Chapter 12

Modeling with Nonlinear Data[1]

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:

1. 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.
2. Next we perform the regression.
3. In some cases, we'll need to compute the summary measures (R^2 and S_E) by hand.
4. 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.

- *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 using StatPro
  - √ Accurately compute R^2 and S_E for nonlinear models containing log(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.

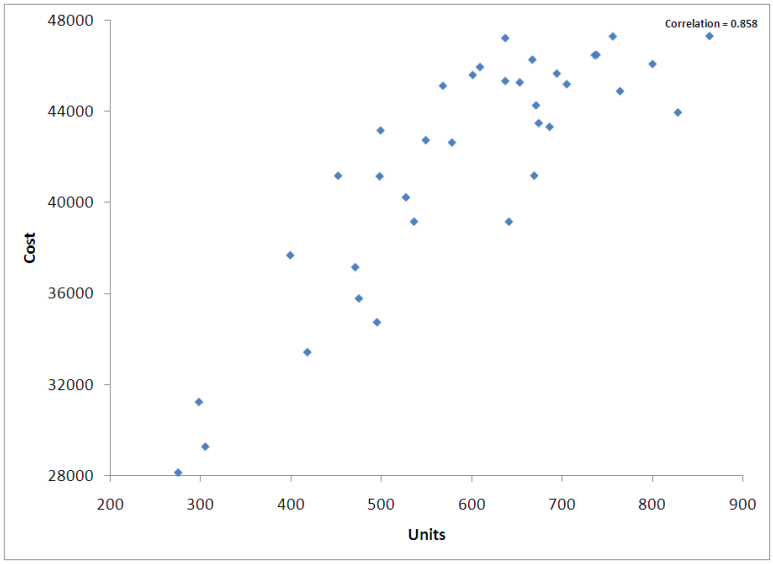

Figure 12.1: Graph of electricity cost vs. units of electricity.

Since the relationship indicates a logarithmic relationship, we examine, in figure 12.2, a graph of Cost vs. Log(Unit). Notice that this graph is "straighter", indicating that we could use linear regression to predict Cost as a function of Log(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 (page 354).

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

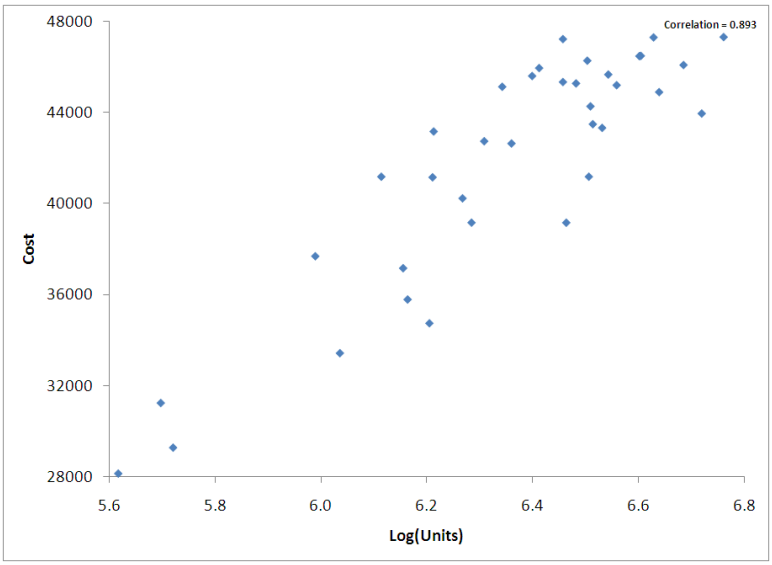

Figure 12.2: Graph of the electricity cost vs. the logarithm of electricity units used. Notice that this relationship is more linear than the one in figure 12.1 (the correlation is higher.) In a sense, we have "straightened" the data by taking the logarithm of the explanatory variable.

