Chapter 7:
Modeling Relationships of Multiple Variables with Linear Regression

Overview

Chapters 5 and 6 examined methods to test relationships between two variables. Many research projects, however, require analyses to test the relationships of multiple independent variables with a dependent variable. This chapter describes why researchers use modeling and then examines one of the most powerful modeling approaches: linear regression. We show how to interpret regression statistics and graph linear regressions using the STATES10 data. Finally, we discuss issues related to data structures and model building.

The Advantages of Modeling Relationships in Multiple Regression

In most studies, building multiple regression models is the final stage of data analysis. These models can contain many variables that operate independently, or in concert with one another, to explain variation in the dependent variable. For example, as we discussed in previous chapters, both gender and education status can predict when a person has a child. Using multiple regression can help a researcher understand, for example, the association between education status and age of having a first born child *above and beyond* the influence of gender. It can also be used to understand how much of the variation in the age of having a first born child can be explained by the combination of those two factors.

However, before one begins a multiple regression, it is critical to follow the stages of analysis already presented in the previous chapters. In the first stage, researchers use univariate analyses to understand structures and distributions of variables, as these can affect the choice of what types of models to create and how to interpret the output of those models. In the second stage, researchers use bivariate analyses to understand relationships between variable pairs. These relationships can "disappear" or emerge in new—and sometimes unexpected—ways once

all the variables are considered together in one model. In some circumstances, the emergence and disappearance of relationships can indicate important findings that result from the multiple variable models. But alternately, disparate results could indicate fragilities in the models themselves. As we discuss later in this chapter, a core concern in regression analysis is creating robust models that satisfy both mathematical and theoretical assumptions. A firm understanding of the data at the individual and bivariate levels is critical to satisfying this goal.

Before discussing model building and the application of multiple linear regression, let us first take a step back and reflect on the reasons why they are needed. Suppose, for example, that a researcher is interested in predicting academic success, a construct that she operationalizes as grade point averages (GPAs). After determining that GPAs are approximately normally distributed, she performs a series of bivariate analyses that reveal the following sets of relationships:

- Women have higher GPAs than men
- Length of time spent studying is positively associated with GPAs
- Students who are members of fraternities have the same GPAs as non members
- Students who are in sport teams have lower GPAs than non-athletes
- Sophomores, juniors, and seniors have higher GPAs than freshmen
- Students who drink heavily have lower GPAs than light drinkers and abstainers

These are interesting findings, but they open new questions. For example, do freshman and athletes receive lower grades because they study less? Are the GPAs of men pulled downward because they are more likely to be members of sports teams? Can the differences between the performances of men and women be attributed to men drinking more heavily? Answering these types of questions requires considering not only the relationships between the dependent variable (GPA) and individual independent variables (gender, drinking, sports, etc.), but also the constellation of variables that correspond with being a student.

One of the great benefits of regression analysis is its ability to document **collective effects -** the interplay among factors on predicted outcomes. For instance, regression models can predict the expected GPAs based on combinations of variables as they may be configured in the lives of individuals (e.g., a non-drinking, female, athlete). They also can measure the amount of variation in the dependent variable that can be attributed to the variables in the model, and conversely, how much of the variation is left unexplained.

In the case of GPAs, a regression model can tell the strength of the six factors in predicting academic success. Do these factors account for many of the differences among students, or only a small fraction of the differences? Regression models measure **explanatory power**, and how well predictions of social behavior correspond with observations of social behavior.

Additionally, multiple regression models are critical to accounting for the potential impact of **spurious relationships**. Recall from Chapter 1 that a spurious relationship occurs when a third variable creates the appearance of relationship between two other variables, but this relationship disappears when that third variable is included in the analysis.

Using the example above, perhaps the differences in performances of athletes and non-athletes may simply be the result of athletes spending less time studying. If the negative association of athletics to GPA “disappears” when studying is taken into account, it leads to a

more sophisticated understanding of social behavior, and more informed policy recommendations.

Finally, one of the great advantages of mulitple regression models is that they allow for the inclusion of **control variables**. Control variables not only help researchers account for spurious relationships, they measure the impact of any given variable above and beyond the effects of other variables. For example, a researcher could document the influence of drinking on GPAs adjusting for the impact of gender, sports, fraternities, and time spent studying. Or consider the relationship between gender and GPA. Suppose the relationship between gender and GPA disappears after taking into account all of the other variables in the model. What would that suggest about theories that posit innate differences in abilities to succeed in college?

Putting Theory First - When to Pursue Linear Regression

Later in this chapter we consider some mathematical principles and assumptions that underpin linear regression. But first we consider some theoretical issues critical to its application.

Linear regressions are designed to measure one specific type of relationship between variables: those that take **linear** form. The theoretical assumption is that for every one-unit change in the independent variable, there will be a consistent and uniform change in the dependent variable. Perhaps one reason why linear regression is so popular is that this is a fairly easy way to conceive of social behavior – if more of one thing is added, the other thing will increase or decrease proportionately. Many relationships do operate this way. More calories results in proportional weight gains, more education results in proportionally higher earnings, etc. In our example above, a linear model assumes that that each additional hour a student spends studying (whether the increase is from 5 to 6 hours a day, or from 1 to 2 hours a day) the incremental effect on the GPA will be constant. This may be true, but it also may not.

Recall the discussion in Chapter 5, that there are many types of relationships between variables. For example, students who experience little anxiety and those who experience excessive anxiety tend to perform more poorly on exams than students who score midrange in an anxiety scale (these individuals are very alert, but not overwhelmed). Because this "inverted U" shaped relationship is nonlinear, the application of linear techniques will make it appear non-existent.

Or consider another example - the impact of class size on academic performance. It is generally understood that a negative relationship exists between class size and academic performance — the smaller the class the more students benefit. However, changing a class from 25 students to 20 students will have almost no effect on student performance. In contrast, the same increment of change from 12 to 7 students can have a much more substantial change in classroom dynamics.[18] The same 5-student difference in class size has different effects, depending on where the incremental change is located .

Conversely, there may be positive effects of time spent studying on GPA, but the benefits of each additional hour may be smaller once a sufficient amount of study time is reached.

[18] This type of *logarithmic* relationship can still be tested using linear regression techniques, but it requires transforming data so that the model corresponds to the way the data are actually configured. Describing these transformations is beyond the scope of this book, but for a description of these methods, see *Applied Linear Statistical Models* by Michael H. Kutner, John Neter, Christopher J. Nachtsheim, and William Wasserman (2004).

Linear Regression: A Bivariate Example

Later in this chapter we detail some criteria for applying linear regression. But before bogging down the discussion in cautions, let us look at its application and interpretation. Note, though, that the structure of the dependent variable is a critical consideration, as linear regressions are performed on scale dependent variables. In our sample illustration, we will be testing the relationship between poverty (independent variable) and the percent of births that are to teenage mothers (dependent variable) using the STATES10 data. Our guiding hypothesis is that places with higher poverty rates will have higher proportions of births occurring to teenage mothers. Before considering why this relationship may exist, we determine *if* it exists.

