * (\text{number of chairs}) \\ &= (1\text{hours/chair}) * (\$7\text{per hour}) * (C\text{chairs}) = 7C \end{aligned}$$

Thus, the profit for making and selling C chairs is

$$\begin{aligned}\text{Profit for chairs} &= \text{Chair revenue} - \text{Material cost} - \text{Assembly cost} - \text{Finishing cost} \\ &= 18C - 5C - 4C - 7C \\ &= 2C\end{aligned}$$

In exactly the same fashion, we compute the following

Item	Revenue	Material cost	Assembly cost	Finishing cost	Profit
Chairs, C	$18C$	$5C$	$4C$	$7C$	$2C$
Tables, T	$54T$	$15T$	$4T$	$28T$	$7T$
Juice carts, J	$36J$	$10J$	$8J$	$14J$	$4J$

Thus, the objective function is given by the expression:

$$\text{Profit} = 2C + 7T + 4J$$

Notice that this expression is easy to interpret. For each chair we sell, we earn \$2 in profit, for each table we sell, we earn \$7 and for each juice cart, \$4. In the absence of constraints, the only way to maximize this profit is to make as many of each product as possible. Based on the coefficients, it would seem that making tables is the best way to go. However, that approach would completely ignore the time constraints on making the different products. At 1 hour per table, we could assemble a total of 200 tables. But it would take $200 \cdot 4 = 800$ hours of finishing time to complete those tables. We must keep finishing time below 350, so we can make at most $350/4 = 87.5$ tables. But this leaves over half of the assembly hours unused. Perhaps there is some way to make just the right number of each product to use the available resources and increase profits.

Converting the constraints

The constraints on time can easily be converted over. For each chair, we use 1 hour of assembly time. Thus, to make C chairs, we use C hours of assembly time. To make T tables, using 1 hour of assembly time per table, we use up T hours of assembly time. To make J juice carts at 2 hours per cart, we use $2J$ hours of assembly time.

Constraint #1 (Assembly time ≤ 250): $C + T + 2J \leq 250$

Similarly, we find the constraint on finishing time:

Constraint #2 (Finishing time ≤ 350): $C + 4T + 2J \leq 350$

The other constraints are simple to write down, but harder to encode as an expression to be used in doing algebra.

Constraint #3 (number of chairs ≥ 0): $C \geq 0$

Constraint #4 (number of tables ≥ 0): $T \geq 0$

Constraint #5 (number of juice carts ≥ 0): $J \geq 0$

Constraint #6 (number of chairs is integer): C integer

Constraint #7 (number of tables is integer): T integer

Constraint #8 (number of carts is integer): J integer

We have now converted all the expressions into mathematical notation. We could now apply the simplex method or some other technique to solve the problem. Notice that one

of the drawbacks to solving the problem this way is that we have all the numbers “hard coded” into the problem. From the final expressions, it is difficult to see how changing some of the initial information, like the number of hours to assemble a cart or the total number of finishing hours available, will change the expressions without starting over from scratch. In the next section, we will formulate our optimization problems for solution in a spreadsheet. This has the advantage of automatically updating the formulas and expressions based on the new information.

Example 16.3. Minimizing Shipping Costs

CompuTek produces laptops in two cities, Spokane, WS and Atlanta, GA. It purchases screens for these from a manufacturer, Clear Viewing, that has two production facilities, one located in Topeka, KS and the other located in Rochester, NY. CompuTek needs these items shipped to its two facilities. The plant in Topeka can produce at most 2000 units/week, while the plant in Rochester can produce 1800 units per week. Given the schedule below for how much it costs to ship a unit of product from each plant to the different cities where CompuTek needs them, how many should be sent from each plant to each city, if CompuTek needs 1000 units in Spokane and 1200 units in Atlanta?

Shipping Costs From	To	
	Spokane	Atlanta
Topeka	\$3	\$2
Rochester	\$4	\$5

This problem is obviously a minimization problem: we want to keep the shipping costs (our objective function) down to the lowest possible amount. We seem to have four input variables: the amounts shipped from each plant to each final city. And we seem to have two explicit constraints. (1) We cannot ship more from a plant than the plant can produce. (2) We need to ship the right number of units to each city so that CompuTek’s order is filled.

Let’s introduce some variables and express our problem in terms of these variables. We’ll use the following variable names for each of the input variables:

TS = the number of units shipped from Topeka to Spokane
 TA = the number of units shipped from Topeka to Atlanta
 RS = the number of units shipped from Rochester to Spokane
 RA = the number of units shipped from Rochester to Atlanta

Then, we have the objective function, which is the total shipping cost (TSC):

$$\begin{aligned}
 TSC = & \text{\#of units shipped Topeka to Spokane} \times \text{Unit price from Topeka to Spokane} \\
 & + \text{\#of units shipped Topeka to Atlanta} \times \text{Unit price from Topeka to Atlanta} \\
 & + \text{\#of units shipped Rochester to Spokane} \times \text{Unit price from Rochester to Spokane} \\
 & + \text{\#of units shipped Rochester to Atlanta} \times \text{Unit price from Rochester to Atlanta}
 \end{aligned}$$

Thus, we see that the total shipping cost is

$$TSC = 3TS + 2TA + 4RS + 5RA$$

We want to make this number as small as possible subject to the following constraints:

We cannot ship more from Topeka than 2000 units:	$TS + TA \leq 2000$
We cannot ship more from Rochester than 1800 units:	$RS + RA \leq 1800$
We need exactly 1000 units in Spokane:	$TS + RS = 1000$
We need exactly 1200 units in Atlanta:	$TA + RA = 1200$
All quantities need to be positive:	$TA, TS, RA, RS \geq 0$
All quantities need to be integers:	TA, TS, RA, RS integer

16.1.3 Exploration 16A: Setting up Optimization Problems

Break down each of the problems described below so that you know (1) what the objective function is, (2) whether it is to be maximized or minimized, (3) the input variables, (4) the explicit constraints, and (5) the implicit constraints. For most of these situations, the problem is very vaguely stated, and you will need to also describe (6) what further information you will need in order to completely set up the problem.

Situation A.

How should a fast food chain allocate its advertising budget among different advertising formats?

Objective function	(max or min?)
Input variables	
Explicit constraints	
Implicit constraints	
Further information needed	

Situation B.

Where should a small town locate its only school?

Objective function	(max or min?)
Input variables	
Explicit constraints	
Implicit constraints	
Further information needed	

Situation C.

A large drug company has \$5 billion available to acquire small, start-up biotechnology companies. Which companies should it acquire?

Objective function	(max or min?)
Input variables	
Explicit constraints	
Implicit constraints	
Further information needed	

Situation D.

Your company has a list of 15 different strategic initiatives that it would like to undertake. Each will tie up some of your skilled labor pool for the next 4 years, and you do not have the resources to take on all 15 projects. Which projects should you select?

Objective function	(max or min?)
Input variables	
Explicit constraints	
Implicit constraints	
Further information needed	

16.2 Using Solver Table

There are many ways to go about solving an optimization problem with constraints, once you have the constraints written down mathematically as we did in the previous section. Some of these techniques involve a lot of algebra to solve equations, others use matrices and linear algebra to solve large systems of equations simultaneously, others are essentially based on guessing and checking, then changing your guess based on some calculus concepts.

Linear programming (LP) can be used to find solutions to optimization problems when the constraint functions and the objective function are all linear. Such problems typically occur in many operations research situations, such as studying the flow of materials through a network. Because the problem involves only linear functions, there are a variety of algebraic tools to solve them. One commonly used technique is called the simplex algorithm. But don't let the name fool you; this can be a very sophisticated technique, involving things like slack variables and shadow prices. Nonlinear and dynamic programming methods are modifications to LP that relax some of the conditions so that you can apply these techniques to nonlinear situations.

A more general method of optimization comes from multivariable calculus. Once you have defined all of your constraints, you have determined a feasible region for your solutions. You can then use calculus techniques to find any solutions inside this region, and can use the method of Lagrange multipliers to find possible solutions on the boundary of this region. The downside is that, since any kinds of constraints and objective functions can be used, there is no guarantee that you can solve the problem algebraically. You may need to resort to numerical tools to approximate the solution.

If your problem has only two independent variables, you can always graph the objective function and the constraints and used graphical techniques to approximate the solution. This method fails, though, when dealing with more than two independent variables because you cannot graph such functions easily.

The method we will discuss in this section is a combination of these techniques built into Excel called solver table. This lets us create an objective function, define our input variables, and develop a variety of constraints, both explicit and implicit. The solver table will then locate solutions based on the tools we tell it to use. It will also help us perform sensitivity analysis to find out how much our solutions will change if the inputs or constraints change. It is one thing to give your boss a report claiming that the company needs to do X under conditions Y, but it is much more useful to give him a report that says "under conditions Y we should do X, but if conditions Z change, we should modify our approach in the following ways..." This second version not only provides knowledge about the current situation, but also helps the decision maker respond to other situations and to anticipate the possible future directions he or she may need to explore.

16.2.1 Definitions and Formulas

Feasibility region This is the set of all possible values of the independent variables for which it even makes sense to talk about the optimization problem. For example, if you are optimizing profits from selling different quantities of different goods, you do not expect to find the optimum profit when you sell a negative number of an item. That

negative value is outside the feasibility region.

Feasible solution After solving the optimization problem, you need to check whether the solution is inside the feasibility region. If it is, you have a feasible solution.

Gradient function The gradient of a function of several variables represents a list, each item of which is one of the partial derivatives of the function. When solving an unconstrained optimization problem, you are trying to find the value(s) of each independent variable that make the partial derivatives in this list equal to zero simultaneously.

Lagrange multipliers This is a method of solving constrained optimization problems in n variables with m constraints. This method is similar to that of solving unconstrained optimization problems, but instead of trying to make n derivatives simultaneously zero we have to solve $n + m$ equations, a considerably harder task.

Linear programming This is a technique for solving optimization problems when you have a single linear objective function and your constraints are all linear functions of the different variables. It is guaranteed to find a solution, but may not find all the solutions or the best possible solution if there are multiple solutions possible. There are many software packages designed to solve linear programming problems. In fact, Excel's solver table may be manipulated to solve LP problems.

Dynamic programming This is a generalized version of linear programming that can be used to solve optimization problems with more complicated objective and constraint functions. By default, this is the way solver table in Excel works.

Sensitivity and sensitivity analysis When solving an optimization problem with constraints, we often need to find not only the solution to a particular problem but also what happens if we change some of the information slightly. For example, we can optimize profits for various anticipated budget and materials constraints, but need to know how this solution will change if the budget is a little less (or more) or the materials are a little harder to get. This additional analysis helps us determine how sensitive the optimum solution is to changes in the inputs and constraints.

Solver Table This is a program in Excel that enables you to solve various types of optimization problems

lpSolve Linear programming package for R similar to Solver Table.

16.2.2 Worked Examples

Example 16.4. Solving an optimization problem with Solver Table

In this example, we will actually solve and interpret the results for the production example described in the last section. We'll be using file **C16 Furniture**. This relates to the production of tables, chairs and carts that we set up in examples 1 and 2. For information on how to set this up in Excel, see the How To Guide for section 16A.

Once the information is entered and the spreadsheet is prepared, you can complete the solution by invoking the solver routine. First, select the cell containing the profit (C19) and then select “Solver” from the “Tools” menu. You should get a dialog box like the one below, already set up for you (because the file was saved with this solver scenario already set up). Notice that the first box contains the “Target Cell” which is our objective function, the profit (\$C\$19). The next row of information tells solver to maximize the target cell. Notice that you could select “Min” for minimization problems, or you could set this target cell to a particular value. This last is useful for scenarios of like the one below:

I need to make \$700 profit. How many of each product do I need to make?

Problems like this are not, strictly speaking, optimization problems. They are much more like the problems one might use Goal Seek to solve, except solver allows multiple input variables, while Goal Seek only allows one.

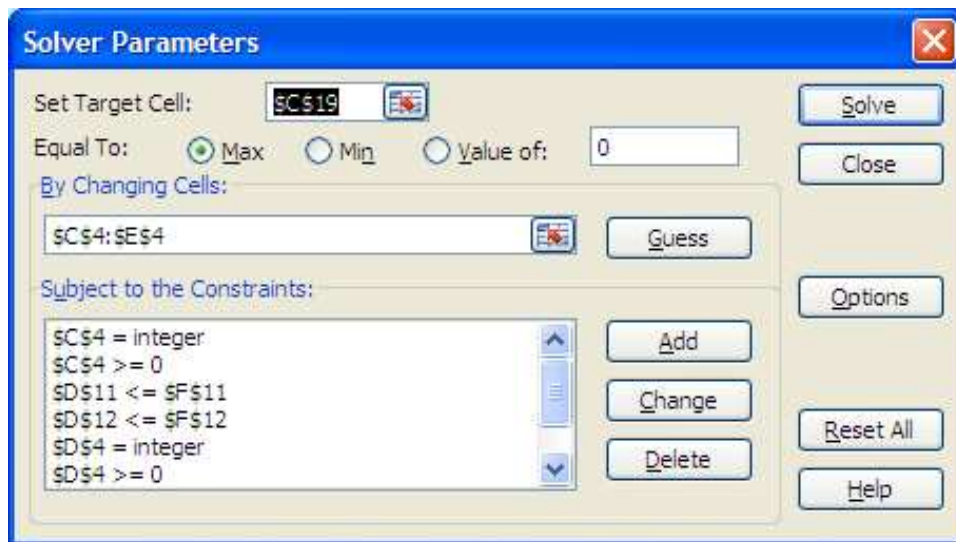


Figure 16.1: Solver dialog box filled in with constraints for solving the furniture problem.

The most difficult part about setting up the solver routine though is the final box, where the constraints are entered. Scrolling down the listed constraints should show you all eight of the constraints identified in example 1. To add another constraint, you simply click the “Add” button and fill in the dialog box (see the How To Guide for details). Once all the constraints are entered, it can be as simple as hitting the “Solve” button to get a solution. In this case, solver returns the following screen.

The “Solver Results” box shows that the routine has found a solution. This solution is displayed in the spreadsheet, and it seems that if we make 62 chairs, 34 tables and 76 juice carts we can maximize our profit at \$666. Notice that this solution leaves us with only 2 unused hours of labor (in the assembly area). Clicking on the “Answer” under “Reports” and then clicking “OK” causes solver to create a new sheet (titled “Answer Report 1”) summarizing the scenario (see the screen below for what this looks like). The report summarizes the initial information and compares this to the final solution. It also summarizes

all of the constraints in the problem and whether the constraint was met exactly or whether some “slack” was allowed.

In this example, we know all of the constraints and the objective function are linear, so we could force Excel to use this information. Clicking on the “options” button in the solver routine, we can check off the “Assume Linear” box. Notice, though, that the solution determined in this way is very different from the solution determined otherwise, even though the maximum profit amount, \$666, is the same!

We could also have left out the constraints that force the quantity of each product to be positive. Instead, we could have selected the “Assume non-negative” option in the “Options” dialog box. The result is the same solution we had originally determined.