transformations, how to use these in the regression routines of StatPro, 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.xls) 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 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)". StatPro allows you to do this automatically through the "Data Utilities/ Transform Variables" function (see the How To Guide for details.) 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	745724176	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:

$$\text{Cost} = 6{,}772.56 + 1{,}448.74*\text{Sqrt(units)}.$$

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:

$$\text{Cost} = -63{,}993.30 + 16{,}653.55*\text{Log(Units)}.$$

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.xls. 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

$$\text{Log(Cost)} = A + B*\text{Log(Units)}.$$

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 Log(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.xls. 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.

$$P = -2 + 0.8723L + 0.1687K$$
$$R^2 = 0.9409$$
$$S_e = 11.1293$$

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 in StatPro. This produces three new variables - $\log(P)$, $\log(K)$ and $\log(L)$. We then 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.

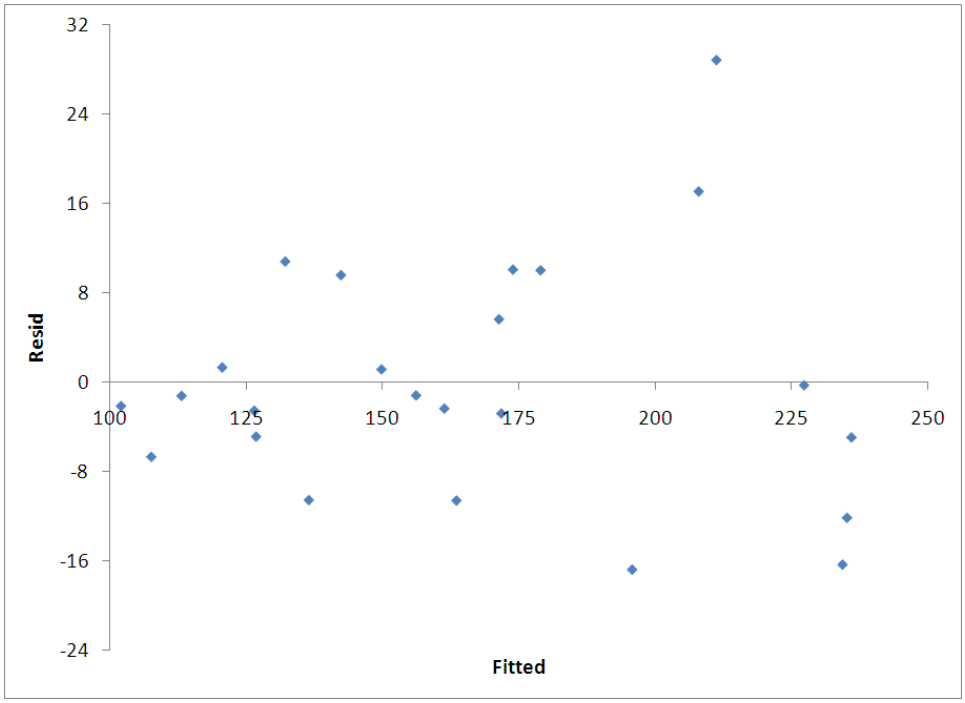

Figure 12.3: Plot of residuals versus fitted values for a linear model of predicting production vs. labor and cost.

$$\log(P) = -0.0692 + 0.7689\log(L) + 0.2471\log(K)$$
$$R^2 = 0.9386$$
$$S_e = 11.3449$$

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.xls 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.