We already identified some ways to look at relationships between two scale variables in Chapter 5 - correlations and scatter plots. The scatter plot of these variables in Figure 7.1 shows that the data points tend to flow from the lower left-hand corner of the graph to the upper right. The correlation of these two variables is .774, a strong positive relationship. The linear regression determines the equation of the line that best describes that relationship. This equation can be used to predict values of the dependent variable from values of the independent variable.

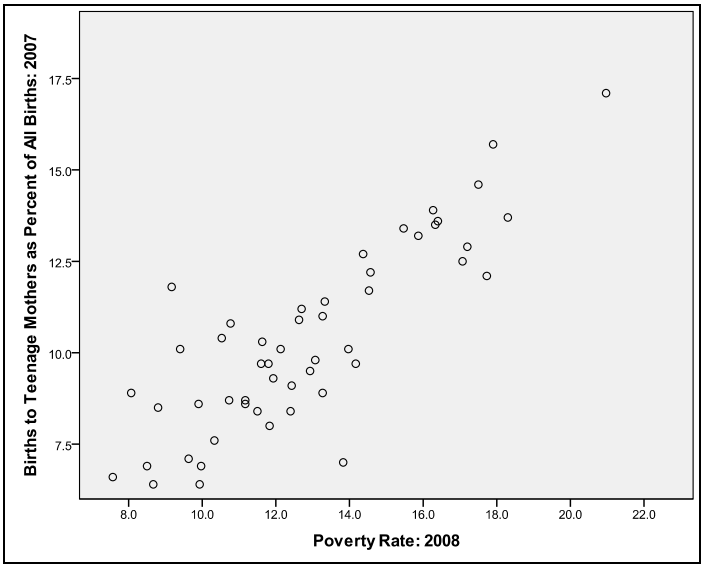

Figure 7.1 Scatter Plot of PVS519 and DMS397

To perform a regression, open the STATES10 data and specify the following variables in the regression menu:

Analyze
Regression
Linear
Dependent: DMS397 (Births to Teenage Mothers as a Percent of All Births: 2007)
Independent(s): PVS519 (Poverty Rate: 2008)
OK

Figure 7.2 contains the resulting regression output. We will concentrate on three groups of statistics from this output: the coefficients, the significance tests, and the R square statistic.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.844[a]	.713	.707	1.3683

a. Predictors: (Constant), PVS519 Poverty Rate: 2008

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	227.944	1	227.944	121.749	.000[a]
	Residual	91.740	49	1.872		
	Total	319.684	50			

a. Predictors: (Constant), PVS519 Poverty Rate: 2008

b. Dependent Variable: DMS397 Births to Teenage Mothers as Percent of All Births: 2007

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.395	.835		1.672	.101
	PVS519 Poverty Rate: 2008	.698	.063	.844	11.034	.000

a. Dependent Variable: DMS397 Births to Teenage Mothers as Percent of All Births: 2007

Figure 7.2 Regression Output

Interpreting the ANOVA F-test

Although this table comes second in the output, the first statistics to look at are the F statistic and its significance value in the ANOVA table. This should look familiar to you—it's the same type of ANOVA table we had in the one-way ANOVA test in Chapter 6. In a regression model, the ANOVA F statistic tests whether the model as a whole is significant. In other words, do the independent variables, taken together, predict the dependent variable better than just predicting the mean for everything?

In the simple linear regression of this example, there is only one independent variable, so the F-test is testing if this one variable, the poverty rate, predicts the percent of births to teen mothers better than if we used the average teenage birth percentage to predict all states' values. We would just use the average for all states if the relationship in Figure 7.1 were flat.

The model clearly does better than a flat line—the p-value (in the Sig. column) is very low, less than .001. So there is less than a 1 in 1,000 chance that the relationship we found in this sample is actually best described by a flat line.

Interpreting Linear Regression Coefficients

The **unstandardized coefficient** of an independent variable (also called **B** or **slope**) measures the strength of its relationship with the dependent variable. It is interpreted as the size of the average difference in the dependent variable that corresponds with a one-unit difference in the independent variable. A **coefficient** of 0 means that the values of the dependent variable do not consistently differ as the values of the independent variable increase. In that case, we would conclude that there is no linear relationship between the variables. In our model, the coefficient for poverty rate is .698. For every one-percent increase in the poverty rate, there is a predicted increase in the percentage of births to teens of .690.

Moving across the row for "Poverty Rate: 2008" in the Coefficients Table, we find a significance **(Sig.) score**. The significance score of .000 indicates that chance is an extremely unlikely explanation, as there is less than a 1/1,000 chance of a relationship this strong emerging, within a data set this large simply because of random chance. Because the relationship is significant, we are confident of an actual linear association between poverty and the proportion of all births attributed to teen mothers.

Because we only have one independent variable in the model, the p-value for its coefficient is exactly the same as the p-value for the ANOVA F-statistic. They're actually testing the same thing. But as we start to add more independent variables to the model, that won't be true. The coefficients will test the unique effect of each independent variable, and the F-test will test the joint effects of all the variables together.

The B column also shows the **constant**, a statistic indicating the **intercept**—the predicted value of the dependent variable when the independent variable has a value of 0. The intercept also has a significance level associated with it, but this statistic is usually ignored. We will show how the intercept is used in predictions and the formulation of a regression line.

In this example, the constant is the predicted percentage of births to teenage mothers if a state had no one living below the poverty line (a poverty rate of 0). Even if a state had no one in poverty, we could still expect 1.395% of births to teenage mothers each year. This means that in an ideal world where poverty were essentially eliminated in the United States, we might expect that teenage mothering could largely disappear as well, since 98.6% (100%-1.395%) of births would occur to women beyond their teenage years. Of course the cross sectional STATES10 data can not establish that this would in fact occur, but as we discuss further below, it does offer a way to estimate the impact of poverty reduction on social behavior. At a minimum, these data establish a strong relationship between poverty and the proportion of births that are to teen mothers.

Interpreting the R-square Statistic

The R-square statistic measures the regression model's usefulness in predicting outcomes – indicating how much of the dependent variable's variation is due to its relationship with the independent variable(s). An R-square of 1 means that the independent variable explains 100% of the dependent variable's variation—it entirely determines its values. Conversely, an R-square of 0 means that the independent variable explains none of the variation in the dependent variable—it has no explanatory power whatsoever.

The Model Summary table for our example shows the R-square is .713, meaning 71.3% of the variation from state to state in the percentages of births to teenage mothers can be explained by variation in their poverty rates. The remaining 28.7% can be explained by other factors that are not in the model. You may have also noticed that next to the R-square statistic is

the correlation between the two variables, R= .844. In bivariate linear regressions, the R-square is actually calculated by squaring the correlation coefficient (.844*.844=.713).

Putting the Statistics Together

In sum, there are four important statistics to attend to in a linear regression model. First, the F statistic tests whether the model as a whole predicts the dependent variable. Second, the regression coefficients measure the strength and direction of the relationships. Third, for each of these regression coefficients, there is a significance score, which measures the likelihood that the relationship revealed in the coefficients can be attributed to random chance. Finally, the R-square statistic measures the model's overall predictive power and the extent to which the variables in the model explain the variation in the dependent variable.