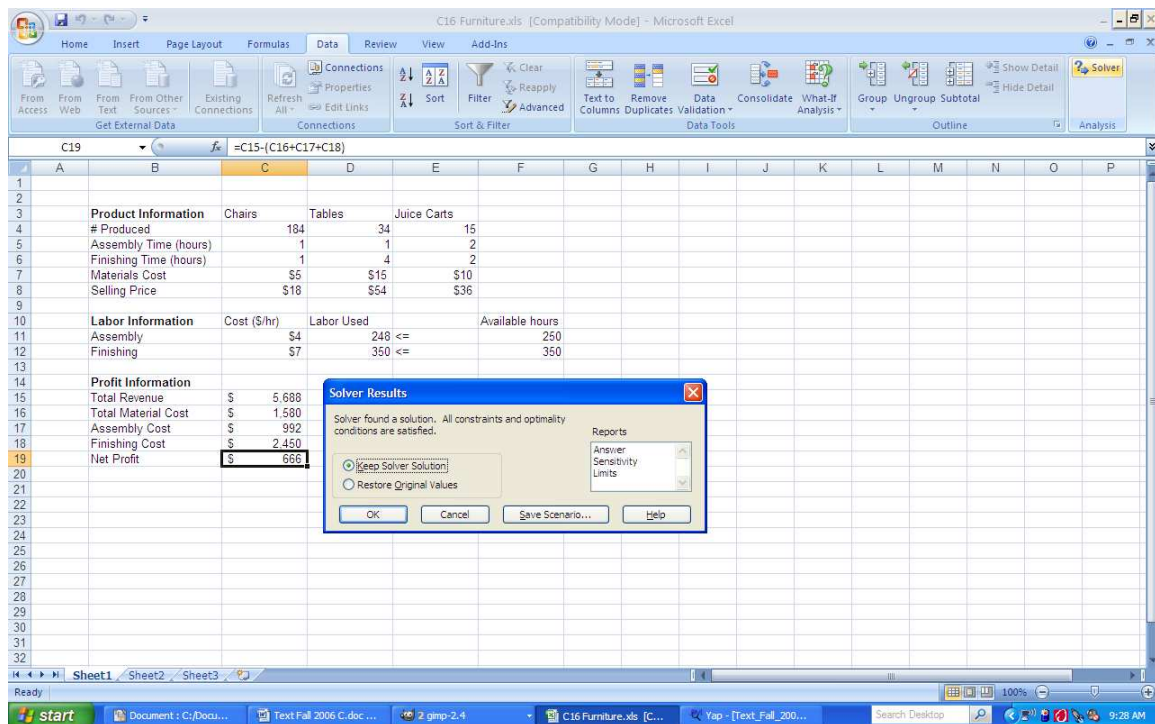


Figure 16.2: Results of solver routine on the furniture problem.

Example 16.5. Solving a minimization problem

Let's solve the problem related to shipping from example 3. We need to organize our spreadsheet carefully so that we can easily change any of the given information in case the scenario changes. We also want to set up the spreadsheet so that the calculations are easy to copy. The screen shot below shows how we have chosen to organize the information.

Notice that the given information about the shipping costs from each plant to the final cities is listed at the top (cells B3:D5), and they are organized into a table much like the one that originally stated the information. This structure will be repeated throughout the setup of this problem. Below that, we have the temporary values for our input variables. These are also organized into a table in cells B8:D10. To the right of the variables are some calculations. Cell E9 contains the total number of units shipped from Topeka ($=C9+D9$),

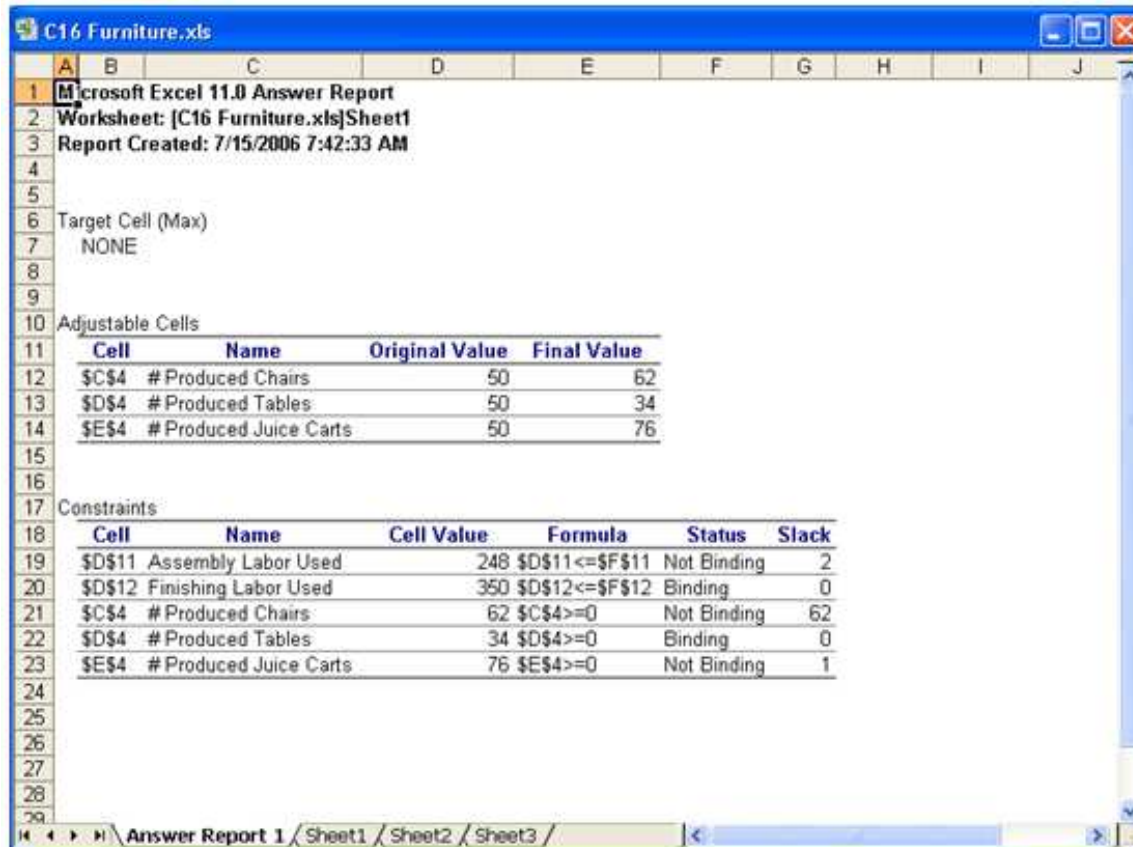


Figure 16.3: Answer report for solver solution of the furniture problem.

which is then copied to cell E10 to get the number of units shipped from Rochester. Cells F9:F10 contain symbols to remind us what the constraints are (\leq) and G9:G10 contain the maximum number of units that can be shipped from the two plants. Below the variables, in cells B12:D13, we have information about the number of units needed in each city.

The next table, cells B15:D17 contains calculations for the total cost of shipping from each plant to each city. For example, in cell C16, the formula “=C9*C4” has been entered; this computes the cost for shipping a total of C9 units from Topeka to Spokane at a cost of C4 dollars per unit. This formula can then be directly copied to the other cells in the table to compute the corresponding amounts of shipping from each plant to each city.

The final piece of the setup is in cell C19: the total shipping cost. This is simply the formula

“=SUM(C16:D17)”. This will be the objective function that we minimize.

Selecting C19 and invoking the solver routine, we make sure to select “minimize” and then we enter our four input variable, found in cells \$C\$9:\$D\$10. There are three constraints:

Constraint	Notation in Solver
Total Shipped \leq Maximum Output	\$E\$9:\$E\$10 \leq \$G\$9:\$G\$10
Total Received = Total Needed	\$C\$11:\$D\$11 = \$C\$13:\$D\$13
All amounts are integers	\$C\$9:\$D\$10 integer

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		Given	To									
3		From	Spokane	Atlanta								
4		Topeka	\$3	\$2								
5		Rochester	\$4	\$5								
6												
7		Variables										
8		Shipping	Spokane	Atlanta	Total from		Max From					
9		Topeka	500	600	1100 <=		2000					
10		Rochester	500	600	1100 <=		1800					
11		Total to	1000	1200								
12		=	=									
13		Needed	1000	1200								
14												
15		Cost	Spokane	Atlanta								
16		Topeka	\$1,500	\$1,200								
17		Rochester	\$2,000	\$3,000								
18												
19		Total cost	\$7,700									
20												

Figure 16.4: Setup for minimizing shipping costs in example 5.

Then, we set the options for the solver routine so that “Assume non-negative” is checked, and solve the problem. The answer report from the solver routine is shown below. It seems we should $TS = 800$, $TA = 1200$, $RS = 200$, and $RA = 0$.

Example 16.6. Special case - two input variables and one constraint

In some situations, optimization problems for multivariable functions are much easier to solve. One of these situations, one that occurs frequently, is when your objective function has two input variables and only one constraint function. In principle, any problem of this kind can be converting into a single variable optimization problem like those solved in chapter 14. Let’s see how this works.

Consider a typical model of production for an economy using a Cobb-Douglas (multiplicative) model like the one below. The two input variables are K , the units of capital invested in the economy, and L , the units of labor invested in the economy.

$$\text{Production} = 300K^{0.6}L^{0.4}$$

In most situations, we would like to maximize the productivity of the economy by setting the labor and capital investments appropriately. Thus, the productivity becomes our objective function. But we cannot choose just any values for K and L . Suppose that each unit of labor costs 85 and each unit of capital costs 130. If costs must be maintained below 100,000, then we want to maximize the productivity above, subject to the constraint that

$$\text{Total Cost} = 130K + 85L \leq 100,000$$

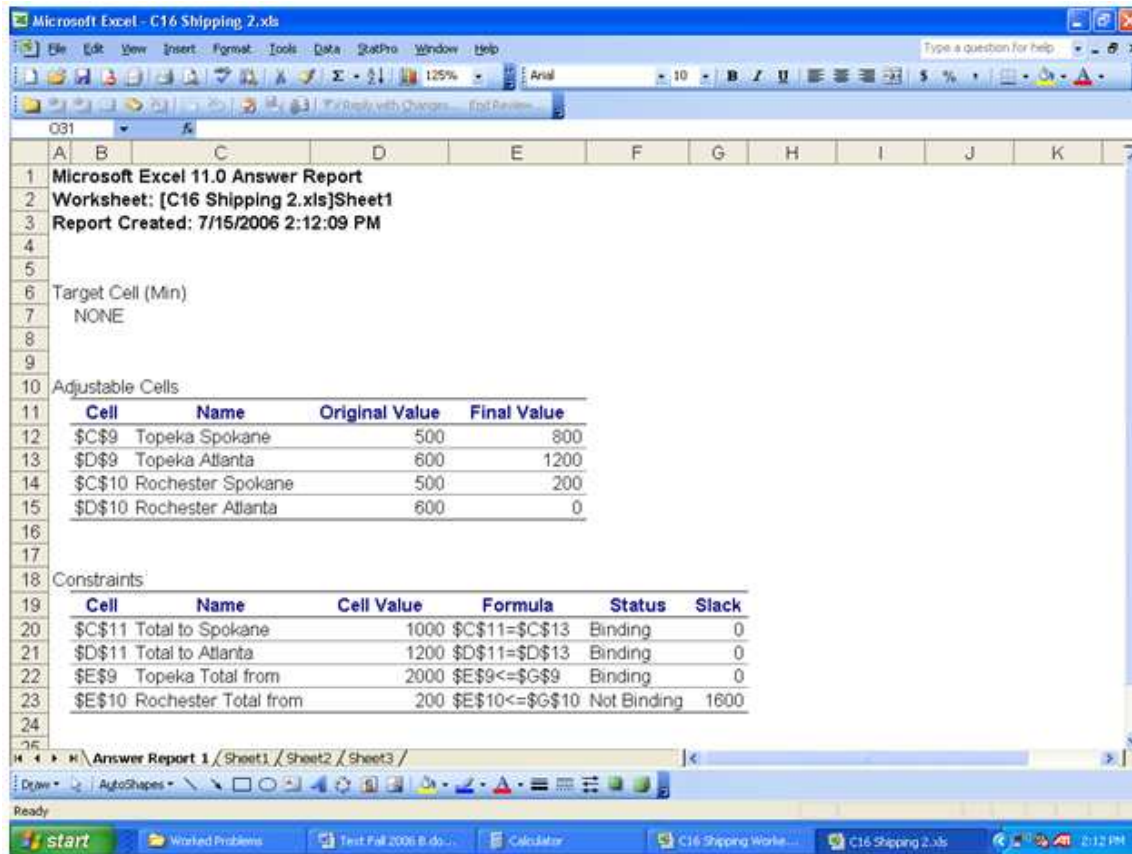


Figure 16.5: Answer report for solver solution of the shipping problem.

Notice that for the type of function we have we can increase the objective function by making the input variables as large as possible. This, and other obvious logic, lead us to conclude that the only place we will get a solution to this problem is if the total cost is exactly equal to its maximum value, 100,000. (If we went for less, then another unit of capital or labor would push the cost closer to the maximum possible cost and would increase the productivity of the economy.) This means that we will take the constraint to hold at equality. That gives us a linear equation, which we can solve for one of the two variables:

$$130K + 85L = 100,000 \rightarrow K = (100000 - 85L)/130$$

If we plug this value of K into the objective function, we get

$$P = 300 \left(\frac{1000000 - 85L}{130} \right)^{0.6} L^{0.4}$$

This is a function of one variable! We can graph it easily and estimate the value of L for which the productivity, P , is maximum. We can then plug this value of L into the equation for K to find the number of units of capital we need to reach maximum productivity. The graph is shown below, indicating that about 500 units of labor are needed. Using the

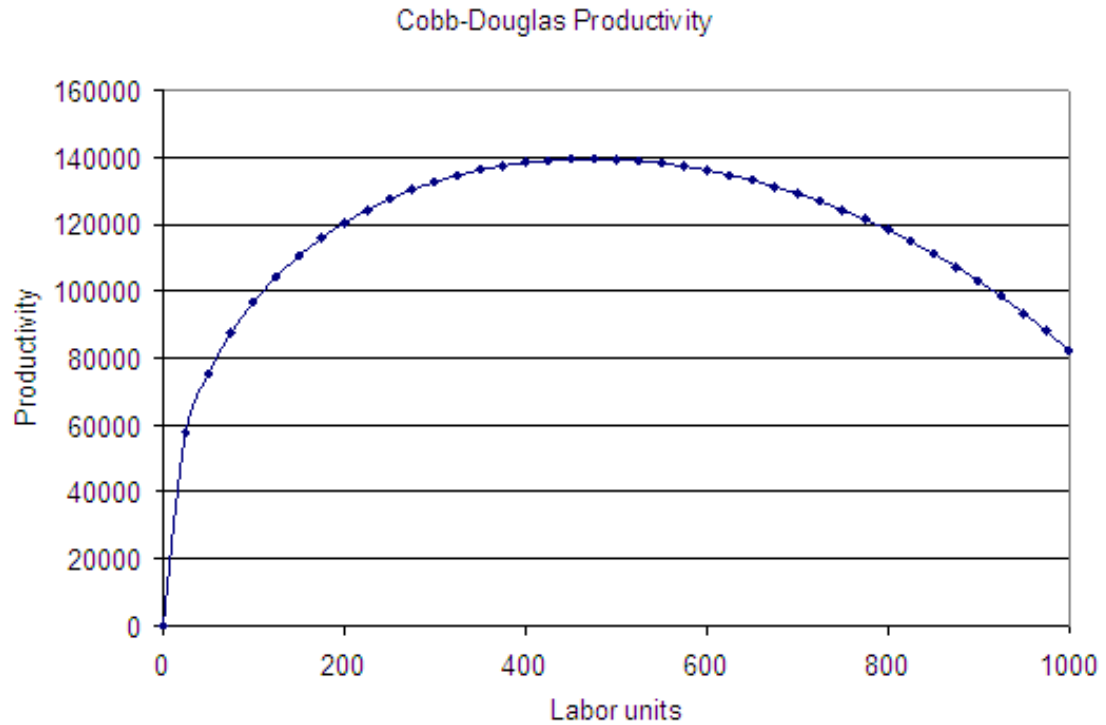


Figure 16.6: Cobb-Douglas productivity function with a constraint that reduces it to a one-variable model.

methods from chapter 14, we find that a better estimate is 470.59 units of labor. This results in 461.53 units of capital needed to achieve maximum productivity.

We could also solve this by setting the derivative of the production function to zero with Goal Seek. Computing the derivative of the objective function (after substituting in the constraint) will require the product rule and the chain rule.