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

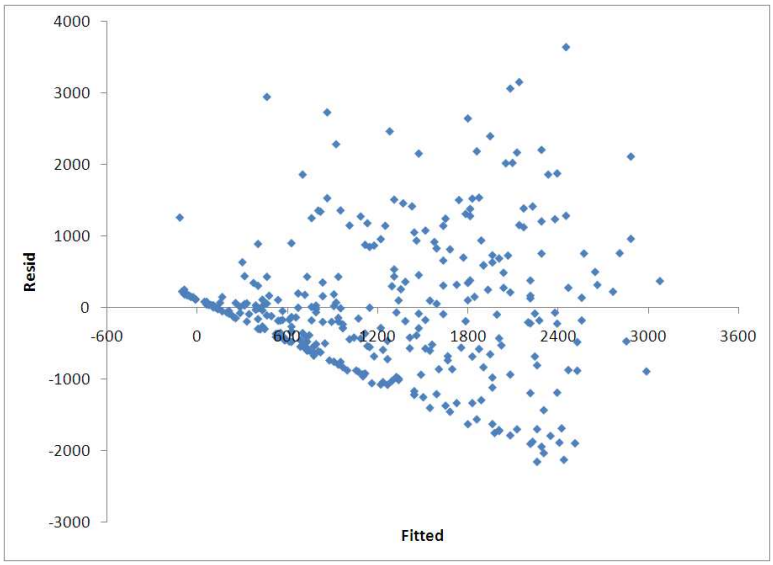

Figure 12.4: Plot of the residuals versus the fitted values when Salary is regressed against Hits.

the model equation log(Salary) = 5.1305 + 0.0151*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.

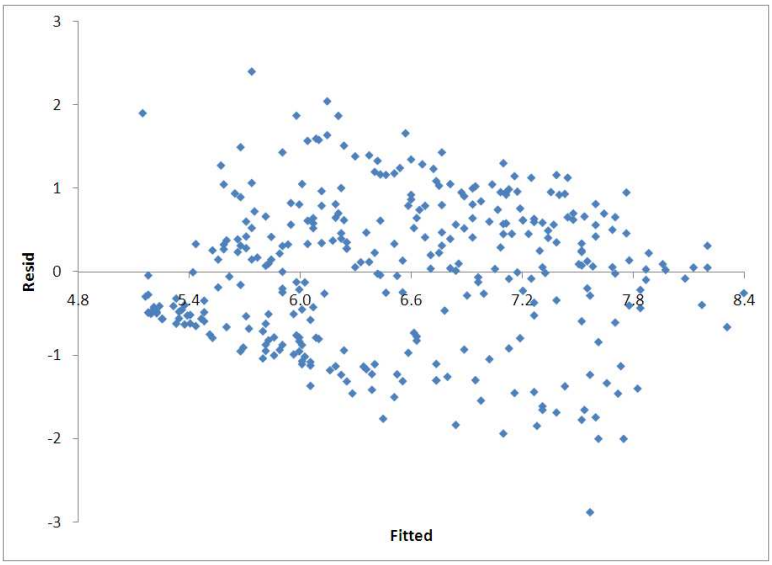

Figure 12.5: Plot of the residuals versus the fitted values when log(Salary) is regressed against Hits.

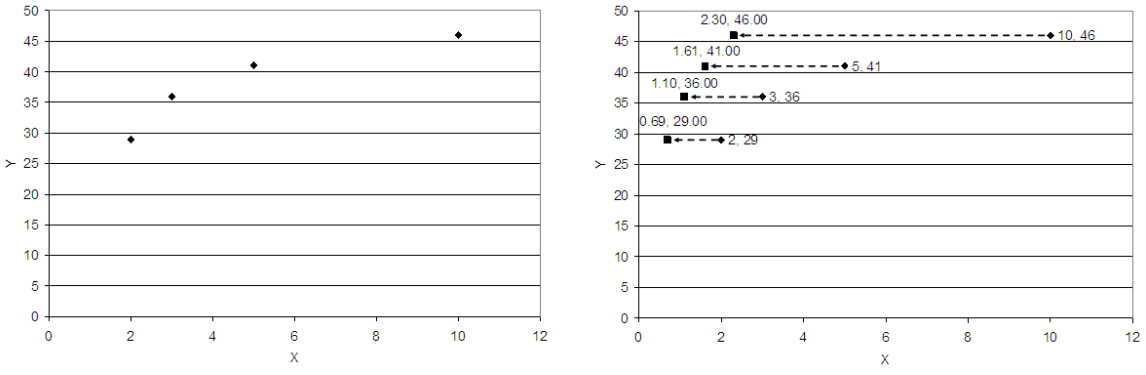

Figure 12.6: Plot of original data (left) and linearized data (right).