In our example, we found that teenage births are positively related to poverty—poorer states have, on average, a higher percentage of births to teenage mothers than affluent states. We also observed that this relationship is unlikely to be due to random chance. Finally, we discovered that poverty has very strong predictive powers, as this one variable accounts for well over two thirds (71.3%) of the variation in teen birth rate within the United States.

Using Linear Regression Coefficients To Make Predictions

One of the most useful applications of linear regression coefficients is to make "what if" predictions. For example, what if we were able to reduce the number of families living in poverty by a given percentage? What effect would that have on teen births? One way to generate answers to these questions is to use regression formulas to predict values for a dependent variable:

$\hat{Y}$	=	A	+	B	(X)
Predicted value of Y (Dependent Variable)	=	Y axis intercept (The constant)	+	Predicted increase of Y for 1 unit increase in X (The slope or coefficient)	Multiply value of X (Independent Variable)

This equation, combined with the information provided by the regression output, allows researchers to predict the value of the dependent variable for any value of the independent variable. Suppose, for example, that we wanted to predict the percentage of births to teenage mothers if the poverty rate is 20%. To do this, substitute 20 for X in the regression equation and the values for the constant and coefficient from the regression output.

$\hat{Y}$	=	A	+	B	(X)
Predicted	=	1.395	+	.698	(?)

$\hat{Y}$	=	A	+	B	(X)
15.36	=	1.395	+	.698	(20)

We calculate that for a state with a poverty rate of 20% , our best prediction for percentage of births to teen mothers is 15.36.

Making predictions from regression coefficients can help measure the effects of social policy. We can predict how much the teenage birth rate *could* decline if poverty rates in states were reduced. What would the predicted percent of births attributed to teenage mothers be if a state could reduce its poverty rate from 15% to 10%?

$\hat{Y}$	=	A	+	B	(X)
11.87	=	1.395	+	.698	(15)

$\hat{Y}$	=	A	+	B	(X)
8.38	=	1.395	+	.698	(10)

Percent of births to teens at 20% poverty rate = 15.36%
Percent of births to teens at 15% poverty rate = 11.87%
Percent of births to teens at 10% poverty rate = 8.38%

It is good practice only to use values in the independent variable's available range when making predictions. Because we used data with poverty rates between 8% and 22% to construct the regression equation, we predict teen pregnancy rates only for poverty rates within this range. The relationship could change for poverty rates beyond 22%. It could level off, or even decrease, or the rates could skyrocket, as some sociological studies indicate. Because our data do not tell us about the relationship for places of concentrated poverty, we must not use the regression line to make predictions about them.

Using Coefficients to Graph Bivariate Regression Lines

Regression coefficients are generally not meaningful to audiences unfamiliar with statistics, but graphs of regression lines are. To illustrate, we create a regression line using the independent variable, PVS519, its coefficient, and the intercept.

We first want to create a scatter plot of the teen birth rate (DMS397) and the poverty rate (PVS519). We also want to superimpose a line on that scatter plot that represents the percentage of all births attributed to teenage mothers as related to poverty rates (PVS519). To do this, we are going to create a scatterplot, then use the *Chart Editor* to add a regression line.

First create the scatterplot with the following commands:

Graphs
- *Chart Builder*
  - *Gallery: Scatter/Dot*
    - Drag *Simple Scatter* to the Chart Preview
    - Drag DMS397 from *Variables* list to *X-axis?* Box
    - Drag PVS519 from *Variables* list to *Y-axis?* Box
  - *OK*

Your screen should look largely like Figure 7.3.

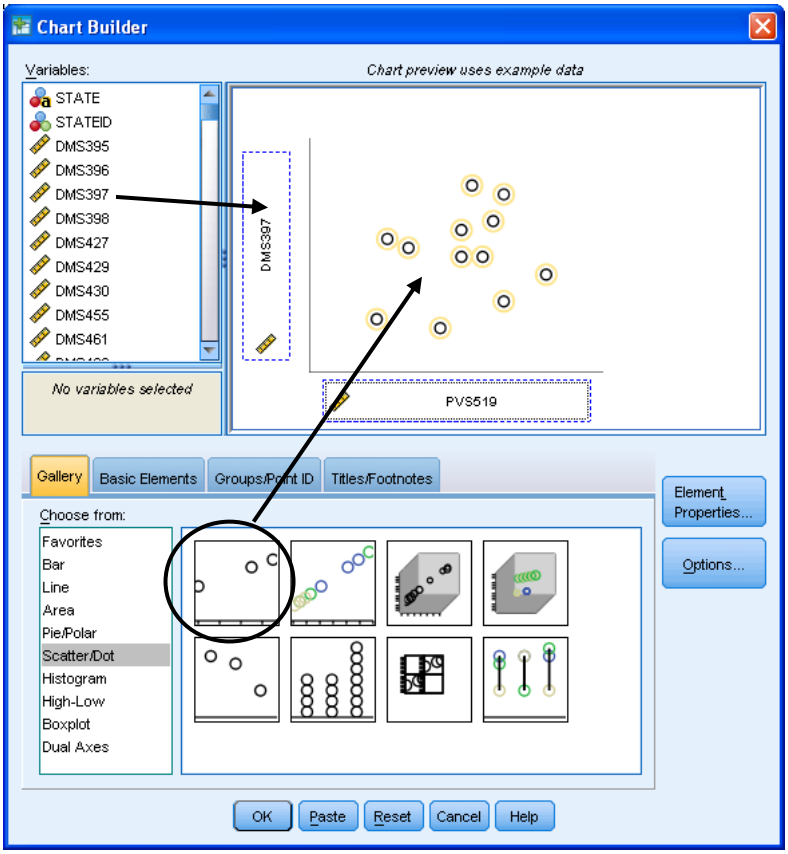

Figure 7.3 Chart Builder Dialog Windows

Once you have the Scatterplot, you can double-click on it to invoke the *Chart Editor*. You can use the *Chart Editor* to change the appearance of the plot, including the scale of the axes, the color of the points and the background. But you will also use it to add the regression line. In a simple regression, like this one, you could just add a regression line in the *Chart Editor* by selecting:

Elements
Fit Line at Total

This approach will not work once we move on to multiple regression models, so we will show you here, in the simpler context, how to compute this line as illustrated in the right panel of Figure 7.4. In the *Chart Editor*, perform the following commands to create a line of the predicted values:

Options
Reference Line from Equation
*Reference Line: Custom Equation: 1.395 + (.698 * x)*
Apply
Close the *Chart Editor*