16.2.3 Exploration 16B: Sensitivity Analysis

In this exploration, we will learn a little about “Sensitivity analysis”. We’ll use the example of producing chairs, tables, and juice carts (Examples 1, 2 and 4) and see how the optimal solution changes as we adjust the information about costs and labor hours. Use the file **C16 Furniture2**. Each of the following scenarios is a slight modification to the existing data; simply change the numbers from the original values and run the solver routine to see what happens. Summarize your results in the tables provided.

Scenario A.

It is probably not realistic to assume that each and every chair is assembled in 1 hour. Suppose your assembly crews are a little faster (0.9 assembly hours per chair) or a little slower (1.1 assembly hours per chair). How does the optimal solution change?

Value being adjusted	Assembly time for chairs			
Change being made	Value	Optimal # chairs	Optimal # tables	Optimal # carts
Lower	0.9			
Original value	1.0	62	34	76
Higher	1.1			

Scenario B.

The market probably changes a little, resulting in slight differences in actual selling prices for the products. What happens if the selling price of juice carts fluctuates between \$35 and \$37?

Value being adjusted	Selling price for juice carts			
Change being made	Value	Optimal # chairs	Optimal # tables	Optimal # carts
Lower	\$35			
Original value	\$36	62	34	76
Higher	\$37			

Scenario C.

Just as your selling prices fluctuate, you should expect your material costs to change slightly. Explore what happens if the material cost for the tables is between \$14 and \$16.

Value being adjusted	Material cost for tables			
Change being made	Value	Optimal # chairs	Optimal # tables	Optimal # carts
Lower	\$14			
Original value	\$15	62	34	76
Higher	\$16			

Scenario D.

Suppose the union demands that the workers in the assembly portion get higher wages. Explore what happens if the wages are raised from \$4 to \$5 or \$6 per hour of assembly work.

Value being adjusted	Labor cost for assembly			
Change being made	Value	Optimal # chairs	Optimal # tables	Optimal # carts
Original value	\$4	62	34	76
Increase \$1	\$5			
Increase \$2	\$6			

16.3 Homework

Mechanics and Techniques Problems

In the situations described below in problems 1 and 2, identify the objective function, whether it is a maximization or minimization problem. Also identify the free variables, the explicit constraints and the implicit constraints.

16.1. A company manufactures two products, widgets and greebles, on two machines, I and II. It has been determined that the company will realize a profit of \$6 on each widget and a profit of \$8 on each greeble. To manufacture a widget requires 9 minutes on machine I and 7 minutes on machine II. To manufacture a greeble requires 13 minutes on machine I and 5 minutes on machine II. There are 5 hours of machine time available on machine I and 3 hours of machine time available on machine II in each work shift. How many units of each product should be produced?

16.2. Harbor Tours operates a fleet of ocean vessels for touring the coast around the Alaska. The fleet has two types of vessels: Cruisers have 60 deluxe cabins and 160 standard cabins, while Corvettes have 80 deluxe and 120 standard cabins. Under a chartered agreement with Glacier Travel Agency, Harbor Tours is to provide Glacier Travel with a minimum of 410 deluxe and 720 standard cabins for their 15-day cruise in July. It costs \$65,000 to operate a cruiser and \$82,000 to operate a corvette for that period of time. How many of each type vessel should be used?

16.3. A financier plans to invest up to \$5 million in three projects. She estimates that project A will yield a return of 12% on the investment, project B will yield a return of 16% on the investment and project C will yield a return of 23% on her investment. Because of the risks associated with the investments, she decides to put not more than 20% of the total investment into project C. She also decided that her investments in projects B and C should not exceed 60% of the total investment. Finally she decided that her investment in project A should be at least 60% of her investment in projects B and C.

In this problem, we are obviously trying to maximize the financier's return on her investments (the objective function) by altering the free variables (the total amount of money invested in each of the three projects). We have four explicit constraints: (1) Investment in B should be less than or equal to 20% of the total investment. (2) Investment in B and C should be less than or equal to 60% of the total investment. (3) Investment in A should be greater than or equal to 60% of the total investment in B and C. (4) The total of all the investments must not exceed \$5 million. We also have some implicit constraints: (1) All three investments must be positive or zero. (2) All three investments must be a monetary value with at most two decimal places, when measured in dollars.

1. Write the objective function to be optimized and the constraints (both explicit and implicit) algebraically in as simple a form as possible.
2. Set up a spreadsheet to help solve this problem (but do not solve it).

16.4. Ivory Keys manufactures upright and console pianos in two plants, Boise and Canton. The output of Boise is at most 300/month, whereas the output of Canton is at most 250/month. These pianos are shipped to three warehouses that serve as distribution centers for the company. To fill current and projected future orders, the warehouse in Seattle requires a minimum of 200 pianos/month, the warehouse in Dallas requires at least 150 pianos/month, and the warehouse in Pittsburgh requires at least 200 pianos/month. The shipping costs from Boise to Seattle, Dallas and Pittsburgh are \$60, \$60, and \$80, respectively. The shipping costs from Canton to Seattle, Dallas and Pittsburgh are \$80, \$70, and \$50, respectively.

Clearly, in this problem we want to minimize the shipping costs (objective function) by altering the free variables (the number of each type of piano shipped from Boise to Seattle, Dallas and Pittsburgh and the number of each type of piano shipped from Canton to Seattle, Dallas and Pittsburgh; that's six variables!). We have five explicit constraints: (1) The number of pianos shipped from Boise cannot exceed 300. (2) The number of pianos shipped from Canton cannot exceed 250/month. (3) Seattle needs at least 200 pianos/month. (4) Dallas needs at least 150 pianos/month. (5) Pittsburgh needs at least 200 pianos/month. Finally, the only implicit constraints are that each shipping value must be positive and an integer.

1. Write the objective function to be optimized and the constraints (both explicit and implicit) algebraically.
2. Set up a spreadsheet to help solve this problem (but do not solve it).

16.5. Return to the furniture production example explored in this chapter. Modify the spreadsheet to include the additional constraint that you must produce at least four times as many chairs as tables. Solve this problem with the additional constraint.

Application and Reasoning Problems

16.6. Consider the situation described in #1 above. Under the stated conditions, the maximum profit of \$192 per shift comes from making 16 Widgets and 12 Greebles. The following questions relate to the way the optimal solution changes under different conditions in the environment. Open file **C16 Problem6**. The problem has been set up for you, and the constraints and settings for Solver are already configured.

1. Suppose one of your work crews finds a way to cut the time for making the Greebles on machine I. Instead of needing 13 minutes, they need only 12 minutes on machine I. See the screen shot below for what this might look like in Excel. Notice that you now have enough time left to make either one more Widget or one more Greeble. Which would be better? Why?

	A	B	C	D	E	F
1						
2		Product	Machine I	Machine II	Profit	Number
3		Widgets	9	7	\$6.00	16
4		Greebles	12	5	\$8.00	12
5						
6		Minutes				
7		Machine I	288	\leq	300	
8		Machine II	172	\leq	180	
9						
10		Profit	\$192.00			

16.7. Open file **C16 Problem6**. The problem has been set up for you, and the constraints and settings for Solver are already configured. Use Solver to help you explore the following change to the situation: Suppose the materials costs for Widgets are expected to increase. This will reduce your profit on each Widget. At what point should we drastically cut back on Widget production? At what point should we cut Widget production altogether?

16.8. Consider the situation described in #2 above. Under the given conditions, the optimal solution is shown in figure 16.13. Data file **C16 Problem7** has this problem set up for you, including the solver table constraints.

	A	B	C	D	E	F
1						
2			Deluxe	Standard	Number	Cost
3		Cruiser	60	160	3	\$65,000
4		Corvette	80	120	3	\$82,000
5						
6		Deluxe	420	\geq	410	
7		Standard	840	\geq	720	
8						
9		Total Cost				
10		\$441,000.00				

1. Given the current optimal solution, how many additional cabins of each type could be rented for the cruise?
2. Suppose the cost for operating the Corvettes is expected to increase. At what point should Harbor Tours supply Glacier Travel Agency with only Cruisers?

16.9. The file **C16 Problem8** shows the set up for the optimization problem below. C-Vite Company has decided to introduce three fruit juices made from blending two or more

concentrates. These juices will be packaged in 2-qt (64 fluid oz) cartons. To make one carton of pineapple-orange juice requires 8 oz each of pineapple and orange juice concentrates. To make one carton of orange-guava juice requires 4 oz of guava concentrate and 12 oz of orange concentrate. Finally, to make one carton of pineapple-orange-guava juice requires 4 oz of pineapple juice concentrate, 4 oz of orange juice concentrate and 8 oz of guava juice concentrate. The company has decided to allot 22,000 oz of pineapple juice concentrate, 28,000 oz of orange juice concentrate and 12,000 oz of guava juice concentrate for the initial product run. The company also stipulated that the production of the pineapple-orange-guava juice should not exceed 900 cartons. Its profit on one carton of pineapple-orange juice is \$1.00; its profits on one carton of orange-guava juice is \$0.90; and its profit on one carton of pineapple-orange-guava is \$0.95.

1. Explain how the spreadsheet sets up this problem for solution. Point out any unnecessary information provided in the problem statement (such unnecessary items are often called “red herrings”).
2. Find the optimal solution using Solver.
3. C-Vite is worried about a bad winter which could damage much of the Florida orange crop. If C-Vite only has 20,000 oz of orange juice concentrate to allot for the blends, how does this change their optimal production plan?
4. After a taste test, C-Vite decides to change the blend in the Pineapple-Orange-Guava juice to 4 oz pineapple, 6 oz orange and 6 oz guava. What optimal product mix should they pursue?

16.10. Oregon Lumber has decided to enter the pre-fabricated housing market. For its initial venture, it plans to offer three models of homes: traditional, deluxe and luxury. Each house is prefabricated and partially assembled at the factory. The final completion of the home takes place on site. The table below shows the costs, in material and labor, and the profit from each type of home.

	Traditional	Deluxe	Luxury
Material	\$6,028	\$8,062	\$10,135
Factory Labor (hr)	245	222	200
On-Site Labor (hr)	178	211	300
Profit	\$3,400	\$4,000	\$5,000

During the first year, Oregon Lumber has \$9 million to spend on materials. They cannot exceed 230,000 hours of labor in the factory, and they cannot exceed 245,000 hours of labor on site. Assuming that the market can sell as many houses as Oregon Lumber makes, how many of each type should be made if they want to maximize their profit?

CHAPTER 17

Area Under a Curve¹

In Chapters 14 and 15 we learned how to find the rate of change function, the derivative, of various functions. We used the derivative in business applications to find such things as marginal cost or marginal profit functions and to study optimization of cost or revenue. In this chapter we do the opposite: For example, instead of beginning with the cost function c and then finding the marginal cost function by applying the rules of differentiation to c , we begin with and find c by reversing the rules of differentiation. This process is called antidifferentiation or finding the indefinite integral. Mathematics and science majors usually take an entire course in differentiation and then follow it with another course devoted to integration. In these courses, students study a rather amazing idea, called the Fundamental Theorem of Calculus; namely, an antiderivative c of a derivative function (found by reversing the rules of differentiation) is intimately connected to the area under the graph of the derivative function $f'(x)$. Although we will study some of the basic rules of antidifferentiation in order to find the area under various curves by using the Fundamental Theorem of Calculus, we will also use spreadsheets to find approximate numerical answers to finding the area, approximations that serve us quite well in real-life situations. This process is called numerical integration. Indeed, for some important functions there is no known way of finding their antiderivative and, as a result, numerical integration is the only way we have of finding the area under the graph of these particular functions.

- Section 17.1 introduces the idea of the integral as both an area and a summation of lots of small parts, something that is commonly done in business applications.
 - Section 17.2 demonstrates the use of the integral to solve problems that arise in business situations.
1. In the first section of the chapter, we will use both numerical integration and the Fundamental Theorem of Calculus to find the area under a curve.

¹©2017 Kris H. Green and W. Allen Emerson

2. In the second section, we apply finding the area under a curve to some business applications.

As a result of this chapter, students will learn

- ✓ How to find antiderivatives, i.e. the indefinite integral, of certain basic functions
- ✓ How to use the Fundamental Theorem of Calculus to compute the area under a curve, i.e. the definite integral
- ✓ How to use numerical integration to compute the area under a curve

As a result of this chapter, students will be able to

- ✓ Use an integration tool to find the definite integral
- ✓ Compute the area between two curves
- ✓ Apply numerical integration to find the total cost of production
- ✓ Compute future and present value of an income stream
- ✓ Compute consumers' and producers' surplus

17.1 Calculating the Area under a Curve

In this section we investigate three interrelated concepts concerning the area under the graph of a function:

1. Approximating the area under the curve by a finite number of rectangles (Example 1 and Exploration 17A) and then seeing how an infinite number of these rectangles could give the exact area under the curve
2. Finding the antiderivative of the function (Example 2)
3. The Fundamental Theorem of Calculus, which connects 1 and 2 above (Example 3)

In (1) above, we approximate the area under a curve by summing the areas of certain rectangles. The more of these rectangles we use, the more closely their sum approximates the true area under the curve. If we sum an infinite number of these rectangles, we will have the exact area under the curve. In (2) above, we find the antiderivative function f , also called the indefinite integral, denoted by $\int f'(x)dx$, by reversing the rules on the derivative function f' . In (3) above, the Fundamental Theorem of Calculus ties the area under the curve of f' to its antiderivative f as follows:

$$\text{The area under } f' \text{ from } a \text{ to } b = f(b) - f(a).$$

That is, the area under the curve $y = f'(x)$ from $x = a$ to $x = b$, $\int_a^b f'(x)dx$, is the difference of the antiderivative function, i.e. the indefinite integral, evaluated at b and at a . Symbolically, the Fundamental Theorem is written: $\int_a^b f'(x)dx = f(b) - f(a)$, where a is called the lower limit of the integral and b is called the upper limit. Another common way of symbolizing the Fundamental Theorem is: $\int_a^b f(x)dx = F(b) - F(a)$ where $F' = f$. Here F is the antiderivative function of the derivative function f .

We turn now to the procedure for finding the approximate area under a curve. We will illustrate the procedure by taking as our derivative function, f' , a marginal cost function that we will denote by $c'(x)$, where $c'(x)$ is the marginal cost of producing x items of a commodity. We wish to find the area under $y = c'(x)$ from a to b (see the left half of figure 17.1). This area will turn out to be the total variable cost of producing from a to b items of the commodity.

We divide the interval from a to b into n equal subintervals of width (see the right half of figure ?? in which we use 5 subintervals) where $\Delta x = \frac{b-a}{n}$.

For ease of notation, we will rewrite these subintervals as follows: $x_0 = a$, $x_1 = a + \Delta x$, $x_2 = a + 2\Delta x$, $x_3 = a + 3\Delta x$, $x_4 = a + 4\Delta x$. If we draw a vertical line segment from $x = a$ to the segment's intersection with the curve (see figure 17.2), the height of this segment represents the marginal cost at a , i.e. the cost of producing one more item when we have already produced a items.

We then create a rectangle with width Δx , the additional items produced beyond a . The area of this rectangle is $c'(x)\Delta x$, which is the approximate cost of producing the Δx items from x_0 to x_1 . (See the right half of figure 17.2.)