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.xls 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?

12.1.4 How To Guide

Using StatPro's Data Transformation Utility

Often it will be necessary to create models of the data that are not linear models. In order to do this using regression analysis, we must first transform the data with a nonlinear function (using StatPro, or typing the formulas into Excel ourselves) and then perform regression using these new, transformed variables as the data.

To have StatPro transform the data, follow these steps:

- Select the data (click anywhere in the data)
- Open the "StatPro" menu on the toolbar
- Select "Data Utilities"
- Select "Transform"
- Choose an appropriate transformation, based on the model you need (see below)

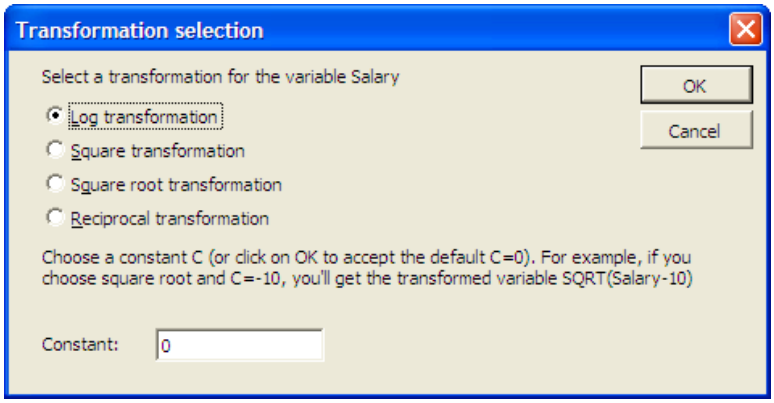

Figure 12.7: Dialog box for transforming a variable in StatPro.

There are four basic transformations available in StatPro (shown in the screen shot above). These can be combined to produce a wide variety of models, shown in the chart below, under "What transformations to use". Also note that the data transformation tool will allow you to shift the data by a known constant, opening up the number of different types of models you can produce to be much larger than your patience would allow you to investigate.

What transformations to use

Okay, now that you know how to use the transformation tool in StatPro, you need to know what variables to transform for each of the nonlinear models that you might encounter. Consult the table below for the type of model you want, and what variables you need to create in order to produce a regression model of that type. We assume that the original data

contains two variables, the response variable and the explanatory variable, and we assume that you are trying to develop a model equation to predict the response as a function of the explanatory.

To create this type of model	Response Variable	Explanatory Variable(s)
Linear model	Response	Explanatory
Square model	Response	Sqr(Explanatory)
Quadratic model	Response	BOTH Explanatory and Sqr(Explanatory)
Logarithmic model	Response	Log(Explanatory)
Exponential model	Log(Response)	Explanatory
General Power model	Log(Response)	Log(Explanatory)
Multivariate Power Model (multiplicative model)	Log(Response)	Log(Explanatory 1), Log(Explanatory 2), etc.
Reciprocal model	Response	Reciprocal(Explanatory)
Square Root model	Response	SqRt(Explanatory)

Transformations without StatPro

It is easily possible to construct these transformations without StatPro. Suppose you have a variable "X" extending from cell B1 to cell B51, with a label for the variable name in B1. In column C we will construct the natural log of this data, using the built-in Excel formula for the natural log: LN. Thus, the natural log of the observation in B2 is "=LN(B2)". Then fill down formula down to all the cells in column C.

To compute the square of a number, use the caret symbol (^) and the power 2 in the formula. Thus, in column D we can construct the square of the data "=B2^2".

To compute the square root, use the formula "SQRT". To compute the reciprocal, use the fact that the reciprocal of X is just 1/X.