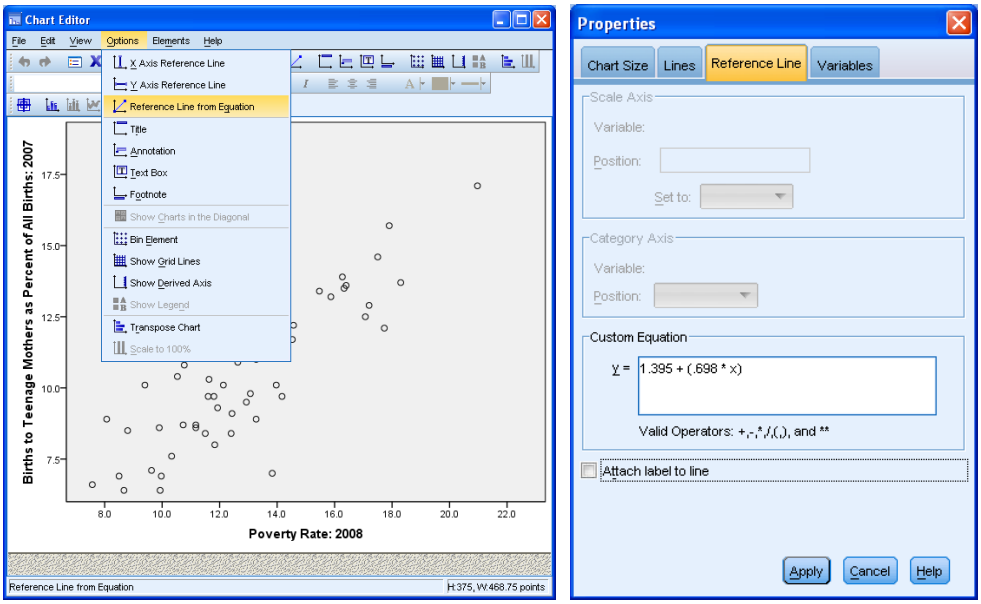

Figure 7.4 Adding Regression Line Using the *Chart Editor*

When you do these commands, there will be default values for a coefficient and intercept. The coefficient is the number multiplied by *x* and the intercept is added to it. We used the coefficient and slope from our regression results.

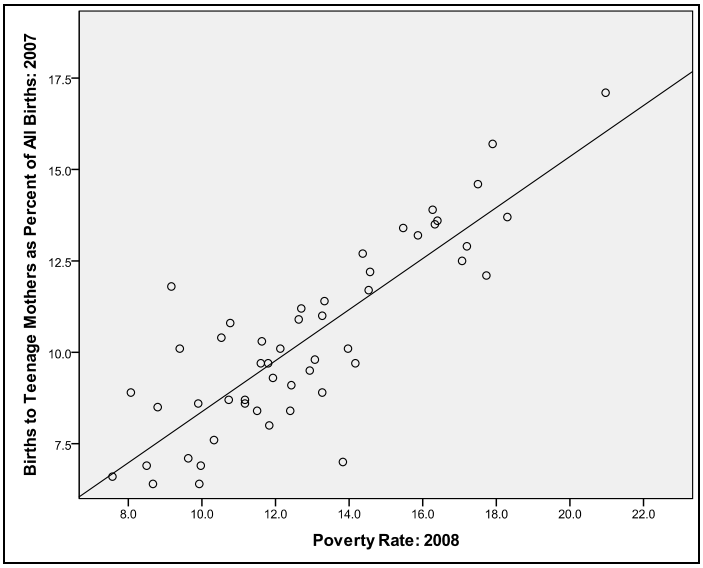

Figure 7.5 A Scatterplot with Predicted Regression Line

After you run these commands, you will observe that SPSS creates a chart that looks fairly similar to the one represented in Figure 7.5. The line we inserted represents the linear regression model, and contains the predicted values of percentage of births attributed to teenage mothers for every value of the poverty rate. Although the line extends far to the left and right of the data, remember not to use the predicted values that fall beyond the scope of the data possessed.

Multiple Linear Regression

In a multiple linear regression, more than one independent variable is included in the regression model. As we discussed earlier, multiple regression examines how two or more variables act together to affect the dependent variable. This allows researchers to introduce control variables that may account for observed relationships, as well as document cumulative effects.

While the interpretation of the statistics in multiple regression is, on the whole, the same as in bivariate regression, there is one important difference. In bivariate regression, the regression coefficient is interpreted as the predicted change in the value of the dependent variable for a one-unit change in the independent variable. In multiple regression, the effects of multiple independent variables often overlap in their association with the dependent variable. The coefficients printed by SPSS don't include the overlapping part of the association. Multiple regression coefficients only describe the unique association between the dependent and that independent variable. The ANOVA F-test and the R-square statistic include this overlapping portion.

This means a variable's coefficient shows the "net strength" of the relationship of that particular independent variable to the dependent variable, above and beyond the relationships of the other independent variables. Each coefficient is then interpreted as the predicted change in the value of the dependent variable for a one-unit change in the independent variable, after accounting for the effects of the other variables in the model.

To illustrate how to perform a multiple linear regression, we will expand the study of the percentage of all births attributed to teenage mothers (DMS397) into a multiple regression analysis. Our interest is in identifying some factors that may relate with the percentage of births to teenagers, particularly those frequently forwarded in the popular press. We will test the following hypotheses:

Hypothesis 1: Percentage of births attributed to teenage mothers is positively associated with poverty.

Independent Variable: PVS519

Hypothesis 2: Percentage of births attributed to teenage mothers is negatively associated with per capita spending for education.

Independent Variable: EDS137

Hypothesis 3: Percentage of births attributed to teenage mothers is positively associated with the amount of welfare (TANF)[19] families receive.

Independent Variable: PVS546

Hypothesis 4: Percentage of births attributed to teenage mothers is positively associated with the percent of the population that is African American

Independent Variable: DMS468

[19] TANF stands for "Temporary Assistance to Needy Families," the most commonly understood program equated with "welfare."

To run this regression, and produce the output reproduced in Figure 7.6, use the following commands:

Analyze

Regression

Linear

Dependent: DMS397 (Births to Teenage Mothers as Percentage of All Births: 2007)

Independents: PVS519 (Poverty Rate: 2008)
EDS137 (Per Capita Spending for Education: 2007)
PVS546 (Average Monthly TANF Assistance per Family: 2007)
DMS468 (Percent of Population Black: 2008)

OK

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.892[a]	.796	.778	1.1902

a. Predictors: (Constant), DMS468 Percent of Population Black: 2008, EDS137 Per Capita State and Local Govt. Spending for Education: 2007, PVS546 Average Monthly TANF Assistance per family: 2007, PVS519 Poverty Rate: 2008

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	254.517	4	63.629	44.915	.000[a]
	Residual	65.166	46	1.417		
	Total	319.684	50			

a. Predictors: (Constant), DMS468 Percent of Population Black: 2008, EDS137 Per Capita State and Local Govt. Spending for Education: 2007, PVS546 Average Monthly TANF Assistance per family: 2007, PVS519 Poverty Rate: 2008

b. Dependent Variable: DMS397 Births to Teenage Mothers as Percent of All Births: 2007

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.176	1.885		1.685	.099
	PVS519 Poverty Rate: 2008	.590	.074	.714	7.956	.000
	EDS137 Per Capita State and Local Govt. Spending for Education: 2007	.001	.000	.136	1.775	.082
	PVS546 Average Monthly TANF Assistance per family: 2007	-.008	.002	-.341	-4.133	.000
	DMS468 Percent of Population Black: 2008	.000	.017	.001	.008	.994

a. Dependent Variable: DMS397 Births to Teenage Mothers as Percent of All Births: 2007

Figure 7.6 Multiple Regression Output

Interpreting Multiple Linear Regression Coefficients

As in the analysis of bivariate regressions, we approach this output with four questions. First, do these independent variables, together, predict the values of the dependent variable better than the mean? If so, what are the natures of the relationships? Third, are the relationships statistically significant? And last, how powerful is the model in explaining the variation in teen birth rates within the United States?