We will deal with the problem of the area of the rectangle underestimating the true area under the curve from x_0 to x_1 shortly. Nevertheless, we continue by constructing a second

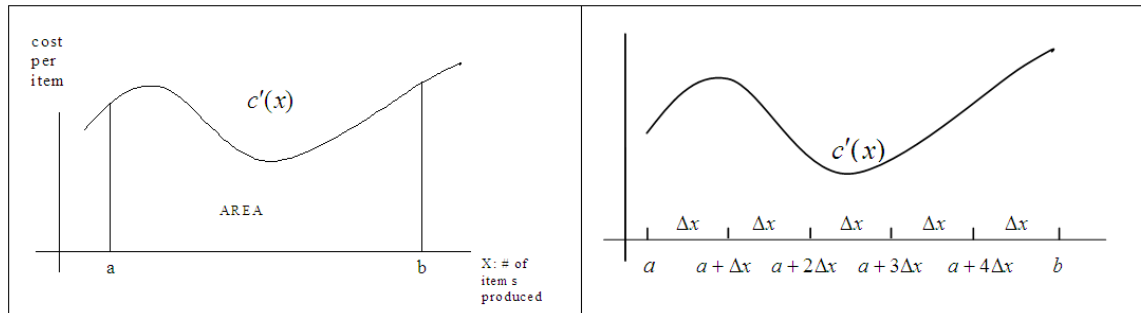


Figure 17.1: Graph showing a marginal cost function between two limits, a and b (left) and showing the interval from a to b broken into 5 subintervals of equal size Δx .

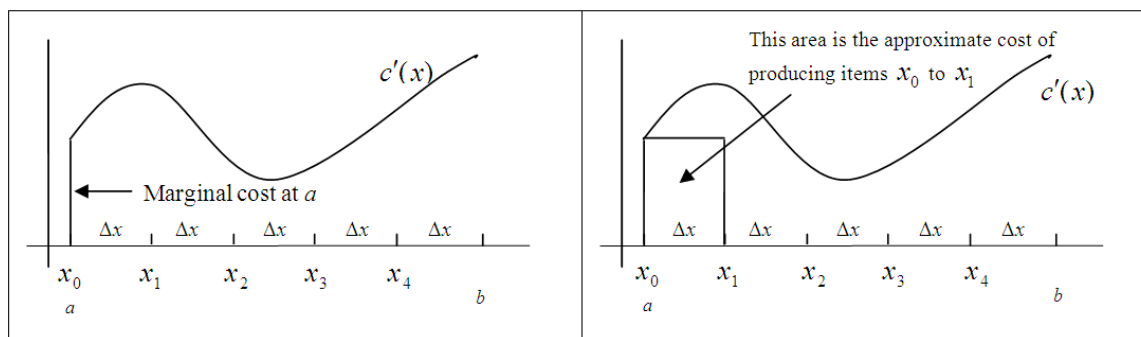


Figure 17.2: Illustration of breaking the area under the marginal cost function into subrectangles that approximate the area.

rectangle with a vertical line segment through x_1 to the curve c' with width Δx . The area of this rectangle is the approximate cost of producing items x_1 to x_2 , which happens to overestimate the true area under the curve from x_1 to x_2 (see Figure 17.5).

Filling in the remaining rectangles in a similar fashion, we find that the sum of these five rectangles is the approximate area under the curve from a to b (see Figure 17.3).

The approximate area under curve from a to b is

$$c'(x_0)\Delta x + c'(x_1)\Delta x + c'(x_2)\Delta x + c'(x_3)\Delta x + c'(x_4)\Delta x.$$

We can imagine constructing an arbitrary number of rectangles n as we have the five above, each of which has width $\Delta x = \frac{b-a}{n}$:

$$\sum_{i=0}^{n-1} c'(x_i)\Delta x = c'(x_0)\Delta x + c'(x_1)\Delta x + c'(x_2)\Delta x + c'(x_3)\Delta x + c'(x_4)\Delta x$$

This is an example of a Riemann Sum ("Rie" rhymes with "me" and "mann" rhymes with "Don." There are many different ways of constructing rectangles under the curve from a to b , some more accurate and efficient than others. Nonetheless, it turns out that each will lead to the same place, the true area under the curve from a to b . This is how:

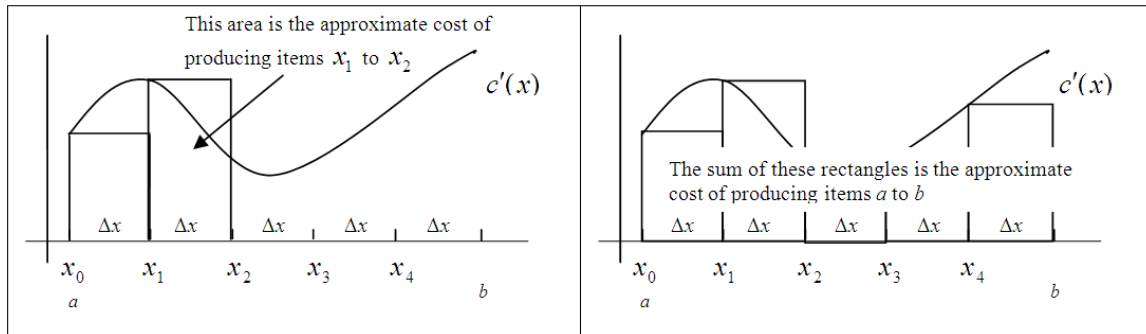


Figure 17.3: Graph showing a marginal cost function between two limits, a and b (left) and showing the interval from a to b broken into 5 subintervals of equal size Δx .

We increase the number n of rectangles and correspondingly shrink the width Δx . The over and under estimations of the rectangles decrease as the number of rectangles increases. Mathematically, though not geometrically, it turns out that an infinite number of “rectangles” can be packed under the curve and summed to give the exact area under the curve from a to b . This is denoted by $\lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} c'(x_i) \Delta x = \int_a^b c'(x) dx$ and is read “the limit of the Riemann sum $\sum_{i=0}^{n-1} c'(x_i) \Delta x$ as n goes to infinity (∞ is the symbol for infinity) is the definite integral of $c'(x)$ from a to b .”

17.1.1 Definitions and Formulas

Indefinite Integral or antiderivative F is an antiderivative function of f if $F'(x) = f(x)$. The antiderivative is also called the indefinite integral of f and is denoted by $\int f(x) dx + C$, where C is a constant.

Constant of integration If $F(x)$ is an antiderivative function of f , then $F(x) + C$, where C is any real number, is likewise an antiderivative function of f since $\frac{d}{dx}(F(x) + C) = \frac{d}{dx}F(x) + 0 = f(x)$. C is called the constant of integration for the indefinite integral.

Definite integral and the limit of a Riemann sum The definite integral from $x = a$ to $x = b$ is the limit of the Riemann sum $\lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} c'(x_i) \Delta x$ as $n \rightarrow \infty$, where $\Delta x = \frac{b-a}{n}$. The definite integral is denoted by $\int_a^b f(x) dx$.

Lower limit and upper limit a is called the lower limit of the integral $\int_a^b f(x) dx$ and b is called the upper limit.

Fundamental Theorem of Calculus The Fundamental Theorem of Calculus states that $\int_a^b f(x) dx = F(b) - F(a)$, where F is an antiderivative of f .

Area Under a Curve If $f(x)$ is positive from a to b , then the definite integral $\int_a^b f(x) dx$ computes the area under $f(x)$ and above the x -axis from a to b .

Numerical Integration Numerical integration approximates the definite integral

$$\int_a^b f(x)dx = \lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} f(x_i)\Delta x$$

by $\sum_{i=0}^{N-1} f(x_i)\Delta x$, where N is a very large number. There are several different methods of numerical integration but this text uses the method of rectangles for simplicity and ease of discussion.

17.1.2 Worked Examples

Example 17.1. Approximating the Area Under a Curve

Let $c'(x) = \frac{1}{2}x + 1$. Approximate the area under the graph of $y = c'(x)$ from $x = 1$ to 3 with four rectangles.

We first find the width of each rectangle: $\Delta x = \frac{3-1}{4} = 0.5$.

We next calculate the Riemann sum from 1 to 3:

$$\begin{aligned} \text{Area} &\approx c'(1)(0.5) + c'(1.5)(0.5) + c'(2)(0.5) + c'(2.5)(0.5) \\ &= \left[\frac{1}{2}(1) + 1 \right] (0.5) + \left[\frac{1}{2}(1.5) + 1 \right] (0.5) + \left[\frac{1}{2}(2) + 1 \right] (0.5) + \left[\frac{1}{2}(2.5) + 1 \right] (0.5) \\ &= [0.5 + 1](0.5) + [0.75 + 1](0.5) + [1 + 1](0.5) + [1.25 + 1](0.5) \\ &= 3.75. \end{aligned}$$

So, the approximate area from 1 to 3 found by summing the areas of the four rectangles is 3.75. If we were to increase the number of rectangles to 16, then 64, then 1000, and then 10000, we would obtain the following results for the approximate area under the curve from 1 to 3:

Number of Rectangles	Width of Each Rectangle	Approximation to Area
$n = 4$.5	3.75
$n = 16$.125	3.9375
$n = 64$.03125	3.984375
$n = 1000$.002	3.999
$n = 10000$.0002	3.9999

According to the Fundamental Theorem of Calculus (to be illustrated in example 3), the exact area under the curve from 1 to 3 is 4, which you might believe from the table, which suggests that as the number of rectangles used to estimate the area increases, the estimated area approaches the number 4. By using a very large number of rectangles, we can approximate the area under a curve quite closely. Although the calculations are staggering when computed with pencil and paper, computers can handle the computations in a fraction of a second (see Exploration 17A). The process we have described is an example

of numerical integration. There are several methods other than using rectangles to compute definite integrals by numerical integration.

Example 17.2. Finding Antiderivatives for Some Basic Functions

If F is an antiderivative function of f , then $F'(x) = f(x)$. But then $F(x) + 2$ is also an antiderivative function of f since the derivative of a constant, like the number 2, is zero. In general, if $F(x)$ is an antiderivative of f , then so is $F(x) + C$, where C is any real number. C is called the constant of integration for F . Note: We will write the constant of integration in upper case in order to distinguish it from a function c or $c(x)$, which we will write in lower case.

General Rule	Examples
$\int a \, dx = ax + C$, where a is any constant	$\int x \, dx = x + C$ since $\int dx = \int 1 \, dx$ $\int -2 \, dx = -2x + C$ $\int 5 \, dx = 5x + C$
$\int x^n \, dx = \frac{1}{n+1}x^{n+1} + C$ where $n \neq -1$	$\int x \, dx = \frac{1}{2}x^2 + C$ since $x = x^1$ $\int x^2 \, dx = \frac{1}{3}x^3 + C$ $\int x^{-3} \, dx = \frac{1}{-2}x^{-2} + C = -\frac{1}{2x^2} + C$ since $-3 + 1 = -2$
$\int x^{-1} \, dx = \ln x + C$	$\int \frac{dx}{x} = \ln x + C$ since $\frac{1}{x} = x^{-1}$ See note below ²
$\int a f(x) \, dx = a \int f(x) \, dx + C$ where a is any constant	$\int 2x^3 \, dx = 2 \int x^3 \, dx = 2 \left(\frac{x^4}{4} \right) + C = \frac{1}{2}x^4 + C$ $\int \frac{2}{x} \, dx = 2 \int \frac{dx}{x} = 2 \ln x + C$
$\int e^{bx} \, dx = \frac{1}{b}e^{bx} + C$	$\int e^{0.2x} \, dx = \frac{1}{0.2}e^{0.2x} + C = 5e^{0.2x} + C$ $\int 10e^{-0.005x} \, dx = 10 \int e^{-0.005x} \, dx =$ $= \frac{10}{-0.005}e^{-0.005x} + C = -2000e^{-0.005x} + C$
$\int (f(x) + g(x)) \, dx$ $= \int f(x) \, dx + \int g(x) \, dx$	$\int (-4x^{-2} + 2x^{-1}) \, dx = \frac{-4}{-1}x^{-1} + 2 \ln x + C = \frac{4}{x} + \ln x + C$ $\int (-x^2 + 4x - 2) \, dx = -\frac{1}{3}x^3 + \frac{4}{2}x^2 - 2x + C$

Example 17.3. Using the Fundamental Theorem of Calculus to Find Total and Variable Costs from Marginal Cost

Suppose we have gathered daily marginal cost for manufacturing a particular item and found its regression model to be $c'(x) = 0.0002x^2 - 0.1x + 30$, where $c'(x)$ is measured in dollars per item and x is the number of units produced. The fixed cost of producing any number of items is \$550.

Part a. Find the total cost in producing the first 350 units per day.

Before using the Fundamental Theorem of Calculus, we need to find the antiderivative of $c'(x) = 0.0002x^2 - 0.1x + 30$.

$$\int (0.0002x^2 - 0.1x + 30) \, dx = \frac{0.0002}{3}x^3 - \frac{0.1}{2}x^2 + 30x + C$$

Since the fixed cost of producing zero items is \$550, we know that $C = 550$. (Substitute $x = 0$ into the antiderivative and set it equal to 550; this gives $C = 550$). Applying the Fundamental Theorem of Calculus, we find the variable cost of producing the first 350 items:

$$\begin{aligned}
\int_0^{350} (0.0002x^2 - 0.1x + 30)dx &= \left[\frac{0.0002}{3}x^3 - \frac{0.1}{2}x^2 + 30x + C \right]_0^{350} \\
&= \left[\frac{0.0002}{3}(350)^3 - \frac{0.1}{2}(350)^2 + 30(350) + 550 \right] - [0 + 550] \\
&= \frac{0.0002}{3}(350)^3 - \frac{0.1}{2}(350)^2 + 30(350) + 550 - 550 \\
&= \$7233.\bar{3} \approx \$7233
\end{aligned}$$

NOTE: In the definite integral, the constants of integration always cancel each other out (see $550 - 550$ above). Therefore, when we compute the definite integral, we will omit the C , the constant of integration.

\$7233 is the variable cost of producing the first 350 items. The total cost must include the fixed cost, \$550 of producing any number of items:

$$\text{Total Cost} = \$7233 + \$550 = \$7783.$$

Part b. What is the variable cost of producing the 151st through the 350th unit?

$$\begin{aligned}
\int_{150}^{350} (0.0002x^2 - 0.1x + 30)dx &= \left[\frac{0.0002}{3}x^3 - \frac{0.1}{2}x^2 + 30x + C \right]_{150}^{350} \\
&= \left[\frac{0.0002}{3}(350)^3 - \frac{0.1}{2}(350)^2 + 30(350) \right] \\
&\quad - \left[\frac{0.0002}{3}(150)^3 - \frac{0.1}{2}(150)^2 + 30(150) \right] \\
&= \$3633
\end{aligned}$$

NOTE: Since the cost of producing the 151st item begins just after having produced the 150th item, the left-hand height of the first rectangle in the Riemann sum begins at $x = 150$.

17.1.3 Exploration 17A: Numerical Integration

Find $\int_0^{350} (0.0002x^2 - 0.1x + 30)dx$ by numerical integration.

Bring up the file **C17 Integration Tool**. Copy it to a new worksheet and save it under other file name. Follow the How To Guide in order to see how to modify this worksheet to perform numerical integration for this function.

Your answer should closely match the answer to this same integral in example 1 (part a), which is the exact area as found by the Fundamental Theorem of Calculus.

17.2 Applications of the Definite Integral

In this section we will apply the definite integral to three useful analytical tools for business:

1. the future value of an income stream,
2. the present value of an income stream, and
3. consumers' and producers' surplus.

The first of these, the future value of an income stream, measures the value of an income stream by calculating the accumulated total of a continuing stream of revenue invested at a continuous rate r over a period of T years. The present value of an income stream is another way of measuring the value of an income stream. The present value is the lump-sum principle that would have to be invested now for a period of T years at a continuous rate of interest r in order to equal the future value of the income stream continually invested at the same rate over the same time T . That is, the bigger the future value of the stream, the more up-front principal P would have to be invested now at the same rate r in order to come out the same as the future value after T years.

Consumers' and suppliers' surplus help management evaluate the unit price of a commodity, i.e. whether it is too low or too high or just about right for the market.

17.2.1 Definitions and Formulas

Demand Function Expresses consumer demand for a product in terms of its unit price p and the number of units x that consumers will buy at price p . The demand function $D(x)$ creates a decreasing (downhill) curve because fewer products will be sold if p is larger; conversely, smaller prices will create more demand.