Computing S_e and R^2 for nonlinear models

This is appropriate only for models in which the y variable (the response variable or dependent variable) is transformed. For example, if the model equation is of the form

$$\begin{aligned} \ln(y) &= A + Bx \\ e^{\ln(y)} &= e^{A+Bx} \\ y &= e^A e^{Bx} \end{aligned}$$

you will need to complete the process below in order to determine the actual S_e and R^2 of the model, since the summary statistics from the regression output are not based on the actual y variable at all. Thus, you only need to go through these steps when producing exponential models (like the example above), power function models, or multiplicative models.

This process will be broken down into several steps, listed in order below. A sample worksheet with all the formulas listed is shown as well, be sure you complete the steps in the

proper order. The numbers on the sheet show the order in which the columns and formulas should be constructed.

Recall that for exponential and multiplicative (power) regression models, the original response variable Y has been transformed by the log function. This means that the standard error of estimate for the model cannot be used to interpret how accurately the model fits the original Y data. The following activity will guide you through the process of finding a standard error of estimate that conforms with the original Y data.

Bring up the data found in C12 Power.xls. We will fit an exponential model to this data and then compute the that is consistent with the original cost data. After we have S_e, we can compute the R^2.

1. The exponential model can be constructed with the following steps

   - StatPro
   - Data Utilities
   - Transform Variable(s)
   - Select "cost" to transform
   - Select Log transformation

	A	B	C	D
1	Data on cost versus production level			
2				
3	Month	Cost	Units	Log(Cost)
4	1	45623	601	10.72817
5	2	46507	738	10.74736

2. Perform regression using Log Cost as the response variable. Be sure to check off the box to have the routine create columns for "fitted values" and "residuals". If you do not do this, start step (2) over. We must have the fitted values in order to complete this process.

   The Standard Error of Estimate of the model with response variable Log(Cost) cannot be directly used in interpreting how accurately the model fits the original Cost data. StatPro will insert a blank column (E) between the calculations. In order to remain consistent with the screen images and formulas below, you should right click on column E (in the original data) and select "Delete" from the context menu that appears.

   Your data sheet now looks like this:

	A	B	C	D	E	F
1	Data on cost versus production level					
2						
3	Month	Cost	Units	Log(Cost)	Fitted Values	Residuals
4	1	45623	601	10.72817	10.6373	0.0908

3. Next, we transform the Fitted Values produced by the model, which approximate the actual Log(Cost) values, back into the original cost units by applying the Antilog function (exp) to the fitted values with Exp(Fitted Values). Thus, in cell G4, type the formula "=EXP(E4)". Then fill this formula down to copy it to all the data.

4. Create the columns indicated below, in order, from left to right. In this case, remember that (a) the "Actual" values are the cost data. Column H "Actual minus EXP(Fitted)" is the residual data that is squared to produce the data in column I.

	E	F	G	H	I
1					
2					
3	Fitted Values	Residuals	Exp(Fit)	Resid	Resid^2
4	10.6373	0.0908	41661.43	3961.57	15694061.87
5	10.7463	0.0010	46458.42	48.58	2360.07
6	10.7049	-0.0280	44575.85	-1232.85	119923.07

5. We are ready to apply the formula for the standard deviation of the sum of the squares of the residuals (SSR):

$$\sqrt{\frac{\text{Sum of Squares of Residuals}}{n-p}} = \sqrt{\frac{\text{SSR}}{n-p}}$$

The bottom right corner of your data sheet should look something like this:

	H	I
37	2387.82	5701660364
38	-36.70	1346.99
39	2485.24	6176425.81
40		
41	Sum of Resid^2	296201375.38
42	Parameters (p)	2.00
43	Data Points	36.00
44	S_ E	2951.58

The number of parameters refers to the number of coefficients that your regression model was used to predict. In this case, there are two parameters. In general, the number of parameters for this step is the number of independent variables plus 1. This number is the p in the formula for standard error. The number of data points is the actual number of observations contained in the data. This is the n in the formula for standard error above.