The first thing we look at is the ANOVA table. The p-value for the Regression model F-test is .000. The model is highly significant, and we can conclude that these four independent variables together predict the percentage of birth attributed to teenage mothers. But do they *all* uniquely predict? If not, which ones do? And are the unique relationships in the direction we hypothesized? To answer these questions, we turn to the Coefficients table.

Hypothesis 1 is supported, re-establishing the relationship identified in the previous bivariate analysis - states that have higher poverty rates tend to have higher percentages of births to teenage mothers. The regression coefficient is positive, (.590) indicating that the more poverty, the higher the percent of births to teens, and the relationship is statistically significant (Sig.=.000). You may notice that the coefficient is smaller than it was in the bivariate regression model (.698). This is the result of the multiple variable model documenting the *unique* effect of poverty rates on teenage births, after accounting for the other variables in the model.

Hypothesis 2 predicts that the more a state spends per capita on education, the smaller the percentage of births will be to teen mothers. Although the regression coefficient is positive (.001), the relationship is not statistically significant (Sig. = .082). Here we find no support for a commonly espoused liberal thesis that allocating more money into education will necessarily result in discouraging teen births.

Hypothesis 3 predicts that the more a state spends on welfare per recipient family, the higher the percentage of births to teenage mothers (this hypothesis tests the conservative thesis that welfare encourages irresponsible behavior). This relationship is statistically significant (Sig.= .000). However, the negative regression coefficient (-.008) shows a relationship *opposite* the one predicted in the hypothesis. The more a state spends on welfare per recipient family *the lower the percent of births attributed to teenage mothers*.

Hypothesis 4 predicts that the greater the proportion of the population that is African American, the higher the percent of births attributed to teenage mothers. If you are curious, try a bivariate linear regression and you will indeed find a statistically significant positive relationship between these two factors (Pearson Correlation = .44). However, when we enter the other variables in a multiple variable model, the regression coefficient is 0 (.000), and it is not significant!

How does this happen? This is a nice illustration of the importance of including control variables in a model. This model suggests that once issues such as poverty and spending on public assistance to the poor are taken into account, the impact of race disappears! This means that teen births may have less to do with the issue of race than it does with issues of poverty and aid to the poor.

Finally, we turn to the question of model strength. In multiple variable regressions, the **adjusted R-square** statistic is used instead of the R-square because adding even unrelated independent variables to a model will raise the R-square statistic. The adjusted R-square statistic compensates for the number of variables in the model and it will only increase if added variables contribute significantly to the model. The adjusted R-square is often used to compare which of several models is best. How good is the model? The Adjusted R-square statistic in the Model

Summary Table means that 77.8% (Adj R-square=.778) of the variation in the teenage birth rate can be attributed to these four variables! This is an excellent model. In fact, it is rare to find a model for social behavior that has such a high explanatory power.

In sum, the model suggests that states that have higher proportions of births attributed to teenage mothers tend to also have higher levels of poverty and extend lower supports to the poor. It also suggests that per capita educational expenditure has no observable effect on the proportion of births attributed to teenage mothers. We also found that the relationship between race and teen births may be spurious, and this apparent relationship disappears when poverty rates are taken into account.

However, it must be emphasized our intent here is not to offer a comprehensive analysis of this challenging research question, but rather to use the question to illustrate a statistical technique. Certainly, additional research is warranted to focus on a variety of related questions concerning causality (does poverty cause increased teen births, or do teen births influence the poverty rates?), levels of analysis (should the analysis be on state level comparisons as we do here, or on individuals?), and measurement (even if overall education funding has no effect on stemming teen births, it may matter how funds are spent).

Graphing a Multiple Regression

Graphing relationships from multiple regressions is more complex than graphing relationships from bivariate regressions, although the approach is the same. Because there are many variables in the multiple variable models, the two-dimensional graphs need to control for the effects of other variables.

To graph a regression line from a multiple variable model requires selecting one independent variable to go on the X axis. The rest of the variables in the model will be held constant. To hold these values as constants, any value could be chosen, but the most common choice is the mean (which we generate using *Descriptives*) for scale variables and the mode for categorical variables. Again, we will use the regression formula, now expanding it to include the other variables in the model.

$$\hat{Y} = A + B_1(X_1) + B_2(X_2) + B_3(X_3) + B_4(X_4)$$

$\hat{Y}$ = Predicted Value of the dependent variable
A = Constant

B_1 = Slope of Variable 1	X_1 = Chosen value of Variable 1
B_2 = Slope of Variable 2	X_2 = Chosen value of Variable 2
B_3 = Slope of Variable 3	X_3 = Chosen value of Variable 3
B_4 = Slope of Variable 4	X_4 = Chosen value of Variable 4

This example will show how to graph the association of welfare benefits and percentage of births to teenage mothers, holding poverty rates, school expenditures and percent African-American population at their means. This requires computing the predicted values of percent of births to teenage mothers based on values of PVS546.

First graph the relationship between DMS397 and PVS546 using the *Scatter/Dot* functions in the *Graph Chart Builder* (Figure 7.7).