Consumers' Surplus Let \bar{p} be a fixed established price for a commodity and \bar{x} be the number of units bought at \bar{p} . The consumers' surplus is the difference between what consumers would be willing to pay for a commodity and what they actually pay for it. The formula for consumers' surplus is: $CS = \int_0^{\bar{x}} D(x) dx - \bar{p}\bar{x}$, where $D(x)$ is the demand function.

Supply Function Expresses producers willingness to supply x units of a commodity at price p . The supply function $S(x)$ creates an increasing (uphill) curve because producers are willing to put more of the commodity on the market if the unit price is higher.

Producer's Surplus Let \bar{p} be a fixed established price for a commodity and \bar{x} be the number of units that producers are willing to supply at price \bar{p} . The producers' surplus is the difference between what the suppliers actually receive \bar{x} and what they would be willing to receive at price \bar{p} . The formula for producers' surplus is: $PS = \bar{p}\bar{x} - \int_0^{\bar{x}} S(x) dx$ where $S(x)$ is the supply function.

Market Equilibrium, equilibrium quantity, equilibrium price The point (\bar{x}, \bar{p}) where the demand curve and the supply curve intersect, i.e. the point at which market equilibrium occurs. This is the highest price consumers are willing to pay for what producers are willing to supply. \bar{x} is called the equilibrium quantity and \bar{p} is called the equilibrium price.

Income Stream Created when a business generates a stream of income $R(t)$, where t is in years, over a period of time T years and this income is invested at an annual rate r compounded continuously. $R(t)$ could be a constant stream or a variable stream but it is invested, nonetheless, on a continuing basis over the T years.

Future Value of an Income Stream The total amount of money that will be accumulated when an income stream $R(t)$ has been invested at an annual rate r compounded continuously for T years. The formula for the future value of an income stream $R(t)$ is: $FV = e^{rt} \int_0^T R(t)e^{-rt} dt$.

Present Value of an Income Stream The principal P that would have to be invested at an annual rate r compounded continuously over T years in order to equal the accumulated value of an income stream over the same period T and the same rate r . The formula for the future value of an income stream $R(t)$ is: $PV = \int_0^T R(t)e^{-rt} dt$.

17.2.2 Worked Examples

Example 17.4. Future Value of an Income Stream

Let

$$\begin{aligned} R(x) &= \text{Rate of income at time } t \\ r &= \text{Interest rate compounded continuously} \\ T &= \text{Number of years the income stream is invested} \end{aligned}$$

We divide the interval $[0, T]$ into n subintervals of length $\Delta t = \frac{T}{n}$ and create n rectangles under the $R(x)$ as seen in the figure below.

The height of the i th rectangle is $R(t_i)$ and its width is Δt . $R(t_i)\Delta t$, the area of the i th rectangle, is the approximately the amount of the income stream to be invested between t_{i-1} and t_i . If we think of investing this small principal at the continuous rate r for a period of $T - t_i$ years (this is the remaining time for the investment from t_i to T), then the amount that will be accumulated after T years is $[R(t_i)\Delta t]e^{r(T-t_i)}$. This formula is derived from the compound interest formula $A = Pe^{rt}$ (see example 4) where $P = R(t_i)\Delta t$ and e^{rt} is replaced by $e^{r(T-t_i)}$. Therefore, the Riemann sum of the future values of the areas of the rectangles is

$$\begin{aligned} [R(t_1)\Delta t]e^{r(T-t_1)} &+ [R(t_2)\Delta t]e^{r(T-t_2)} + [R(t_3)\Delta t]e^{r(T-t_3)} + \dots + [R(t_n)\Delta t]e^{r(T-t_n)} \\ &= [R(t_1)\Delta t]e^{rT}e^{-t_1} + [R(t_2)\Delta t]e^{rT}e^{-t_2} + [R(t_3)\Delta t]e^{rT}e^{-t_3} + \dots + [R(t_n)\Delta t]e^{rT}e^{-t_n} \end{aligned}$$

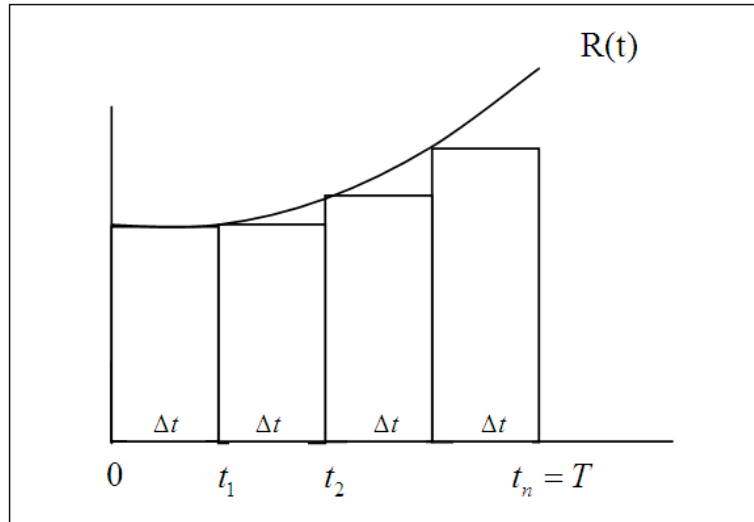


Figure 17.4: Using integration to find the future value of an income stream.

Factoring out e^{rT} and Δt , we have

$$= e^{rT} \left[[R(t_1)]e^{-t_1} + [R(t_2)]e^{-t_2} + [R(t_3)]e^{-t_3} + \dots + [R(t_n)]e^{-t_n} \right] \Delta t$$

Letting $n \rightarrow \infty$, we have the following result:

The future value after T years of an income stream $R(t)$ dollars per year, invested at the rate r per year compounded continuously, is

$$FV = e^{rT} \int_0^T R(t)e^{-rt} dt$$

To illustrate this idea, consider Glow Health Spa. Glow Health Spa recently bought a full-body dermal treatment machine that is expected to generate \$50,000 in revenue per year for the next 5 years. If this income stream is invested at 8% per year compounded continuously, what is the future value of this income stream in 5 years?

$R(t)$ = 50,000 where t is in years
 Note: the income stream $R(t)$ is a constant amount per year in this example.
 r = .08
 T = 5 years
 Future Value = $e^{0.08(5)} \int_0^5 50000e^{-0.08t} dt = 1.49 \int_0^5 50000e^{-0.08t} dt$
 = $1.49(206054) = \$307,020$
 where the definite integral is calculated using the Basic Integration Tool.

Example 17.5. Present Value of an Income Stream

The present value of an income stream of $R(t)$ dollars per year over T years, earning interest at the rate of r per year compounded continuously, is the principal P that would have to be invested now to yield the same accumulated value as the investment stream would earn if it were invested on a continuing basis for T years at rate r .

In equation form, we have

$$Pe^{rT} = e^{rT} \int_0^T R(t)e^{-rt} dt.$$

Dividing both sides of this equation by e^{rT} , we have $PV = \int_0^T R(t)e^{-rt} dt$, the present value of the income stream $R(t)$.

To illustrate the present value of an income stream, we will consider the present value of Health Glow's income stream (see above).

$$PV = \int_0^5 50000e^{-0.08t} dt = \$206054$$

Where we have used the Basic Integration Tool in the last step. This means that in order to equal the future value (\$307,020) of investing an income stream of \$50000 for a period of 5 years at 8% compounded continuously, Health Glow would have to invest a lump sum now of \$206054 for the same time period at the same rate of interest.

Health Glow's revenue stream was constant (\$50000) over the five-year period. Revenue streams need not be constant, however. Suppose Health Glow generated an increasing income stream given by $R(t) = 50,000 + 2000t$. We find the future value and present value of this income stream as follows:

$$\begin{aligned} FV &= e^{0.08(5)} \int_0^5 (50000 + 2000t)e^{-0.08t} dt = (1.49)(225287) = \$335,678 \\ PV &= \int_0^5 (50000 + 2000t)e^{-0.08t} dt = \$225,287 \end{aligned}$$

Example 17.6. Consumer Surplus

Let

$p = D(x)$	be the demand function for a commodity
\bar{p}	the established market price of the commodity
\bar{x}	the number of items sold at (i.e. the consumer demand at \bar{p})

Consumers' surplus is the difference between what consumers would be willing to pay p and the actual price \bar{p} they pay. If we plot $p = D(x)$ and the straight line $p = \bar{p}$ on the same axes, then the consumers' surplus is the area between $D(x)$ and $p = \bar{p}$, i.e. $D(x) - \bar{p}$, from 0 to \bar{x} .

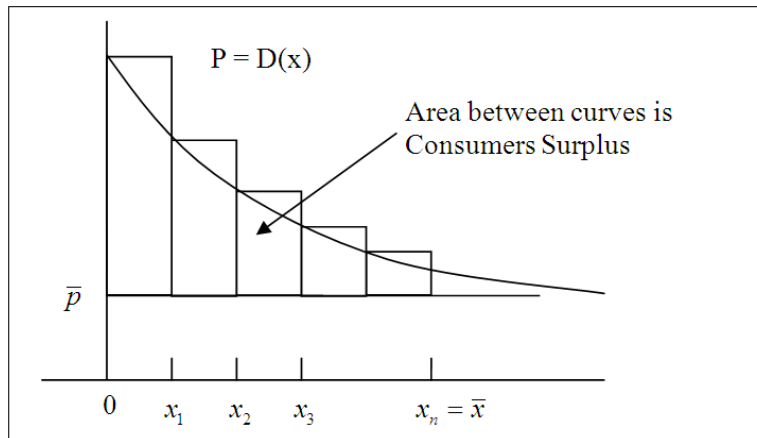


Figure 17.5: Integrating to compute the area between the demand curve $p = D(x)$ and the curve $p = \bar{p}$, the established market price of a commodity, determines the consumer surplus.

We find the approximate area between the two curves (the straight line $p = \bar{p}$ is considered to be a curve) by placing rectangles between them (see the figure below). The height of each rectangle is $D(x) - \bar{p}$ and the width is Δx . The area of the i th rectangle is $[D(x) - \bar{p}]\Delta x$.

The approximate area between the curves is the Riemann sum

$$\sum_{i=1}^n [D(x_i) - \bar{p}]\Delta x = [D(x_1) + D(x_2) + \dots + D(x_n)]\Delta x - \underbrace{[\bar{p}\Delta x + \bar{p}\Delta x + \dots + \bar{p}\Delta x]}_{\substack{n \text{ terms add to} \\ \bar{p}(n\Delta x) = \bar{p}\bar{x} \\ \text{because } n\Delta x = \bar{x}}}$$

Taking the sum as $n \rightarrow \infty$, we have

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n [D(x_i) - \bar{p}]\Delta x = \lim_{n \rightarrow \infty} \sum_{i=1}^n D(x_i)\Delta x = \int_0^{\bar{x}} D(x) dx - \bar{p}\bar{x}$$

So, $CS = \int_0^{\bar{x}} D(x) dx - \bar{p}\bar{x}$. To illustrate this concept, suppose the demand function is given by $P = D(x) = -0.01x^2 - 0.3x + 25$ where

p is the wholesale price in dollars
 x is the demand in thousands
 $\bar{p} = 10$ the established market price (in dollars)

We need to find the intersection of $p = D(x)$ and $p = 10$ in order to find the area between the two curves. See the figure below.

Using Goal Seek in Excel or `uniroot` in R, we find the demand at \$10 is $\bar{x} \approx 26$ (rounded down to the nearest whole number). Substituting in the formula above and using the Basic Integration Tool for the definite integral, we have

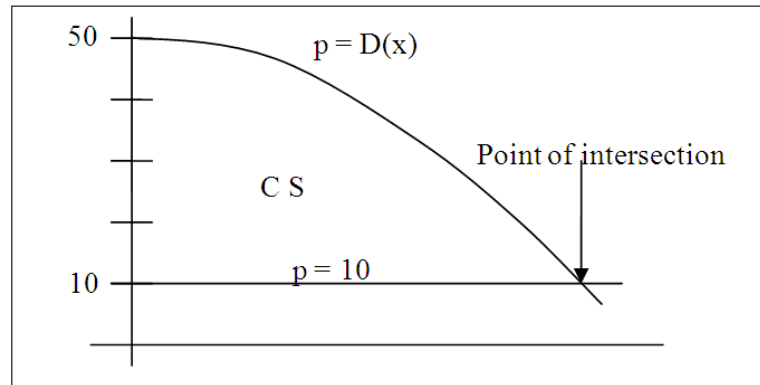


Figure 17.6: Finding the point of intersection to determine the limits of integration for computing consumer surplus.

$$\int_0^{26} (-0.01x^2 - 0.3x + 25) dx - (10)(26) = 500 - 260 = \$240 \text{ thousand}$$

Example 17.7. Producers' surplus

Let

$p = S(x)$ be the supply function for a commodity
 \bar{p} be the established market price of the commodity
 \bar{x} be the number of items producers are will to supply at \bar{p} (i.e. the consumer demand at \bar{p})

The producers' supply is the difference between what the suppliers actually receive and what they are willing to receive. The producers' surplus is the area between the line $p = \bar{p}$ and the supply curve $S(x)$.

Similar to what we did for the consumers' surplus, we find the area between the two curves to be the definite integral

$$\text{Producers' Surplus} = \bar{p}\bar{x} - \int_0^{\bar{x}} S(x) dx$$

To illustrate this concept, suppose the supply function is given by $P = S(x) = 0.1x^2 + 3x + 15$ where

p is the price in dollars
 x is the demand in thousands
 $\bar{p} = 50$ is the established market price

Before we can find the Producers' Surplus $= \bar{p}\bar{x} - \int_0^{\bar{x}} S(x) dx$, we need to find the intersection of $p = 50$ and $p = S(x)$ in order to find the area between the two curves. Using Goal

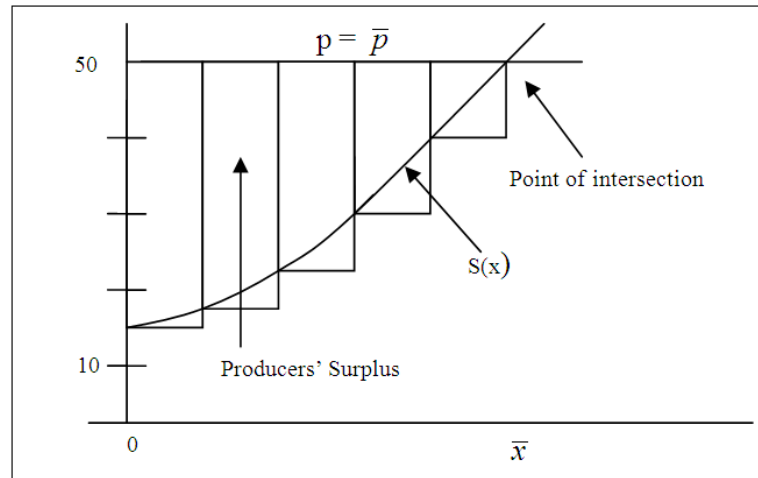


Figure 17.7: Using integration to find producer surplus, the area between the fair market price curve $p = \bar{p}$ and the supply curve $p = S(x)$.

