Chapter 9

Multiple and Categorical Regression[1]

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 this chapter, 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.

The second part of this chapter will take 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 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*
  - $\surd$ How to read the measures of validity for multiple regression output
  - $\surd$ 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

- *As a result of this chapter, students will be able to*
  - √ Set up a multiple linear regression using StatPro
  - √ 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 using StatPro
  - √ 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 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 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 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, scatterplots, for picturing the data. Without a scatterplot of the actual data (Y vs. all the X variables) we cannot get Excel 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. That is, by holding 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 to the 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) adds exactly $27,000 to the 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 normal linear model, except that it includes several explanatory variables. If the explanatory variables are labeled X_1, X_2, ... and the response variable is Y, then a multiple-linear model for predicting Y would take the form

$$Y = A + B_1X_1 + B_2X_2 + \ldots + B_NX_N$$

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 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.

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.

$p < 0.05$ Keep the variable
$p > 0.05$ Drop the variable

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 loose $p+1$ 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.

$$\text{Multiple}R^2 = \frac{SSE}{SST} = \frac{SST-SSE}{SST} = 1-\frac{SSR}{SST}$$

SSR=Sum of the squares of the residuals (unexplained variation)
SSE=Explained amount of variation
SST=Total variation in y

Multiple R^2 is the coefficient of simple determination R-Squared between the responses y_i and the fitted values $\hat{y}_i$.

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".

Adjusted R^2 Adding more explanatory variables 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 SSTO 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.

9.1.2 Worked Examples

Example 9.1. Reading multiple regression output and generating an equation
In the last chapter, if you did the memo problem, you have encountered Ms. Carrie Allover, who needed help determining how each of the possible variables in her data are related to the number of riders each week on the commuter rail system she runs for a large metropolitan area. (See data file "C09 Rail System.xls".) However, in the last chapter, we were forced to examine the data one variable at a time. Now, we can try to build a model that incorporates all of the variables, so that each is controlled for in the resulting equation. If we have StatPro produce a full regression model, 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 the regression output is very similar to the output from 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 260:

$$\begin{aligned}\text{Weekly Riders} = & -173.1971 - 139.3649 * \text{Price per Ride} + 0.7763 * \text{Population} \\ & -0.0309 * \text{Income} + 131.0352 * \text{Parking Rate}\end{aligned}$$

Example 9.2. Interpreting a multiple regression equation and its quality
The rail system model (previous example, see the data file C09 Rail System.xls) 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 really 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 is a little shaky is the y-intercept. It's p-value is 0.4389, far above the acceptable level. This means that we could reasonably expect that the y-intercept is actually 0, and that would make a lot of sense in interpreting the model. 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.

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" graphs that StatPro produces (if you checked the appropriate boxes in the regression routine). These graphs are shown below.

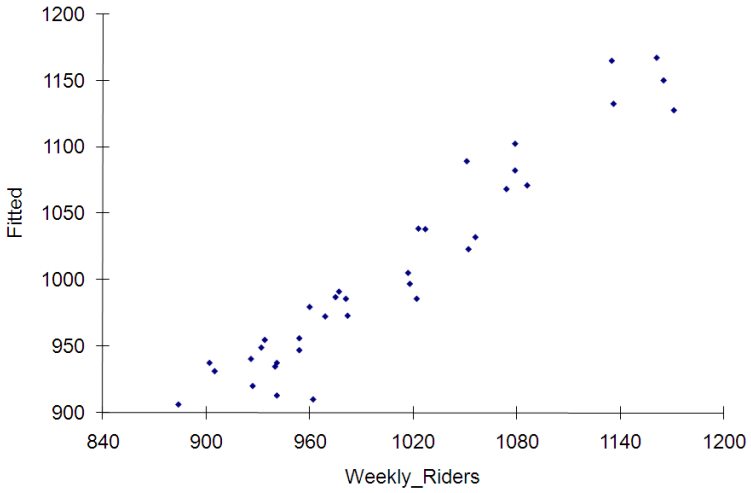

Figure 9.1: Graph of fitted values versus the actual values for the rail system example.