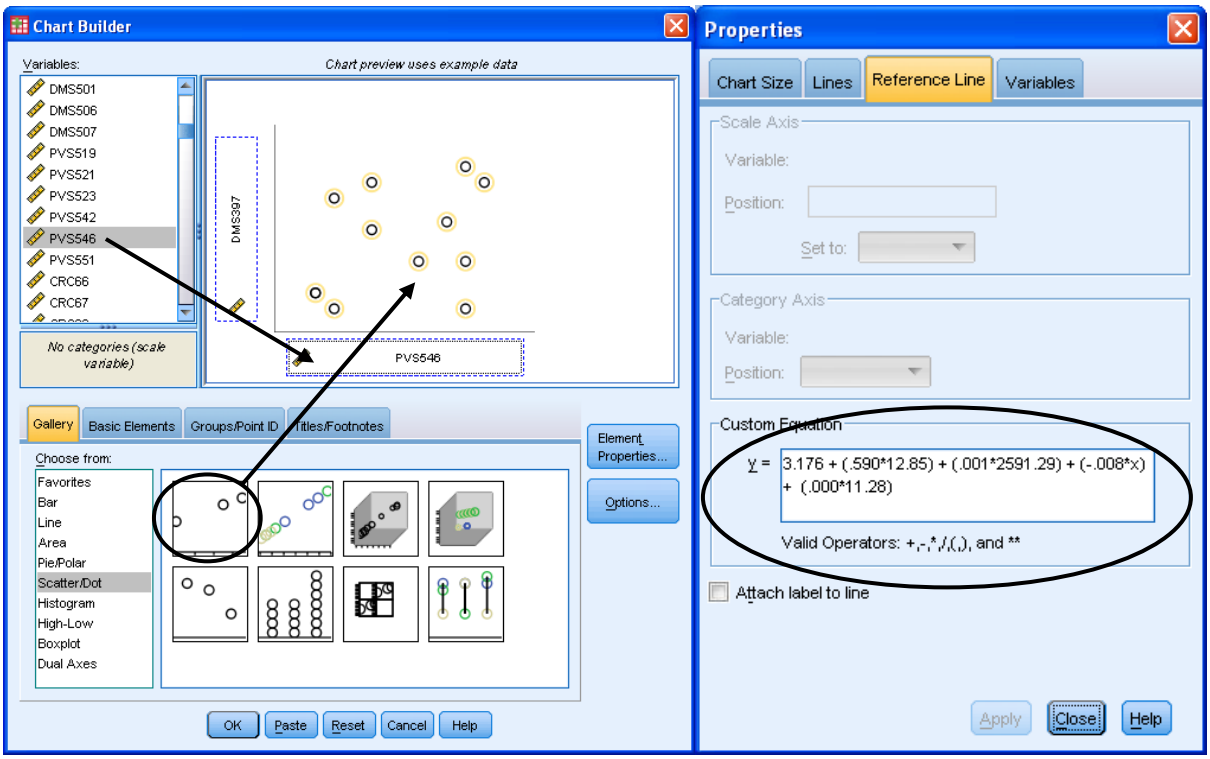

Figure 7.7 Scatter Plot Dialog Box and Reference Line Window from the *Chart Editor*

Double-click on the scatterplot to invoke the *Chart Editor*. Once again, choose:

Options

Reference Line from Equation

Reference Line: Custom Equation: 3.176 + (.590*12.85) + (.001*2591.29) + (-.008**x*) + (.000*11.28)

Apply

Close the *Chart Editor*

Sources of numbers in the above equation:

Constant (A) = 3.176

Variable	B	Mean Value
PVS519	.590	12.85
EDS137	.001	2591.29
PVS546	.008	325.49
DMS468	.000	11.28

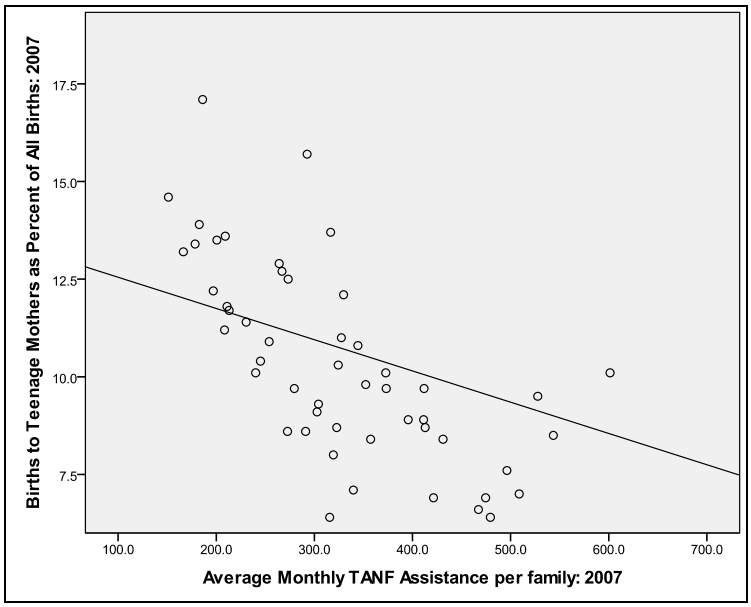

Figure 7.8 Scatter Plot of PVS546 and Predicted Regression Line

Figure 7.8 shows the scatterplot with the multiple regression line. The regression line will often look a little "off" because the predicted values on the line are adjusted for the other variables in the model.

Other Concerns In Applying Linear Regression

Just like ANOVA, linear regression has assumptions concerning the origin and structure of the dependent variable. Linear regression results are only meaningful if these assumptions have been met. Linear regression assumes that the residuals follow a normal distribution with constant standard deviation, as outlined below.

Residuals

The coefficients and significance values that result from regression analysis are calculated under the assumption that a straight line is a good model for the relationship. How well a line serves as a model for the relationship can be checked by looking at how the actual observations sit in relation to the predicted values along the line. This is measured by the vertical distance from the prediction to the actual observation and is called the **residual** (illustrated with the dashed lines in Figure 7.9). Regression coefficients are calculated so that the resulting line has the lowest possible accumulation of residuals, minimizing the overall distance between the observations and the predictions.

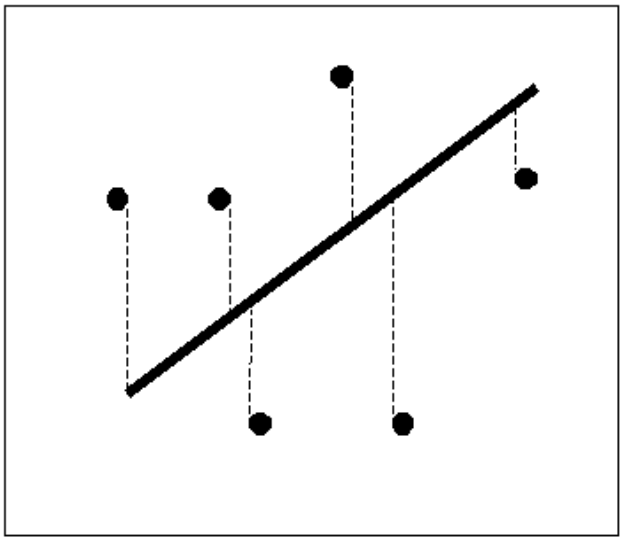

Figure 7.9 Data Points, a Regression Line and Residuals (dashed lines)

Constant Variation

Sometimes the regression line is not well suited to the data, as in Figure 7.10, which shows a "fan effect." This regression line is much better at predicting low values of the dependent variable than it is at predicting high values. It is therefore not an appropriate model, regardless of the strength of the coefficients or their significance values. In a well-fitting regression model, the variation around the line is constant all along the line (Figure 7.5 is a good example).

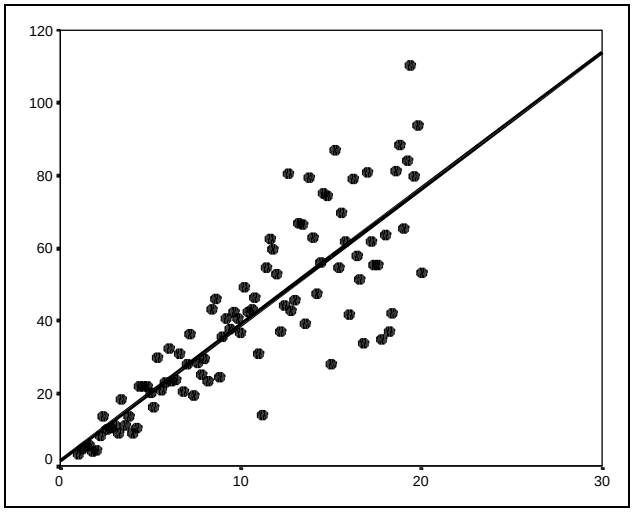

Figure 7.10 Scatter Plot of a Fan Effect

Normality of Residuals

The residuals should follow a normal distribution, with a mean of 0. Recall that a normal distribution is shaped like a bell—it is symmetric, and most points are in the middle, with fewer and fewer farther from the mean. Since the residuals measure where the points fall in relation to the line, a symmetric distribution of the residuals indicates that the same number of points fall above and below the line. Since a residual of 0 means a point is right on the line, a mean of 0 indicates the line is in the middle of the points. Once again, some are above and some are below. And the bell shape means that most are close to the line, and there are fewer points farther from the line.