Seek in Excel or `uniroot` in R, we find the demand at \$50 is $\bar{x} \approx 9$. Substituting into the formula above and using the Basic Integration Tool for the definite integral, we have

$$(50)(9) - \int_0^9 (0.1x^2 + 3x + 15) dx = 450 - 281 = \$169 \text{ thousand}$$

17.2.3 Exploration 17B: Consumers' and Producers' Surplus at Market Equilibrium

1. Bring up the file **Exploration 17B**. This file contains data from consumer and producer market surveys of a particular company's products. Find the equations for the demand function and the supply functions.

$$D(x) =$$

$$S(x) =$$

2. How do you find the point where market equilibrium occurs? (Hint: Use the difference of the demand and supply functions in Goal Seek as described in the How To Guide.)

$$\bar{x} =$$

$$\bar{p} =$$

3. Compute the consumers' and suppliers' surplus at market equilibrium (\bar{x}, \bar{p}) .

4. Graph CS , PS , and $p = \bar{p}$ on the same axes where \bar{p} is the equilibrium price (see the How To Guide).

5. Suppose the figure below illustrates the consumers' and producers' surplus at market equilibrium of a different commodity than above. Sketch the horizontal line $p = p_L$, where p_L is the established price of a commodity that is lower than the equilibrium price \bar{p} . What are the implications for the company in this situation?
6. The figure below illustrates the consumers' and producers' surplus at market equilibrium. Sketch the horizontal line $p = p_H$, where p_H is the established price of a commodity that is higher than the equilibrium price \bar{p} . What are the implications for the company in this situation?

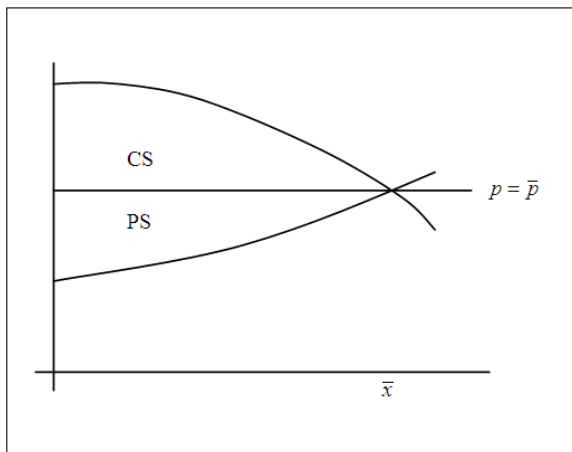


Figure 17.8: Graph for exploring what happens if the established price is lower than market equilibrium.

17.3 Homework

Mechanics and Techniques Problems

17.1. A standard normal distribution is a normal distribution whose mean is 0 and whose standard deviation is 1. The function that gives rise to the standard normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

1. Verify that the area under the standard normal distribution from $-\infty$ to $+\infty$ is 1. In the Basic Integration tool, use -50 for $-\infty$ and 50 for $+\infty$. Use 3.141593 for π .
2. Verify the Rules of Thumb (also called the “Empirical Rule”) for the standard normal distribution (see Definitions and Formulas in Chapter 4B).

17.2. A few years ago, you set up an internet business that initially brought in \$30,000 revenue and \$800 per year thereafter. At start up time, you immediately invested this income stream at 6% interest compounded continuously. You want \$1,000,000 to accrue from this investment in order to retire. After how many years will you be able to retire? Hint: Set $x_0 = 0$ in the Basic Integration Tool, replace the function in C9 with the appropriate function for this problem, and change x_N (which is T in the integral you need) until you hit your goal.

17.3. The management of Fitter Than Thou Health Spa is considering renovating its exercise room and buying new equipment. It has developed two plans. Plan 1 costs \$700,000 to renovate the room, buy the equipment and then install it. It is expected to generate an

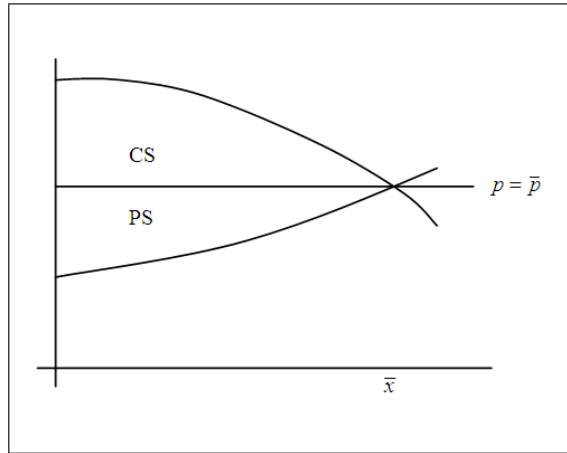


Figure 17.9: Graph for exploring what happens if the established price is higher than market equilibrium.

income stream of \$550,000 per year over the next 5 years. Plan 2 requires less initial outlay at \$250,000 but will generate an income stream of only \$470,000 per year for the next 5 years. If the interest rate is expected to hold at 8% per year for the next 5 years,

1. calculate the present value of the income stream of each plan for the 5 year period; and
2. determine which plan will generate a higher net income after the 5-year period.

17.4. The demand function for a collapsible pull-along sports carrier is

$$p = d(x) = \ln(75 - 0.005x^2)$$

where p is the unit price in hundreds of dollars and x is the quantity demanded per week. The corresponding supply function is

$$p = s(x) = \sqrt{1 + 0.03x}$$

where p is the unit price and x is the number of carriers the supplier is willing to make available at price p .

1. Find the consumers' surplus and the producers' surplus if the unit price is set at the equilibrium price.
2. Graph the consumers' and producers' surplus on the same axes when the unit price is set at the equilibrium price.

Application and Reasoning Problems

17.5. At last year's annual association fair of suppliers of Digiview camcorders, surveys were taken of how many of a new digital model the association members would be willing to supply at various prices. A supply function $S(x)$ was generated by regressing on this data. In a similar fashion, a demand function $D(x)$ was generated from surveys taken in malls across the country as to the price consumers would be willing to pay for this new model. After determining the consumers' surplus and the producers' surplus, the association recommended at that time that the digital model's market price be set at the equilibrium price. Recently, however, the association journal ran an article expressing alarm over the relatively large producers' surplus compared to the consumers' surplus. Without modifying the demand function, the article recommended that suppliers need to rethink how much they would be willing to supply of the model at various prices and that new surveys need to be taken at this year's fair in order to establish a new market price.

1. What do you think the article recommended?
2. Explain the reasoning behind the recommendation.

17.6. Suppose the demand and supply curves for a commodity are known. Discuss the state of the market for this commodity from both the consumers and producers points of view when

1. the established price is set below the equilibrium price \bar{p} , and
2. the established price is set above the equilibrium price \bar{p} .

17.7. Discuss how we can determine when market equilibrium has been reached by adding together the areas for the consumers' surplus and the producers' surplus for increasing values of x (i.e. the area between the supply and demand curves)? Hint: see figure 17.10.

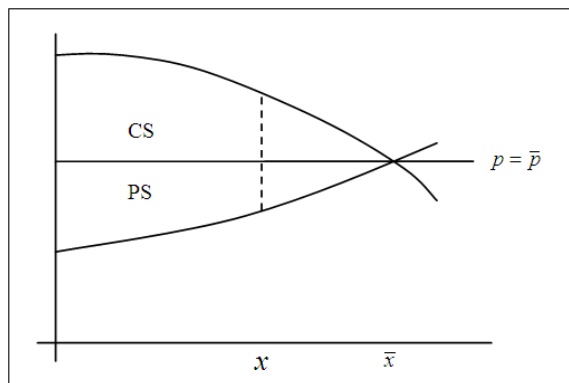


Figure 17.10: When do we reach market equilibrium?

Part VI

Appendices

APPENDIX A

Memo Problems

A.1 Carnivorous Cruise Lines

To: Analysis Staff
From: Director of Marketing
Date: May 11, 2008
Re: Salena Way RFP

I have received an RFP (Request For Proposal) from Salena Way, Director of Carnivorous Cruise Lines. Her RFP is enclosed in hard copy and also attached electronically (see page 427).

After you read and think about Ms. Way's problem, I want each of you to send me a preliminary proposal for how to deal with it. I will give you some feedback and you can resubmit your revision to me (I will post the deadlines on our intranet web site). I will then pass on your revised proposal to our marketing team, who will cost it out. I will write a cover letter and submit the final proposal to Ms. Way myself.

Our marketing team will need your proposal to include the following, so make sure you address each of them:

1. What is the perceived problem(s) and its consequences?
2. Possible reasons for the problem (The RFP suggests three possibilities. Make sure you address these and maybe consider one or two other possibilities).
3. A plan for gathering data to help identify the problem. You need to include a rough timeline for the whole data collection and analysis process.
4. Use your possible reasons and possible solutions (1 and 2 above) as a way of ensuring that your data collection gets you what you might need; that is, use these as a reality check to refine your thinking.
5. Identify any possible difficulties, problems or expenses (there will indeed be some) that might hamper your data collection and analysis process.

To: Salena Way, Director of Carnivorous Cruise Lines
 From: Director of Marketing, Oracular Consultants
 Date: May 1, 2008
 Re: RFP Regarding Entertainment Attendance

As you may be aware, cruise ship traveling has become big business. Our cruise line is now competing for customers of all age groups and socioeconomic status levels. We offer all types of cruises, from relatively inexpensive 3-4-day cruises in the Caribbean, to 12-15-day cruises in the Mediterranean, to several-month, around-the-world cruises. These have several features that attract customers, many of whom book 6 months or more in advance: (1) they offer a relaxing, everything-done-for-you way to travel, (2) they serve food that is plentiful, usually excellent, and included in the price of the cruise, (3) they stop at a number of interesting ports and offer travelers a way to see the world, and (4) they provide a wide variety of entertainment, particularly in the evening.

This last feature, the entertainment, presents a difficult problem for our ship's staff. A typical cruise might have well over a thousand customers, including elderly singles and couples, middle-aged people with or without children, and young people, often honeymooners. These different types of passengers have varied tastes in terms of their after-dinner preferences in entertainment. Some want traditional dance music, some want comedians, some want rock music, some want movies, some want to go back to their cabins and read, and so on. Obviously, our cruise entertainment director wants to provide the variety of entertainment our customers desire within a reasonable budget because satisfied customers tend to be repeat customers. The question is how to provide the right mix of entertainment.

As a part of an internal quality control study my department has been conducting, I recently took one of our 12-day cruises. The entertainment seemed to be of high quality and there was plenty of variety. A seven-piece show band played dance music nightly in the largest lounge, two other small musical combos played nightly at two smaller lounges, a pianist played nightly as a piano bar in an intimate lounge, a group of professional singers and dancers played Broadway-type shows about twice weekly, and various professional singers and comedians played occasional single-night performances. (There is also a moderately large onboard casino, but it tended to attract the same people every night and it was always closed when the ship was in port.) Although this entertainment was free to all passengers, much of it had embarrassingly low attendance. The nightly show band and musical combos, who were contracted to play nightly until midnight, often had fewer than a half dozen people in the audience, sometimes literally none. The professional singers, dancers, and comedians attracted larger audiences, but there were still plenty of empty seats. In spite of this, the cruise staff posted a weekly schedule, and they stuck to it regardless of attendance. In a short-term financial sense, it doesn't make much difference. The performers get paid the same whether anyone is in the audience or not, the passengers have already paid (indirectly) for the entertainment part of the cruise, and the only possible impact on our cruise line (in the short run) is the considerable loss of liquor sales from the lack of passengers in the entertainment lounges. The morale of the entertainers was not great; entertainers love packed houses (and so do we at Carnivorous!). Of course, as they usually argue somewhat philosophically, their hours are relatively short and they are still, after all, getting paid to see the world.

We need to get to the bottom of this. Off the top of my head, could it be that we have a problem with deadbeat passengers, or low-quality entertainment, or a mismatch between the entertainment offered and the entertainment desired? How do I go about finding out? Should we keep a strict schedule, or should we play it more by ear? We need a proposal that identifies the problem(s) and then offers a solution(s) within a reasonable time frame for a reasonable price.

(Adapted from *Data Analysis and Decision Making with Microsoft Excel* by Albright, Winston, and Zappe, Duxbury Press, New York, 1999)

A.2 Carnivorous Cruise Lines, Part 2

To: Analysis Staff
From: Marketing Director
Date: May 15, 2008
Re: RFP from Ms. Way

The marketing team wants some further details for the proposal we developed for Ms. Way regarding the possible issues with entertainment at Carnivorous Cruise Lines. So here is what I want you to do:

1. Design data collection forms that I can include in the final proposal. These forms might be questionnaires, attendance counts, sales figures, however you are proposing to go about collecting data.
2. Develop a spreadsheet for each different type of data collection form and enter some test data (maybe 15 observations; you shouldn't go overboard, but you should do enough to "show off" the range of values of your variables). Since you have more than one spreadsheet, put them all in one workbook with individual, relevant names, so that I don't overlook them.
3. Include comments below your data on the spreadsheet that
 - (a) Provide explanations for any codes you are using (where appropriate) and
 - (b) Give the units of your variables when they are not obvious.
4. All data collection forms and their accompanying spreadsheets must be incorporated into a Microsoft Word document and sent to me.
5. Just under your mockup data in the spreadsheet, list each variable and identify the type of data, e.g. nominal categorical, discrete continuous, etc.

A.3 Carnivorous Cruise Lines, Part 3

To: Carnivorous Crusie Lines Project Team
From: Director of Marketing
Date: May 14, 2008
Re: Preliminary analysis of venue attendance

As you know, we have won the contract on Carnivorous Cruise Lines and have begun data collection. The enclosed file contains attendance data for four venues over a one-week cruise. We want to analyze the data for this one cruise in order to determine the venues and days to focus our attention on for subsequent cruises. Since each venue has a different capacity, use a percent of capacity (as a decimal) to measure the attendance. I would like to see two separate analyses: One analysis that compares each venue and each night to the overall cruise data and one analysis which looks at each venue individually with respect to its own data. We need the two analyses to answer the following questions:

- Are there any venues performing so poorly (with respect to the entire data set for the cruise) that are candidates for elimination?
- Are there certain nights on which a particular venue does poorly with respect to its own performance and should be considered for possible closing on those nights?
- After deciding on which venues or nights might need to be eliminated what effects would such a change be likely to have on the other venues and nights?

Attachment: Data file C03 Venues

A.4 Matching Managers to a Company

To: Job Placement Staff
From: Project Management Director
Date: May 18, 2008
Re: Placement of Managerial Clients

Since our company does management consulting, we have two middle-management clients who have come to us looking for management positions. Each of the clients is qualified to work at the four large companies in the local region. I need you to analyze the four companies in the attached data file and make a recommendation to each client as to which company each would be better suited to. The data file contains a list of the salaries at each of the four companies. There are about the same number of managers in each company with roughly the same ratios of middle- to upper managers in each.

Each of our clients has just moved out of the lower 25% management rank in his or her previous position. They are, however, quite different. Manager A is a confident go-getter who enjoys leaving the competition behind. Manager B, on the other hand, prefers to run with the pack. He wants to do well, of course, but stability and security are important.

To get started, you might consider generating comprehensive summary statistics and side-by-side box plots for these four companies. Based on what you learn from the box plots make a recommendation of a company for each client. Be sure to provide as much evidence as possible.

Attachment: Excel data file C04 Companies

Follow-up Memo Problem

To: Job Placement Staff
From: Project Management Director
Date: May 20, 2008
Re: Follow-up on Placement of Managerial Clients

Our clients like the recommendations that we made, but they would each like an additional option. You are to make two recommendations for each client; that is, for A, you are to recommend two companies and for B you are to recommend two companies. If you haven't already, make sure you consider and discuss almost every part of the boxplot in making your recommendations. Moreover, be sure you point out the comparative advantages and disadvantages of your selected companies for each client according to that client's profile. I expect your work to be thoughtful and comprehensive.

Attachment: Data file C04 Companies

A.5 Service at Beef n' Buns

To: Beef n' Buns Store Managers
From: Chad R. Chez, Regional Manager of Beef n' Buns
Date: May 18, 2008
Re: Response to poor service

As you know, we have been analyzing the customer service data that each of you has compiled and submitted to me. What we have found from an analysis of the data is that service times seem to be dependent on two things:

1. Whether the order is placed from the breakfast menu in the morning or from the lunch menu during the day
2. Whether the order is placed at the counter or at the drive through

The attached data file contains the service times from your restaurant and has been reworked to reflect these venues (a venue is a combination of a menu type and a location for the order). We need a detailed description of the service times in each venue. Analyze the patterns of the service times in all the venues and state whether these seem reasonable or whether they might reveal possible problems giving rise to customer complaints. It will be critical to this ongoing study that you first identify the venue or venues that may require more investigation. Once you have identified these venues you need to propose possible reasons for the problems that will lead us to further data collection. What additional data do we need in order to get at the underlying problems with customer service?