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.

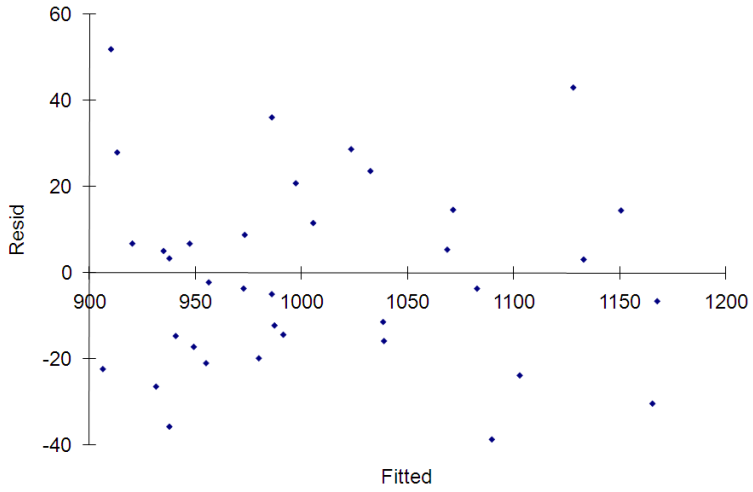

Figure 9.2: Graph of the residuals versus the fitted values for the rail system example.

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 that are outside the range included in the actual data (extrapolation)

2. predict values for the response variable, based on values of the explanatory variables in between the values in the actual data (interpolation)

3. find values of the explanatory variables that produce a specific value of the response variable (solving an equation)

Suppose, for example, that Ms. Allover (see the data file C09 Rail System.xls) wants to use the regression model 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

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.xls"):

- 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.

9.1.4 How To Guide

Multiple regression in StatPro

StatPro makes multiple regression analysis fairly easy. Follow the steps below with any set of data to develop a regression model along with its associated measures and graphs.

1. Select the region of the worksheet that contains the data.
2. Select the StatPro routine to apply to the data.

   The routine for simple regression is under "Regression/ Multiple..."
3. Verify that the data region is correct.
4. Select the variables to which the routine will apply.

   StatPro will give you two screens on which to select variables. The first screen allows you to select the dependent (response) variable. The second allows you to select the explanatory (independent) variables. To select several explanatory variables, hold down the control (CTRL) key while you select them with the mouse.

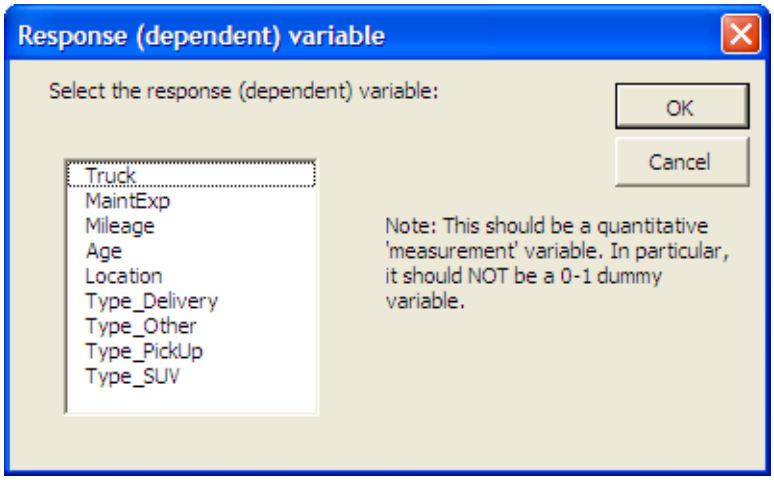

Figure 9.3: Selecting variables in multiple regression.