One way to check the normality of residuals is to save and plot residuals from the regression command. To do so for our multiple variable example:

Analyze
Regression
Linear
Dependent: DMS397 (Births to Teenage Mothers as Percentage of All Births: 2007)
Independents: PVS519 (Poverty Rate: 2008)
EDS137 (Per Capita Spending on Education: 2007)
PVS546 (Maximum Monthly TANF Benefit for Family of Three in 2007)
DMS468 (Percent of Population Black: 2008)
Save
Residuals: Check *Unstandardized*
Continue
OK

This will create a new variable RES_1, which can be graphed using the Graph command (Figure 7.11).

Graphs
Chart Builder
Gallery: Histogram
Drag *Simple Histogram* to *Chart preview*
Drag *RES_1* from *Variables* list to *X-axis?* box [Hint:RES_1 will be at the bottom of the list!]
Element Properties
Display Normal Curve
Apply
OK

As Figure 7.11 shows, the residuals form a reasonably normal distribution, which is a good indication that the regression is working well.

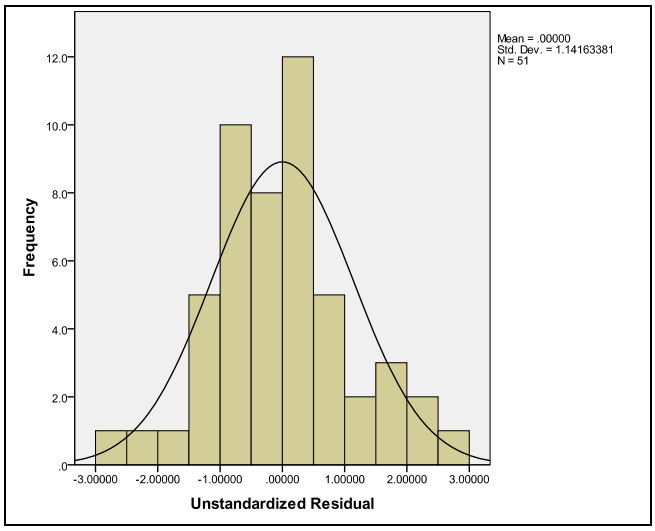

Figure 7.11 Histogram of Normally Distributed Residuals

Building Multiple Variable Models

How does a researcher decide which independent variables to include in a model? One seemingly efficient way is to load all of the variables in the data set into the model and see which create significant findings. Although this approach seems appealing (albeit lazy), it suffers from a number of problems.

Degrees of Freedom

The first problem is degrees of freedom. **Degrees of freedom** refer to the number of observations in a sample that are "free to vary." Every observation increases degrees of freedom by one, but every coefficient the model estimates (including the constant) decreases the degrees of freedom by one. Because every independent variable in a regression model lowers the degrees of freedom, it reduces the test's ability to find significant effects. Therefore, it is in the researcher's interest to be highly selective in choosing which variables to include in a model.

The following strategies can make decisions easier. First, researchers should select only variables for which there is a theoretical basis for inclusion. Then they should explore the data with univariate and bivariate analyses, and only include variables that have potentially informative results, or which are needed to serve as controls.

In large models, we also suggest introducing new variables sequentially. Rather than including all of the variables at once, start by introducing a small group of variables. As you enter more variables in the model, observe not only how they operate, but also how the coefficients and significance scores change for the other variables as well. If the model is statistically stable, these values will tend to remain in the same general range and the relationships will retain a consistent directionality. If values start fluctuating in a remarkable manner, this suggests that the findings should not be trusted until further exploratory work is performed to determine the reasons for these fluctuations.

Collinearity

Collinearity occurs when two or more independent variables contain strongly redundant information. If variables are collinear, there is not enough distinct information in these variables

for the multiple regression to operate correctly. A multiple regression with two or more independent variables that measure essentially the same thing will produce errant results. An example is the poverty rate (PVS519) and the percent of children living in poverty (PVS521). These are two different variables, but they are so strongly collinear (correlation of .983) that they are nearly indistinguishable in the regression equation.

Collinear variables create peculiar regression results. For example, the two correlated poverty variables have significant bivariate relationships with teenage birthrate. However, if we run a multiple regression (you can try this) of PVS519 and PVS521 with the percents of births attributed to teenage mothers (DMS397), *both* variables become *insignificant*. Unlike the issue of spurious relationships, here the disappearance of the relationship is not explained away, but is rather the result of a mathematical corruption of the model.

The implication is that researchers should be careful about putting highly correlated variables into regression equations. To check for collinearity, start by examining a correlation matrix that compares all independent variables with each other.[20] A correlation coefficient above .80 is an indicator that collinearity *might* be present. If it is, variables may need to be analyzed in separate regression equations, and then how they operate together when included in the same model. It is only fair to mention, however, that collinearity this extreme is quite rare.

Dummy Variables

Recall that the coefficients for independent variables reflect incremental differences in the dependent variables. This means the values of independent variables need to have meaningful increments. This works easily for scale variables, but not categorical ones. Categorical independent variables, such as gender or race, can be made to have meaningful increments through **dummy coding** (also called indicator coding). Each dummy variable splits the observations into two categories. One category is coded "1" and the other "0." With this coding, a one-unit difference is always the difference between 1 and 0, the difference of being in one category compared to the other. The coefficient is then interpreted as the average difference in the dependent variable between the two categories. Dummy variables are always coded 1 and 0. Although values of "1" and "2" seem logical and are also one unit apart, there are mathematical conveniences that make "1" and "0" preferable.

Dummy coding is easy for a variable with two categories, such as employment status: (employed vs. unemployed), or gender (men vs. women). Consider this example. Suppose we wanted to assess who has saved more for retirement, employed people or unemployed people. Since we are comparing two groups, the employed and the unemployed, employed would be coded as 1 and unemployed would be coded as 0. If the regression coefficient for employment status were 4,600, it would mean that in general, employed people had saved an average of $4,600 more than unemployed people.

Some categorical variables have more than two categories, such as race, or geographic location (East, West, North, South). To handle this situation, each group except one needs a dummy variable. To dummy code the variable "RACE" with "White" as the excluded category, the following variables would be necessary: the variable "BLACK" would be coded 0/1, with all African American respondents coded as 1 and all other respondents with known ethnicity coded 0. Likewise, a new variable "HISPANIC" would be coded 0/1 with all Hispanic respondents

[20] While high correlations among independent variables can indicate collinearity, it can also miss it. It is possible to have collinearity without high correlations among independent variables and vice-versa. For a more thorough check for collinearity, use SPSS's Collinearity Diagnostics, available under "Statistics" in the Regression command.

coded 1 and all others coded 0. We would do this for all but one racial group called the **reference group** that has the value of 0 in all of the computed dummy variables. Then all of these new dummy variables would be included in the model.