Attachment: Data file C05 BeefNBuns 2

A.6 Portfolio Analysis

To: Financial Planning Services Department
 From: John E. Cash
 Date: May 18, 2008
 Re: Portfolio development for clients

As you know, one of the many services offered by Oracular Consulting is financial planning. We have recently acquired two new clients. The financial planning department has pre-filtered the current market and provided a list of ten (10) stocks that have been performing well in recent months. The data on these stocks for the last three years is provided in the attached file, presented as frequency data based on the number of days the stock provided a particular daily return.

The two clients are quite different. Client A is young and energetic. She has a long time before retirement, and is willing to take risks in order to gain a lot. Client B is older and has much less time before retirement. He needs a stock portfolio that will provide income in his retirement, so he is not willing to accept a lot of risk, but he of course would like a steady return. Both clients recognize the need to diversify their portfolio in order to plan for the future.

Using the frequency data provided, fill out a chart like the one below to help present the data to the clients in a convenient, easy-to-compare format. Then put together two portfolios, one for each client, composed of 4 stocks, showing the percentage of the investment in each of the stocks in the portfolio. Justify your choices carefully, and provide the clients with both an estimated rate of return for their portfolio and a range of possible returns that they can reasonably expect. To estimate the range of likely returns, simply use the high and low expected values for each stock in the portfolio, weighted by the percentage of the investment in that stock.

	Stock 1	Stock 2	Stock 3
Mean			
Std. deviation			
Minimum			
Q1			
Median			
Q3			
Maximum			

Attachment: Data file C06 StockPerformance

A.7 Truck Maintenance Analysis

To: Analysis Staff
From: Project Management Director
Date: May 27, 2008
Re: Truck maintenance data

Our services have been retained by Metro Area Trucking to analyze the records they have maintained on the trucks in their fleet. The company has locations around the Rochester area, some inside the city limits and some outside the city limits. The director of operations, Ms. Mini V. Driver, at the company has asked that we determine how the different locations affect the maintenance costs on the trucks.

She has provided data on each of the trucks in the fleet. The data includes information on last year's maintenance expenses for the truck, the mileage of the truck, the age, the type of truck, and whether it is based at one of the in-city or out-of-city locations. As a first look at the data, you should separate the data into trucks that are based in the city and trucks outside the city. Mini Driver suspects that mileage and age are the most important factors, so use everything at your disposal to explain how these two quantities affect the maintenance costs. I need a full report, including graphs, tables, and formulas, as well as an analysis and explanation of what each piece of information means.

Attachment: Data file C07 TruckData

A.8 Commuter Rail Analysis

To: Analysis Staff
From: Project Management Director
Date: May 27, 2008
Re: Commuter Rail Analysis

Ms. Carrie Allover, the manager of the commuter rail transportation system of our fair city has contracted us to analyze how various factors affect the number of riders who use the rail system. Her Supervisory Board wants this information for long-range planning. Accordingly, she has sent along some data on the weekly ridership (number of people who use the train during a week) of commuters taking the train into the city, as well as some data on various factors thought to have an influence on the ridership. These data contain the following variables: Weekly riders, Price per ride, Population, Income, and Parking rate. The latter variable, Parking rate, refers to the cost of downtown parking.

To deal with Ms. Allover's requests, you will have to build several regression models with Weekly riders as the response variable, but before you proceed with this I want some common sense predictions on whether the coefficients of each of these explanatory variables will have a positive or negative sign; that is, whether the variable will have a positive or negative effect on the weekly ridership. Of course, you have to provide an explanation for your prediction. Some of these will be clear cut, but there may be a couple that are not so easy to predict and you won't know the answer until you actually run the model. But don't change your analysis if you prove to be wrong. Ms. Allover needs this kind of verbal, up-front analysis (whether right or wrong) so that she will be prepared to deal with possible responses, as well as misunderstandings, on the part of the Board.

After you have explained how you anticipate each variable will affect the number of weekly riders, go ahead and formulate the different models, one for each possible explanatory variable. Explain what each of the models means, using the coefficients in the regression output. In particular, describe how each explanatory variable actually affects the response variable, Weekly riders, including all appropriate units. This is extremely valuable information, Ms. Allover insists. Let's provide her with a brief analysis of how well the models fit the data as well as how accurate we can anticipate the predictions of the models will be.

Attachment: Data file C08 Rail System

A.9 Gender Discrimination

To: Analysis Staff
From: Director Project Management Director
Date: May 27, 2008
Re: Gender Discrimination at EnPact

EnPact, a company which performs environmental impact studies, is a medium-sized company. Currently, they are being audited by the Equal Opportunity Employment Agency for possible gender discrimination. Our firm has been brought in to conduct a preliminary analysis. A database of employee information is available in the attachment below. These data include employee salaries, genders, education, job level, experience, and age.

First, I want you to construct a full regression model for these data. Next, you should work toward the best possible model by dropping insignificant variables, one at a time according to the following rules:

1. Always drop the least significant variables first because this may change the significance of the remaining explanatory variables.
2. If you decide to drop a category of a categorical variable from the model, you must drop all the other categories of that categorical variable as well. This is an all-or-nothing proposition for categorical variables at this stage of our analysis.
3. Only drop a single numerical variable or a group of related dummy variables at each stage of the model-building process.
4. Any variables whose significance is questionable (that are close to the border, $p = 0.05$) should be kept, but noted for further investigation in your report.
5. Furthermore, you may detect outliers in the residual plots. At this stage of our analysis, do not delete them; further investigations may determine that these should be kept in the data. However, notes should be made in your report to identify any outliers.

Your final report on these data must discuss what your model tells you about the significant influences on the salaries at EnPact and should explain how gender might be implicated in the salary structure.

Attachment: Data file C09 EnPact

A.10 Truck Maintenance Expenses, Part 2

To: Analysis Staff
From: Project Management Director
Date: May 27, 2008
Re: New Truck Contract

As you know, we have been doing some work for Ms. Mini Driver, the Director of Operations at MetroArea Trucking, on how location affects the maintenance expenses for the trucks in the fleet. We have received an additional contract to further analyze the fleet's maintenance expenses. Ms. Mini Driver would like us to analyze the entire truck data set (see attachment), which includes last year's maintenance expense, the mileage, age, and type of truck, as well as the location (based either in city or out of city) of where the truck is based. Ms. Mini Driver wants us to provide her with an analysis of what factors affect maintenance expenses and how much each affects the expenses.

I'd like you to develop your own optimal regression model by choosing your own variables and going through your own model-refining process before seeing what a stepwise regression routine produces for an optimal model. This process should give you a better feel for how the variables contribute to the maintenance expense, which should be helpful when you interpret your models.

1. Start with a full model without any interaction terms and record your findings in the chart below. I would like you to begin this way because there are situations when interaction terms aren't really worth their trouble, whereas in others they are.
2. Run the reduced model with the significant variables that you get from the full model, again without any interaction terms. Record your findings in the chart.
3. Start over with a full model with all interaction terms. Record your findings.
4. Run a reduced model with the significant variables only. Record your findings.
5. Now run a full model with all interaction terms using a stepwise regression routine. Record your findings.
6. Write a memo to me stating what you think the model should be and why, including a description of how you went about finding your model. Be sure to include your supporting evidence (you will find the chart helpful here). Comment on the quality of your model and then interpret your model, explaining which variables significantly affect maintenance expenses and how much each affects the expenses.

Attachment: Data file C10 Truck

Model	R^2	Adj R^2	S_e	List of significant variables
Full Model With no Interactions				
Reduced Model with no interactions				
Full Model with all interactions				
Reduced Model with significant interactions				
Stepwise regression				

A.11 DataCon Contract

To: Analysis Staff
 From: Project Management Director
 Date: May 28, 2008
 Re: DataCon Contract

We have received a contract from DataCon, a large data analysis provider that does general data analysis and management contracting for a wide variety of manufacturing and service sector businesses. They have subcontracted some of their business to us. They want us to fit some predictive models for four sets of data they have sent along. They want to see a best-fit nonlinear trendline for each data set, as well as the best model that we come up with, superimposed on both the scatterplot of the data and the best-fit trendline. DataCon management wants not only simple trendlines but also good fitting models constructed from shifting and scaling the basic functions because models built from basic functions are more transparent and easier to analyze than typical trendline models.

As usual, direct your memo to me. Include the following:

- A brief introduction
- Complete information about each model, including what shifting and scaling you included, how you found optimal values for these, what the final parameter values for the best fit were, and the final equation of the model
- Graphical representation of the typical trendline and the best model on the same graph
- Correctly computed values for R^2 for the best models and a description of how well they seem fit as compared to the automatic trendlines
- A few summary comments, including any special considerations you want to pass along about what you found.

Attachment: Data file C11 DataCon Data

Here are some suggestions for dealing with this assignment:

1. Start by fitting the best built-in trendline for your software (don't forget to record its equation and its R^2) to a scatterplot of the data set. The table below (or one like it) will help organize the information.
2. Now try fitting your own shifted and scaled basic function on top of the scatterplot and the best-fit trendline, comparing your computed R^2 to the R^2 of the trendline.
3. You might not be able to construct a better model in every case, but get as close as you reasonably can. That's all DataCon really wants or needs.

4. Here are a couple of tips:

- (a) Don't even try to do your own fit for a polynomial function (used when the scatter plot has a turn(s), etc) because built-in routines for a polynomial fit are already clear and understandable. Your job, in this case, is to find that polynomial.
- (b) If you are fitting your own exponential function, don't bother with horizontal shifts because mathematically such shifts can be absorbed by the scaling parameter.
- (c) If the best trendline is a power function with a fractional power, for example, $x^{0.42}$, you might suggest using $x^{0.5}$ for your own power function because $x^{0.5} = x^{1/2} = \sqrt{x}$, which is much easier to understand (remember, this is what DataCon wants).

DATA SET 1		EQUATION	R^2
	My Best Fit		
	Best Trendline Fit		
DATA SET 2			
	My Best Fit		
	Best Trendline Fit		
DATA SET 3			
	My Best Fit		
	Best Trendline Fit		
DATA SET 4			
	My Best Fit		
	Best Trendline Fit		

A.12 Insurance Costs

To: Analysis Staff
 From: Top Modeler
 Date: May 28, 2008
 Re: Operating Costs for Insurance Company

Our clients' management team would like us to compare a straight-forward linear model with the multiplicative model that we came up with for our original submission. They want to know if there is anything to be gained from their basing their management decisions on the more complicated multiplicative model. Or is a linear model almost as good? As we all know, simpler is better. But if there is indeed something to be gained from using the more complicated multiplicative model then we should point out exactly what it is. Otherwise, we should recommend that they use the simpler linear model.

Actually, this request should enable us to sharpen our analysis considerably. For example, we can now compare the R^2 and S_e that we calculated for our multiplicative model to the R^2 and S_e generated by the linear model (we don't have to calculate these latter ourselves, however, since they are valid for linear models). Also, we can compare the fitted vs. observed graphs and the residual vs. fits graphs of the two models to see if we can detect a difference in goodness of fit or accuracy.

Attachment: Data file C12 Insurance

Here's how you might go about dealing with this assignment:

1. Run a linear regression model, along with the two diagnostic graphs (fits, residuals).
2. Compute the cost predicted by the linear model with 100 home and 2000 auto policies.
3. Do a **marginal cost analysis** for the linear model (if one more home policy is sold, then the cost will increase by what dollar amount, holding the number of auto policies at 2000; do a similar thing for auto policies).
4. Run your multiplicative model, generate your two diagnostic graphs and calculate your own R^2 and S_e .
5. Compute the cost predicted by the multiplicative model at the 100 and 2000 levels.
6. Do a parameter cost analysis for the multiplicative model (if the number of home policies increases by 1%, then the cost will increase by what %, holding the number of auto policies at the current level; do this for levels of 100 home and 2000 auto policies, then do the similar thing to analyze how costs change if the number of auto policies changes).

7. Do a nice summary presentation and analysis for your two models, including side-by-side graphs and maybe a table or two showing R^2 , S_e , the costs predicted by the two models at the 100 and 2000 levels, and your marginal and parameter change analysis - lay it all out for the client.

8. Make a summary statement as to which model you recommend for our client and why.

A.13 Revenue Projections

To: Analysis Staff
 From: Project Management Director
 Date: May 29, 2008
 Re: Revenue Projections at Dream Grills

One of our smaller clients, Dream Grills, sells its one product, the Dream Grill 5000, in two forms: assembled and unassembled. Based on economics theories about substitute commodities, they have been making projections and analyses for their business plan based on the following models of their revenue.

$$\begin{aligned}
 R(Q_A, Q_U) &= Q_A P_A + Q_U P_U \\
 P_A &= 462 - 0.1Q_U - 0.35Q_A \\
 P_U &= 372 - 0.20Q_A - 0.16Q_U
 \end{aligned}$$

In these models, the P and Q refer to the price and the quantity of the two items; the subscripts A and the U refer to the “assembled” and “unassembled” versions of the product. Thus, the quantity P_A is the price of the assembled grills, based on the quantities of each version that are sold.

The company has collected revenue and quantity sales data for the last 50 weeks. Formulate a regression model for the revenue and compare the two models, yours and theirs, using graphical and analytical tools you feel are appropriate to illustrate the differences.

Attachments: Data File C13 Revenue

A.14 Profit Analysis

To: Analysis Staff
From: Project Director
Date: May 29, 2008
Re: Profit analysis for MacroSoft Software Company

A small, but up-and-coming software firm called MacroSoft has contacted us concerning a new software package they have developed. The CEO of the company, Bob Doors, has asked us to analyze three different production scenarios and to report on the findings. For each of the scenarios, he wants us to assume that the average cost of producing q million copies of the software is given by the function (with $q > 0$): $\bar{C}(q) = 0.01q^2 - 0.6q + 10$.

The units of this average cost function are in millions of dollars per million copies. Further, he expects that users will pay \$9.95 per copy of the software. Each of the three scenarios is described below. Mr. Doors has asked that the report contain both analytical calculations and spreadsheet calculations to verify these.

- Scenario A. In this production scenario, the company needs to know how many copies to produce (and then sell) in order to minimize the average cost for producing each copy of the software.
- Scenario B. In this production scenario, the company needs to know the total number of copies that it should produce (and sell) in order to minimize the total cost for producing the entire quantity of software.
- Scenario C. In this scenario, cost is no object. The company is interested in maximizing the profit earned from manufacturing and selling the software, no matter how many copies it takes to do it and regardless of the costs involved.

Your final report should include advice for manufacturing under each scenario and an overall comparison of each scenario, including: average cost, total cost, revenue, and profits. These should be in a nice table, and should be clearly explained for Mr. Doors - I know him, and he doesn't read anything that isn't fully explained and absolutely clear. Further, he would like a final recommendation on which of the scenarios his company should follow at the present. Again, keep in mind that this is a start-up company with limited production capacity.

Attachment: No attachment - you should create your own file to analyze this problem.

A.15 Loan Analysis

To: Analysis Staff
From: Cassandra Nostradamus, CEO
Date: May 30, 2008
Re: Loan options

Oracular Consulting is planning to purchase \$1,000,000 in computer equipment and software to upgrade the main server and our web presence. Since we do not want to reduce our liquid assets by this amount, we are considering several different loan possibilities. The terms of these loans are described below.