Thus, about 2/3 of the actual cost data will fall within $2951.58 of the cost predicted by the exponential model.

To compute the R^2 value, we need to first compute the mean of the original y data, in this case, the cost data. It's also useful to have the standard deviation of the y data in order to compare the and determine how accurate the model really is. Once we have the actual mean of the y-data, we need to create another column to the right of our previous calculations. This column should compute the square of the actual values minus the mean of the y data. Thus, if you have the mean of the y-data in cell B41, the formula for the first entry in column J (J4) will be "=(B4 - B41)^2". Fill this formula down the column to get all the data.

Next, compute the total variation, the sum of all the results in column K. We have placed this information in cell J46. To get R^2, we then use the formula for R^2:

$$R^2 = 1 - \frac{\text{SSR}}{\text{Total Variation}}$$

Thus, in cell J47, we enter "=1 - I41/J46" to get the actual R^2 for this model.

NOTE: The following figures show a final result of these calculations, all on one spreadsheet, with all the formulas written out and explained. However, the formulas are different than this example. The reason is because in the example above, we have deleted the blank column E that StatPro inserts during the regression. In the image below, it has not been deleted.

CELL	Formula	CELL	Formula
B41	=average(B4:B39)	J44	=sqrt(J41/(J43-J42))
B42	=stdev(B4:B39)	K46	=sum(K4:K39)
J41	=sum(J4:J39)	K47	=1-J41/K46

COLUMN	Information
D	Created by "Data Utilities/Transform Variables" - This is the Log of the COST Variable.
F and G	Created by the Multiple Regression routine ONLY IF you remember to click the last check box in the diagnostic options during the regression setup. The "Fitted Values" are fitted LOG(Cost) values. The residuals are the difference between Log(Cost) and Fitted - not what we need.
H	These are the model's actual predictions for the Cost. To get these, we need to "unlog" the Fitted Values Column. We do this with the exponential function, which is EXP in Excel. Thus, in cell H4 we enter =EXP(F4) and copy this down the H column.
I	This column shows the real residuals - the difference between Cost and Exp(Fit). In I4 type =B4-H4 and copy this to the whole column.
J	This column computes the square of each of the residuals. Simply enter the formula =I4^2 and copy this down the column.
K	This column computes the square of the difference between each data point's Y-value (COST) and the mean of all the actual Y-Values. This is the variation of the data point. To compute this, in cell K4 enter =(B4-B41)^2 and fill down the column.