In this example, we would exclude Whites as the reference group, and the regression output would allow us to compare other ethnicities to Whites.[21] For example, if we were predicting how much money people saved for retirement and the variable BLACK had a coefficient –3,200, it would mean that African Americans save, on average, $3,200 less than Whites. If the variable "ASIAN" had a coefficient of 6,700, it would mean that Asians save $6,700 more for retirement than Whites. In each case, the dummy variable is interpreted relative to the reference group in the regression.

Outliers

Linear regressions can be greatly influenced by **outliers**—atypical cases. Outliers can "pull" the equation away from the general pattern, and unduly sway the regression output. But what is the appropriate way to deal with an outlier?

In some situations it is reasonable to simply delete an outlier from the analysis. For example, perhaps the outlier was a mistake – a data entry or data recording error. Another example is when there are special circumstances surrounding a specific case. For example, Nevada has a very high divorce rate. Because of its laws, it attracts people from outside the state for "drive-by divorces." Because Nevada's divorce rate does not accurately measure the rate of divorce for Nevada residents, as do the divorce rates in other states, it may be reasonable to completely remove Nevada from any analysis of divorce rates.

In other situations, outliers should remain in the analyses, as an outlier may be the most important observation. In these circumstances, it is a good idea to run the analysis twice, first with the outlier in the regression and second with it excluded. If the outlier is not exerting an undue influence on the outcomes, both models should reasonably coincide. If, however, the results are vastly different, it is best to include both results in the text of the report.

Causality

Finally, researchers should use great caution in interpreting the outcomes of linear regressions as establishing causal relationships. As discussed in Chapter 1, three things are necessary to establish causality: association, time order, and nonspuriousness. While regressions of cross sectional data can reveal associations, they usually do not document time order. Note how we were careful to say that poverty is associated with teen births, but did not assert that it causes them. (Although this might be true, we just don't have enough evidence to claim it).

One of the strengths of multiple linear regressions is that researchers can include factors (if they are available) that can control for spurious effects. However, there always remains the possibility that a spurious factor remains untested. Even though multiple variables may be included in the statistical model, it is still possible to have spurious relationships if important

[21] In deciding which group to omit, it is important that there be a sufficient number of cases in the data set to allow a meaningful comparison (e.g., we would not select Whites as a comparison group if there were only a few represented in the data). It is also important that the reference group be useful in its potential for meaningful comparisons. For this reason, we chose not to use "Other" as a comparison group because it does not represent a cohesive category of observations, but rather a mish-mash of racial/ethnic groups that don't fit into the larger categories present in the data set.

variables are left out. Only a large body of research would be able to account for enough factors that researchers could comfortably conclude causality.

Summary

Multiple linear regression has many advantages, as researchers can examine the multiple factors that contribute to social experiences and control for the influence of spurious effects. They also allow us to create refined graphs of relationships through regression lines. These can be a straightforward and accessible way of presenting results.

Knowing how to interpret linear regression coefficients allows researchers to understand both the direction of a relationship (whether one variable is associated with an increase or a decrease in another variable) and strength (how much of a difference in the dependent variable is associated with a measured difference in the independent variable).

Knowing about the F-test and R-square helps researchers understand the explanatory power of statistical models. As with other statistical measures, the significance tests in regressions address the concern of random variation and the degree to which it is a possible explanation for the observed relationships.

As regressions are complex, care is needed in performing them. Researchers need to examine the variables and construct them in forms that are amenable to this approach, such as creating dummy variables. They also need to examine findings carefully and test for concerns such as collinearity or patterns among residuals. This being said, linear regressions are quite forgiving of minor breaches of these assumptions and can produce some of the most useful information on the relationships between variables.

Key Terms

Adjusted R-Square
B
Coefficient
Collective effects
Collinearity
Constant
Constant variance
Control variables
Degrees of freedom
Dummy variables
Explanatory power
Intercept
Linear relationship
Normality of residuals
Outliers
Reference group
Residuals
R-square
Slope
Spurious factors
Unstandardized coefficient

Chapter 7 Exercises

Name_____ Date_____

1. Using the STATES data, test the hypothesis that states with large African American populations receive lower educational funding than predominantly White states. Test this hypothesis by performing a bivariate linear regression on EDS137 (Per Capita State and Local Gvt Spending for Education: 2007) and DMS468 (Percent of Population Black: 2008). Fill in the following statistics:

Regression Coefficient B for DMS468 _____

Significance level _____

Is the relationship significant? Yes No

R Square _____

In your own words, describe the relationship between EDS137 and DMS468. How would you explain these findings?

2. Using the STATES data set, examine the relationship between states' rate of U.S. Military Fatalities in Iraq and Afghanistan as of January 2010 (DFS90) and their 2006 Public High School Graduation Rate (EDS131).

Regression Coefficient B for EDS131 _____

Significance _____

Is the relationship significant? Yes No

R Square _____

In your own words, describe the relationship between DFS90 and EDS131. How would you explain these findings?

3. Perform a regression on the relationship between the property crime rate (CRS48) as predicted by the percent of the population of a state that is living below the poverty level (PVS519).

Regression Coefficient B for PVS519 _____

Significance _____

Is the relationship significant? Yes No

R Square _____

In your own words, describe the relationship between CRS48 and PVS519. How would you explain these findings?

4. Perform a multiple regression to predict the crime rate (CRS31). Include as indicators of predictors the homeownership rate (ECS445), the divorce rate (DMS506), and the personal bankruptcy rate (ECS105).

Constant _____

Adjusted R Square _____

Regression Coefficient B for ECS445 _____

Significance _____

Is the relationship significant? Yes No

Regression Coefficient B for DMS506 _____

Significance _____

Is the relationship significant? Yes No

Regression Coefficient B for ECS105 _____

Significance _____

Is the relationship significant? Yes No

In your own words, describe these relationships and why they might be (or are not) associated with crime rates.

5. Using the output from the regression in Exercise 4, write the formula which would be used to generate a line showing the association of the divorce rate with the crime rate. Hold constant ECS445 and ECS105 at their mean values. You will need to generate the mean values using the *Descriptive Statistics - Descriptives* command.

$$\hat{Y} = A + B_1(X_1) + B_2(X_2) + B_3(X_3)$$

6. Create and print a scatterplot between CRS31 (*Y-axis*) and DMS506 (*X-axis*). Using the *Reference Line from Equation* option in the *Chart Editor*, add the predicted regression line for DMS506, holding constant ECS445 and ECS105 at their mean values. [Hint: See the instructions for creating Figure 7.8].

7. Using the output from the previous regression in Exercise 4, write the formula which would be used to generate a line showing the association of the home ownership with the crime rate. Hold constant DMS506 and ECS105 at their mean values. You will need to generate the mean values using the *Descriptive Statistics - Descriptives* command.

$$\hat{Y} = A + B_1(X_1) + B_2(X_2) + B_3(X_3)$$

8. Create and print a scatterplot between CRS31 (*Y-axis*) and ECS445 (*X-axis*). Using the *Reference Line from Equation* option in the *Chart Editor*, add the predicted regression line for ECS445, holding constant DMS506 and ECS105 at their mean values. [Hint: See the instructions for creating Figure 7.8].