	Loan A	Loan B	Loan C	Loan D
APR	6%	5%	3%	2%
Number of Years	2	3	5	10
Payments per year	12	12	4	4

Analyze the four loans and provide a well-reasoned recommendation as to which loan (or loans) would be the best choice. It would certainly be nice to choose a loan that we can pay off as quickly as possible, but that may require very high monthly payments. If we are willing to pay large monthly payments, then we can take a short term for the loan, but if we need to lower the payments, we need to make a decision on some other characteristics. The three obvious ones are to compare the length of the loan, the total interest paid over the lifetime of the loan, or the monthly equivalent payments for the loan (the amount we pay each period, pro-rated to a monthly budget amount).

Attachments: None - create your own to display the results

A.16 Advertising Costs

To: Analysis Staff
From: Advertising Director
Date: May 30, 2008
Re: Advertising for Oracular

Oracular Consulting wants to increase its exposure to the public in order to increase business. We are planning to split our advertising budget among three venues:

- Full-page ads in the Albatross Airlines magazine that every executive will see when flying
- Web-based ads in Business Breaks Online which most executives subscribe to
- Print ads in the Sunday business section of a large national chain of newspapers

We have a maximum of \$50,000 to spend on advertising per month. Ads in the airline magazine cost \$25,000 for a one page ad each month. The magazine is seen by about 600,000 people each month. The newspaper ads cost \$210 per square inch for an ad that runs in all four Sunday issues in one month. The smallest ad available is four square inches; the largest is a half-page ad (115 square inches). The newspaper has a circulation of 900,000. Market research has shown that the number of people who pay enough attention to the ad to count as having seen it is dependent on the number of square inches of the ad according to the formula (where “units” stands for the number of units, in square inches, of the ad):

$$\text{Exposure} = -68.05 \cdot \text{Units}^2 + 14504.6 \cdot \text{Units}$$

The web ads are randomly generated whenever someone visits the web site. The price we pay is based on the priority our ad is assigned. For every 0.1% chance of our ad being shown, we pay \$60. The website will not allow us to purchase more than 25% of the ad space. The web site gets an average of 800,000 hits per month. Obviously, we want Oracular to be seen by as many people as possible. Keep in mind, though, that many of the circulation numbers above are approximate, so give us some idea of what range of results we can expect if these numbers change. Also, you should be aware that Albatross Airlines has been having some difficulties lately, and the number of passengers is expected to drop; they will probably raise prices for ads to compensate a little. We need to know at what point we should stop advertising in their magazine. Finally, the monthly advertising budget fluctuates as much as \$10,000 per month; how will that effect our decisions about what ads to place?

Attachments: None

A.17 Pricing Dispute

To: Analysis Staff
From: Project Management Director
Date: May 30, 2008
Re: Pricing Dispute

Ted Bair, one of the managers of Cool Toys for Tots, has requested our help in resolving a pricing conflict within the company over a new digital doll it wants to market. Ted's group has spent valuable resources in gathering data from consumer and producer market surveys in order to help establish a rational price for this doll. Other managers, however, have a gut feeling based on their many years of experience in the business that the selling price should to be \$55? per doll. Ted would like to determine the price based on research data, whatever price that may turn out to be. His group could do this for themselves but he thinks it would be better if someone outside the company did the analysis and made a recommendation. So here's what I think we should do:

1. Based on the survey data (see the attachment), find the demand and supply functions.
2. Calculate the consumers' and producers' surplus for the equilibrium point.
3. Determine the consumers' and producers' surplus for the company based on the intended pricing of \$55 per doll and the demand at this price.
4. Present graphs of the consumers' and producers' surplus from 2) and 3)
5. Make a recommendation as to how the company should set its price \bar{p} and what the demand \bar{x} would be at this price.

If it's possible to make a compromise or "diplomatic" recommendation one way or the other, I'm sure Ted would appreciate it and pass along more business our way.

Attachments: Data File C17 Pricing

APPENDIX B

Sample Rubric for Evaluating Memo 7

The rubric below provides a sample of how instructors can easily use a checklist approach to grading the memo assignments in this text. Bascially, each memo has three categories in which students should demonstrate excellence: Mechanics and Techniques, Application and Reasoning, and Communication and Professionalism. These are discussed in general terms in the preface. For each category, there are items listed in a checklist format at two levels of accomplishment: Expected and Impressive. To meet the minimum requirements for a memo, students should have the expected items checked off as being present in the memo in a clear and easily understandable way. Then, for each category in which the student's work is impressive, the grade is bumped up.

For example, one could define the following grade scale for memo problems, where the entries define the number of categories (out of 3) that must be at that level. Then, the intermediate grades can be awarded for partial success in a category.

Grade	Expected	Impressive
D	3	0
C	2	1
B	1	2
A	0	3

	Expected Level	Impressive Level
M & T: 0 E- E E+ I- I	<ul style="list-style-type: none"> <input type="radio"/> File was correctly titled for Bb <input type="radio"/> File name correct <input type="radio"/> Contains a correlation matrix <input type="radio"/> Contains 4 scatterplots <p>A Table of Results contains</p> <ul style="list-style-type: none"> <input type="radio"/> 4 correct regression equations <input type="radio"/> A correct R^2 for each equation <input type="radio"/> A correct S_e for each equation <input type="radio"/> Correctly states which relationship has the strongest positive or negative correlation 	<ul style="list-style-type: none"> <input type="radio"/> The 4 regression equations are correctly ranked according to best-fit <input type="radio"/> Graphical analysis indicates how much confidence we have that we can model this relationship with a linear trendline <p>For Revision Only</p> <ul style="list-style-type: none"> <input type="radio"/> Errors made in original are adequately corrected
A & R: 0 E- E E+ I- I	<p>A reasonable preliminary prediction is made for the effect on Weekly Riders of</p> <ul style="list-style-type: none"> <input type="radio"/> Price per Ride <input type="radio"/> Population <input type="radio"/> Income <input type="radio"/> Parking Rate <p>Correctly compares preliminary predictions with correlation matrix results for effects on Weekly Riders of</p> <ul style="list-style-type: none"> <input type="radio"/> Price per Ride <input type="radio"/> Population <input type="radio"/> Income <input type="radio"/> Parking Rate <input type="radio"/> Slopes and y-intercepts are stated in terms of the problem context (not Xs or Ys) <input type="radio"/> Correct units are given for slopes <input type="radio"/> Correct units are given for Y-intercepts <input type="radio"/> Correct units are given for S_e 	<ul style="list-style-type: none"> <input type="radio"/> Provides an adequate and correct explanation of R^2 <input type="radio"/> Explains what R^2 tells us about the model <p>Correct R^2 analysis is used to rank Regression Model of Weekly Riders vs.</p> <ul style="list-style-type: none"> <input type="radio"/> Price per Ride <input type="radio"/> Population <input type="radio"/> Income <input type="radio"/> Parking Rate <input type="radio"/> Provides an adequate and correct explanation of S_e <input type="radio"/> Explains what S_e tells us about the model <p>Correct S_e analysis is used to rank Regression Model of Weekly Riders vs.</p> <ul style="list-style-type: none"> <input type="radio"/> Price per Ride <input type="radio"/> Population <input type="radio"/> Income <input type="radio"/> Parking Rate <p>Correctly interprets the slope of the model for</p> <ul style="list-style-type: none"> <input type="radio"/> Price per Ride <input type="radio"/> Population <input type="radio"/> Income <input type="radio"/> Parking Rate <p>Correctly interprets the y-intercept of the model for</p> <ul style="list-style-type: none"> <input type="radio"/> Price per Ride <input type="radio"/> Population <input type="radio"/> Income <input type="radio"/> Parking Rate
C & P: 0 E- E E+ I- I	<ul style="list-style-type: none"> <input type="radio"/> Assignment was submitted on time <input type="radio"/> Submitted in memo form <input type="radio"/> The writing is competent (grammar, spelling are basically correct) <input type="radio"/> There is an adequate introduction to the problem situation <input type="radio"/> The introduction clues the reader as to what to expect in the memo <input type="radio"/> The presentation of the proposal is adequate and complete (must include everything the memo requires) <input type="radio"/> Charts are not fragmented <input type="radio"/> All axes and text on graphs are readable. <input type="radio"/> All charts are legible <input type="radio"/> All parts of memo are adequately addressed <input type="radio"/> Supporting computer output is embedded in the memo 	<ul style="list-style-type: none"> <input type="radio"/> The writing adequately deals with the complexity and depth of the analysis <input type="radio"/> Text and graphics are well integrated in a way that facilitates the readers understanding <input type="radio"/> Memo includes a conclusion summarizing the results of the analysis (executive summary) <input type="radio"/> Conclusion states which model is the best-fit model <input type="radio"/> Conclusion states how accurate we can anticipate the predictions of the best-fit model will be. <input type="radio"/> Overall, the graphs, charts, and text have a professional appearance.

- 5W+H, 24
- Antiderivative, 405
- Archival Data, 24
- Area Under Curve, 405
- Average Cost, 350
- Calculus, Fundamental Theorem, 405
- Charts
 - Axis, 167
 - Box-and-whisker Plot, 100
 - Boxplot, 100
 - Histogram, 120
 - Scatterplot, 167
 - Side-by-side boxplot, 100
 - Surface Plot, 315
- Coefficient, 190
- Concavity, 251
- Confidence Interval, 230
- Constant, 190
- Consultant, 15
- Consumers' Surplus, 410
- Critical Point, 346
- Cross-sectioning Data, 81
- Data, 15
 - Binning, 51
 - Bins, 120
 - Categorical, 41
 - Nominal, 41
 - Ordinal, 41
 - Code book, 41
 - Coding, 50
 - Column, 49
 - Computed Variable, 50
 - Cross-sectional, 50
 - Experimental Unit, 51
 - Factor, 41
 - Field, 49
 - Identifier, 49
 - Normal, 114
 - Numerical, 40
 - Continuous, 41
 - Discrete, 40
 - Outlier, 71
 - Qualitative, 41
 - Quantitative, 41
 - Raw Field, 50
 - Record, 49
 - Row, 49
 - Time Series, 50
 - Variable, 49
- Data Mining, 81
- Database
 - Field, 81
 - Record, 82
- Degrees of Freedom, 71
- Demand Function, 410
- Derivative, 338
 - Chain Rule, 359
 - Gradient, 387
 - Product Rule, 359
 - Quotient Rule, 359
 - Sum Rule, 347
- Difference Quotient, 337

- Dimensions, 314
- Discriminant, 315
- Distribution, 120
 - Bimodal, 121
 - Cumulative, 148
 - Negatively Skewed, 121
 - Positively Skewed, 121
 - Symmetric, 121
 - Uniform, 121
- Dynamic Programming, 387
- Elasticity, 292
- Email Protocol, 8
- Equilibrium Price, 411
- Exact Multicollinearity, 220
- Excel
 - Solver, 387
 - Sumproduct, 138
- Exponents, Properties of, 292
- Extrema, 347
- Factor, 305
- Factoring, 305
- Fitted Values, 195
- Frequency Table, 120
- Function, 179
 - Basic, 252
 - Cobb-Douglas, 282
 - Composition, 359
 - Exponential, 252
 - Inverse, 292
 - Linear, 252
 - Logarithmic, 252
 - Multiple Linear, 210
 - Multiplicative, 282
 - Polynomial, 265
 - Power, 264
 - Quadratic, 265
 - Quadratic, Multiple, 305, 315
 - Reciprocal, 254
 - Square, 253
 - Square Root, 253
- Future Value, 411
- Goodness of fit, 82
- Graphs
 - Horizontal Shift, 266
 - Translation, 266
 - Vertical Scaling, 266
 - Vertical Shift, 266
- Income Stream, 411
- Inequality, 377
- Integral
 - Constant, 405
 - Definite, 405
 - Indefinite, 405
 - Limits of Integration, 405
 - Numerical, 405
- Interaction Terms, 236
- Interest
 - Compound, 366
 - Continuously Compounded, 367
 - Simple, 366
- Interview, 24
- Joint-interaction, 305
- Lagrange Multipliers, 387
- Level-dependent, 251
- Linear Equations, 178
- Linear Programming, 387
- Loaded words, 24
- Logarithms, Properties of, 292
- Marginal Analysis, 292, 338
- Marginal Cost, 338
- Marginal Profit, 338
- Marginal Revenue, 338
- Market Equilibrium, 411
- Maximum
 - Global, 347
 - Local, 347
- Minimum
 - Global, 347
 - Local, 347
- Model, 69
 - Empirical, 69
 - Linear, 190
- Non-constant Variance, 283
- Non-proportional, 251

- Observation, 24
- Optimization, 347
 - Constraint, 377
 - Explicit, 378
 - Implicit, 378
 - Feasibility Region, 386
 - Feasibility Solution, 387
 - Maximize, 378
 - Minimize, 378
 - Objective Function, 378
 - Variables, 378
- Parameter, 69
- Parameter Analysis, 292
- Parameters, 264
- Parsimony, 230
- Percent Change, 291
- Pivot Table, 81
 - Column Field, 81
 - Data Field, 82
 - Row Field, 82
- Population, 50
- Present Value, 411
- Principal, 366
- Problem
 - Cause, 15
 - Chain of cause and effect, 16
 - Communicative Context, 24
 - Effect, 15
 - Long term, 15
 - Short term, 15
 - Fishbone diagram, 25
 - Law of unintended consequences, 16
 - Perceived, 15
 - Problem Context, 24
 - Situation, 15
- Producer's Surplus, 410
- Proportional, 190
- Quadrants, 167
- Quotient, 337
- Rate of Change, 291
- Raw Data, 82
- Reference Category, 220
- Regression
 - R^2 , 196
 - S_e , 197
 - Adjusted R^2 , 211
 - Degrees of Freedom, 197, 210
 - Diagnostics
 - Fitted vs. Actual, 197
 - Residuals vs. Fitted, 197
 - Full Model, 212
 - Identity, 196
 - Multiple, 210
 - Multiple R^2 , 211
 - Multiple R, 211
 - Observed Values, 196
 - P-value, 230
 - Predicted Values, 195
 - Residuals, 196
 - Simple, 190
 - Stepwise, 210
 - Variables
 - Insignificant, 230
 - Significant, 230
 - Variation
 - Explained, 196
 - Total, 196
 - Unexplained, 196
- Regression Best-fit Model, 212
- Relationship
 - Linear, 179
- Relationships
 - Direct, 168
 - Indirect, 168
 - Negative, 168
 - Positive, 168
 - Strong, 170
 - Weak, 170
- RFP, 25
- Rules of Thumb, 114
- Sample, 50
- Second Derivative, 338
- Second Derivative Test, 347
- Self-interaction, 305
- Sensitivity Analysis, 387
- Sigma, 70
- Slope, 178

- SSE, 196
- SSR, 196
- SST, 196
- Statistic, 69
 - Average, 70, 94
 - Central Tendency, 69
 - Correlation, 168
 - Correlation Matrix, 169
 - Deviation, 70
 - Estimated Mean, 137
 - Estimated Standard Deviation, 137
 - Interquartile Range (IQR), 100
 - Maximum, 100
 - Mean, 70
 - Median, 94
 - Minimum, 100
 - Mode, 95
 - Outliers, 100
 - Percentiles, 147
 - Q1, 100
 - Q3, 100
 - Quartiles, 100
 - Range, 100
 - Skewness, 120
 - Standard Deviation, 70
 - Standard Score, 113
 - Total Variation, 70
 - Trimmed mean, 95
 - Weight, 137
 - Weighted Average, 137
 - Z-score, 113
- Sum, 70
- Summary Data, 137
- Supply Function, 410
- Survey, 24
- Term, 305
- Timeline, 24
- Total Change, 291
- Trendline, 179
- Units, 291
- Variables
 - Base, 236, 305
 - Controlling, 210
 - Dependent, 167
 - Dummy Variables, 220
 - Explanatory, 190
 - Independent, 168
 - Indicator, 220
 - Interaction Variables, 235, 305
 - Response, 190
- Writing
 - Memos, 6
 - Reports, 6
- Y-intercept, 178