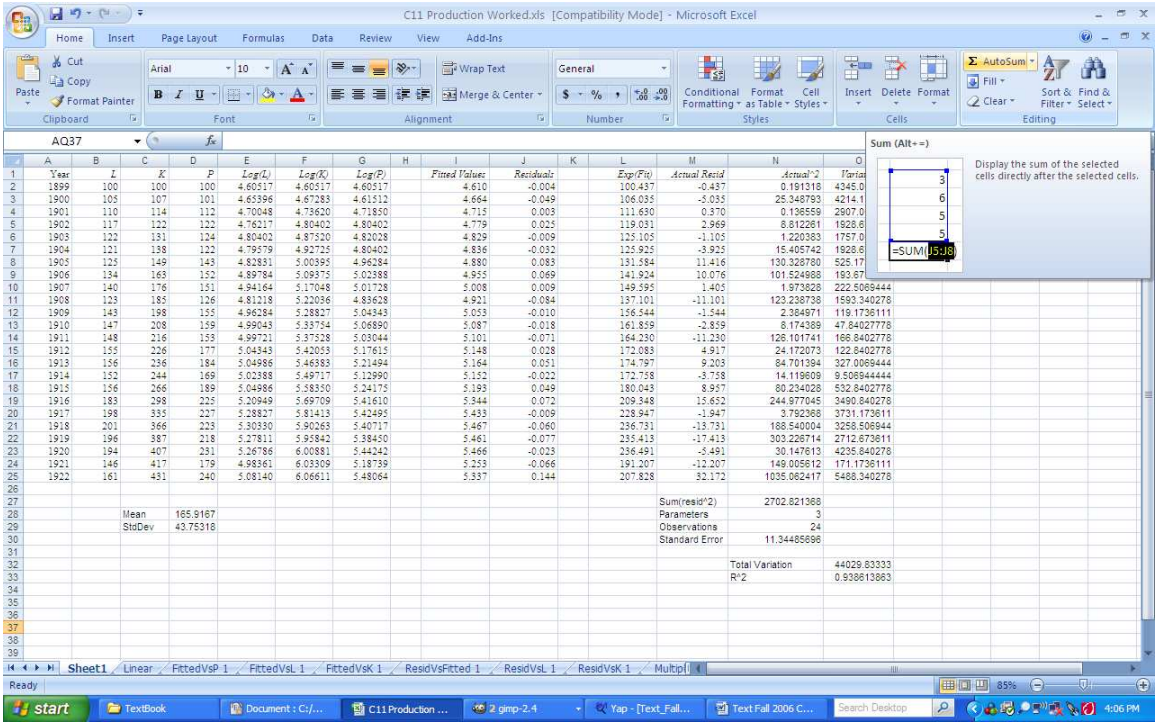

Figure 12.8: Screen showing the set up for calculating R^2 and S_e for nonlinear models.

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 358. 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 $0.30 \to 30\%$ 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.

$$\begin{array}{lcl} E1 & b^0 & = 1 \\ E2 & b^r b^s & = b^{r+s} \\ E3 & (b^r)^s & = b^{rs} \\ E4 & \frac{b^r}{b^s} & = b^{r-s} \end{array}$$

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.

$$\begin{array}{lcl} L1 & \ln(e^r) & = r \\ L2 & e^{\ln(a)} & = a \\ L3 & \ln(a) + \ln(b) & = \ln(ab) \\ L4 & \ln(a) - \ln(b) & = \ln\left(\frac{a}{b}\right) \\ L5 & r \cdot \ln(a) & = \ln(a^r) \end{array}$$

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 "$y = \ldots$" 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 Excel, enter the formula "=EXP(A)". Also note that we can use property E3 to rewrite the functional form as $y = e^A \left(e^B\right)^x$. 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: $y = \left(e^A\right) x^B$.

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.xls. 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 Cost = 23651 + 31*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 $\ln(\text{Cost}) = 10.1592 + 0.0008 * \text{Units}$, 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: $y = A + B * \ln(x)$
   (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 $-(98.35)/2 \cdot (-0.06) = 820$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 increases 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.xls. 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.xls. 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).

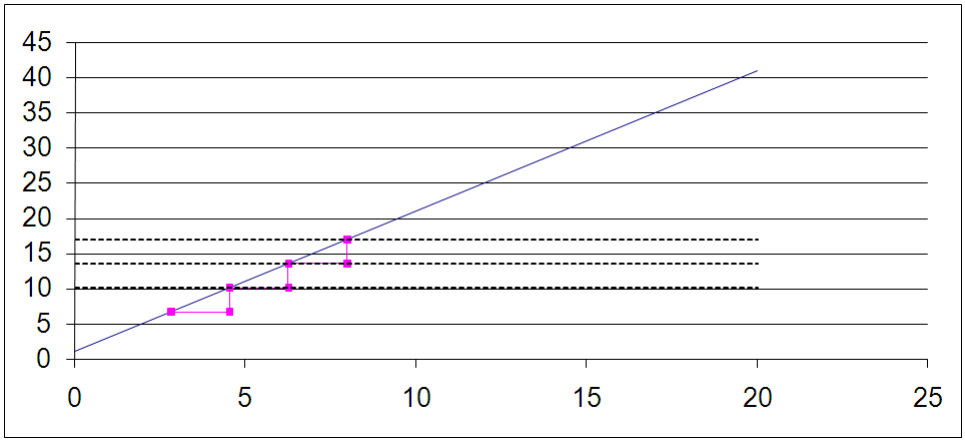

Figure 12.9: Screen shot of C12 StepByStep.xls.