5. Fill in the details of the routine.

   For multiple regression, the routine will ask you which diagnostic graphs you want produced (the dialog box is shown below). Usually, though, you will only need the first and third graphs ("Fitted versus actual" and "Residuals versus fitted"). These two scatterplots allow you to determine whether the model, in this case a linear model, is a good choice for your data. You may also want to have the regression routine calculate the predicted ("fitted") values and the residuals. If so, be sure to check the box at the bottom of the dialog box. Unlike simple regression, these are not calculated automatically. If you select the second and fourth options, StatPro will produce quite a few graphs; each choice creates one graph for each explanatory variable.

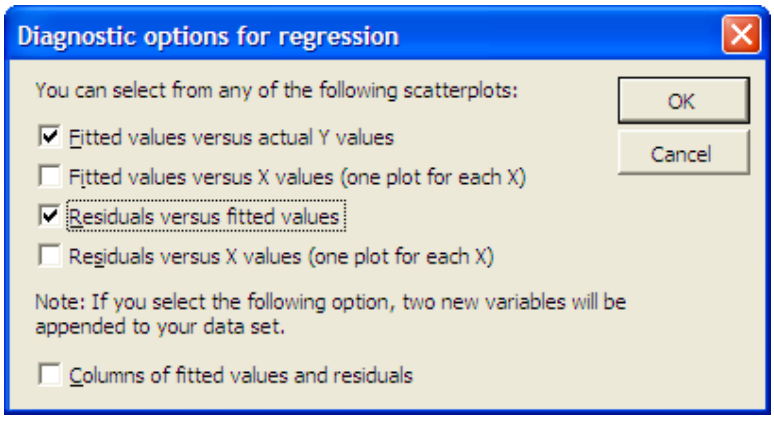

Figure 9.4: Selecting diagnostic graphs in multiple regression.

6. Select the placement for the output of the routine.

   For regression output, we definitely suggest putting the output on its own sheet. Name the sheet so that you can tell which model it contains (in case there are several different explanatory variables to choose from).

   Now that you have the regression coefficients, you can easily create Excel formula to compute the predicted y-values, based on given x-values. This will let you explore the model's predictive power. See below for how to do this most easily (to avoid re-typing the coefficients).

Stepwise regression in StatPro

Stepwise regression is very similar to performing multiple regression. In steps 1 - 4, nothing should be changed, except to use "StatPro/ Regression Analysis... /Stepwise..." Also, in step 4, select all of the explanatory variables that you think might be important. Usually, all of the non-identifier variables are selected. In step 5, there will be two more screens to fill in. For right now, leave everything as is and just click "OK".

For this routine, definitely have the output placed on a new worksheet. We advise that you call the new sheet "Step" so that you know it is the stepwise regression model. When the routine is finished (it may take a little longer than the multiple regression routine) you will have a lot more information. Basically, the routine will make several multiple variable models, carefully adding one explanatory variable (from the list you chose) at a time and seeing how it affects the summary measures and the p-values. Scroll down the output screen to see how each variable was added. The last step will show you the best model the routine could develop, based on the variables you gave the routine and the parameters you set in step 5 to determine when to throw variables out.

Easily predicting values from a multiple linear model

For this example, suppose you have values of the variables on the data sheet in cells C42:F42 and suppose the regression output is on a worksheet called "FullModel" with the coefficients

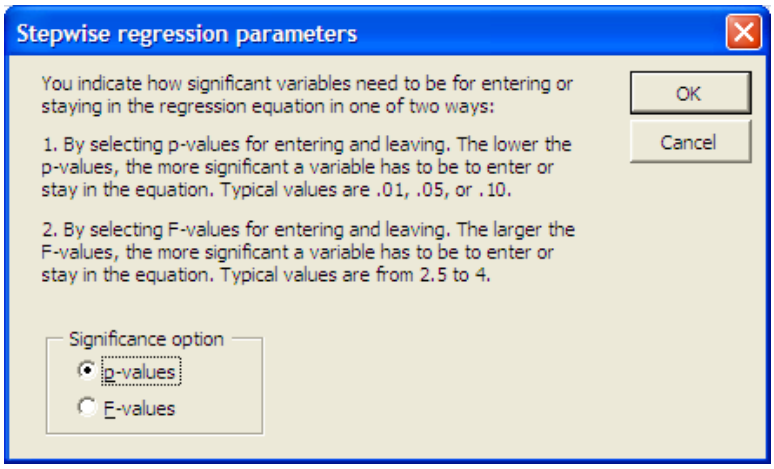

Figure 9.5: Selecting the p-value option for stepwise regression in StatPro.

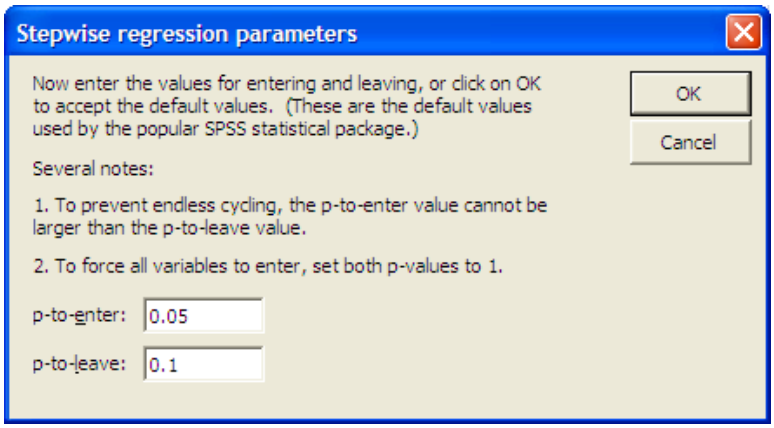

Figure 9.6: Setting significance for p-values in stepwise regression.

in cells C16:C20 (the constant is in C16). The following formula will calculate the value of the response variable that corresponds to the values of the explanatory variables in C42:F42:

```
= Constant Cell + MMULT(X values, Coefficients)
```

So for our setup, use "=FullModel!C16 + MMULT(C42:F42, FullModel!C17:C20)".

Be sure that you have the variables in the same order in both places. Also be sure that the X-values are listed first in the model. The x-values must also be listed in a row (not a column) for the formula to work. If the x-values are in a column (like the coefficients) use the formula

```
= Constant Cell + SUMPRODUCT(X values, Coefficients)
```

Also remember that if you want to calculate values for several sets of x-values by copying the formula to other cells, be sure that the cell references to the coefficients are absolute and not relative (the reference to the constant term must also be absolute).

9.2 Modeling with Qualitative Variables

For Excel/StatPro, as well as any other statistical package, 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 (such as StatPro) to be able to create regression models, the various values in each category 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.

Exact multicollinearity This is an error that can occur if some of the explanatory variables are exactly related by a linear equation.

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. (StatPro can automate this process; you can also use IF statements to create a formula to calculate the values.)

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

Annual Salary = 4667 - 2345*(0) + 845*Age = 4667 + 845*Age

When GenderF has value 1, (female employee), the salary is

Annual Salary = 4667 - 2345*1 + 845*Age = 2322 + 845*Age

We can now see that the single regression equation with dummy variables is actually two separate equations, one for each gender:

For a female employee: Annual Salary = 2322 + 845*Age
For a male employee: Annual Salary = 4667 + 845*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

The data file "C09 Truck data.xls" 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
    - Response variable:
    - Numerical variable:
    - Categorical variables broken down by categories:
    - Dummy Variables: (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?

9.2.4 How To Guide

Generating dummy variables in StatPro

StatPro has a variety of tools available to manipulate your data. Under "Data utilities/ Create dummy variables..." you can convert either a categorical variable into several dummy variables (one for each category) or a numerical variable into a single dummy variable, based on a cutoff value.