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

$$\Delta y = f(x_1 + n\Delta x) - f(x_1) = n\left[f(x_1 + \Delta x) - f(x_1)\right].$$

However, for a nonlinear function, this is not true.

12.2.4 How To Guide

Creating a column-oriented, one-variable data table

Excel makes it easy to create tables of data from formulas using a data table. Essentially, you create a column of values you want to substitute into the formula, and enter a sample computation of the formula. Then you tell Excel which cell is the input cell and it substitutes all your values into the formula at once, producing a table of outputs corresponding to each input.

	A	B	C	D
1	A	2	-1	
2	B	2	2	
3	C	6	-3	
4				
5	X	4		
6				
7	X	46	-11	
8	1			
9	2			
10	3			
11	4			
12	5			
13	6			
14	7			
15	8			
16	9			
17	10			

Figure 12.10: Setting up a column-oriented, one variable data table.

Suppose you want to make a table of data so that you can compare two functions. If your functions are both quadratic, we might set up the spreadsheet as shown below. In cells B1:C3, we have entered two sets of A, B, and C values for a quadratic function of the form

$$y = Ax^2 + Bx + C$$

In cell B5 we have entered a sample x value. This x value will be used to generate sample calculations for the data table in cells B7 and C7. These cells contain the calculation of the functions, using the formulas

```
B7 = B1*B5*B5+B2*B5+B3
C7 = C1*B5*B5+C2*B5+C3
```

Once you have the sample calculations, you then create a list of the values you want to substitute into the formulas for x (the sample x in cell B5 will be replaced by the x values in your list). In the screen shot in figure 12.10, these values are listed in cells A8:A17.

To complete the data table, highlight the region of cells containing your list of values and the sample calculations, in this case A7:C17. Activate the data ribbon, click on "What if analysis" and select Data Table on the menu and you will see a dialog box like the one shown in figure 12.11.

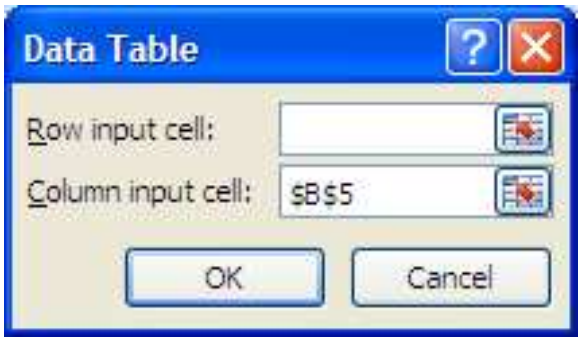

Figure 12.11: Dialog box for data tables.

Since your table is set up as a column-input table, in the "Column input cell" box enter a reference to the cell with the sample value of x used in your formulas. In this case, the sample value is in B5, so enter that. When you click "OK" the table will automatically compute the result of substituting values into the formulas.

FINAL NOTE: If you select any cell of the data table, you will not see a calculation or a value. Instead, you will see something like =TABLE(,B5). This is because the table is live, so changing the cells in B1:C3 or the input cells in A8:A13 or the formulas themselves in B7 or C7 will instantly propagate through the table. To freeze the values, first select the table, then copy it, and instead of pasting it, use "Edit/ Paste Special". Selecting "Values" from the list will copy just the resulting values and not the formulas.

12.3 Homework

12.3.1 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.

4. Use parameter analysis to interpret the model. 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)."

12.2. 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.

4. Use parameter analysis to interpret the model. 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)."

12.3. 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.
4. Use parameter analysis to interpret the model. 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)." (Hint: You will need to make two statements, one for each of the explanatory variables, for this model.)

12.3.2 Application and Reasoning Problems

Coming soon

12.3.3 Memo Problem

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.XLS"

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 polices 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.