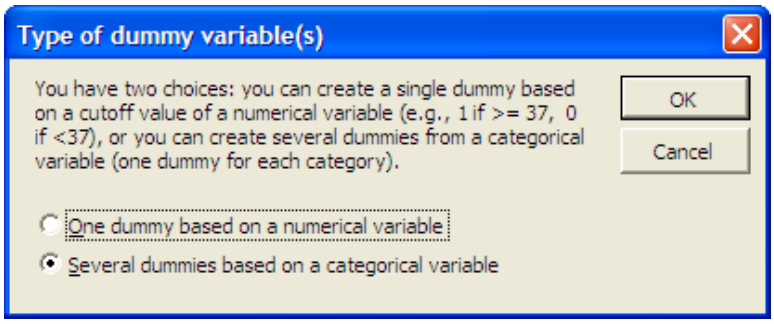

Figure 9.7: Two options for making dummy variables in StatPro. Usually you will want to choose to make several dummy variables from a single categorical variable (the second choice, shown selected.)

Option #1: One dummy based on a numerical variable This option will not be used too often. After you select a variable (pick a numerical one or a categorical variable that is coded with numbers, like "Education level" or "Job Grade"), you will need to determine the definition of the dummy variable. Simply enter a cutoff value and select a criterion. The name of the new variable will be a description of the cutoff criterion added to the end of the original variable name. The new variable will be added to the data set at the far right in a new column.

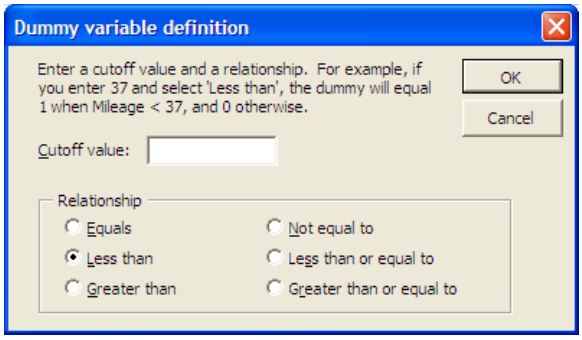

Figure 9.8: Dialog box for choosing how to construct a single dummy variable from a numerical variable.

Option #2: Several dummy variables from a categorical This will be the option you use most often. Once you select a categorical variable, the routine automatically creates

several dummy variables from it at the far right of your data set. Each is named by adding a category to the end of the name of the original categorical variable. Thus, a "Gender" variable which was originally coded "Male" or "Female" would be turned into two new variables "GenderMale" and "GenderFemale".

Dummy variables with IF functions

If you do not have StatPro available, it is easy to create a formula to make a dummy variable. Suppose you have a variable "Gender" coded as "Male" or "Female" in column B2:B40, with the variable name "Gender" in cell B1. Let's say you want to create a "GenderMale" variable in column F. First type the variable name "GenderMale" in cell F1. Then, in cell F2, enter the formula

```
=IF(B2=''Male'', 1, 0)
```

After the formula is entered, select the cell and double-click the fill handle to copy the formula down the column. You can enter whatever category name you want in the formula in place of "Male", but be certain that you type it exactly as it is coded in the original categorical variable.

The IF function has the following syntax.

```
IF(condition, value if true, value if false)
```

The condition can be any sort of logical condition and can include checking for whether a cell is equal to a particular value, greater than a particular value, or whatever. See the help files on "IF" for more information.

Multiple regression with dummy variables

This is just like normal multiple regression with the following modifications:

1. First create all the dummy variables that you want to use in the model.
2. When selecting explanatory variables, leave one dummy variable out from each categorical variable. Otherwise, you will get the "exact multicollinearity" error.
3. When selecting explanatory variables, be sure to not use both the original categorical variable and the dummy variables.

9.3 Homework

9.3.1 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.xls".

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.xls". 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?

9.3.2 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 (page 263). 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 Excel to help answer this.

9.4. The data file "C09 Homes.xls" 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				

9.3.3 Memo Problem

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.XLS"