Chapter sixteen

Multivariate Regression

Specification: building a multivariate model

What is a Correctly Specified Model?

How Does the Researcher Try to Correctly Specify the Model?

Model building

How Can You Include Nominal and Ordinal Variables in a Regression Model?

How Do You Compare Regression Coefficients?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_16,

One of the statistical tools most commonly used in criminal justice and criminology is regression modeling. A regression model allows the researcher to take a broad approach to criminological research problems. It is based not simply on understanding the relationships among variables, but on specifying why changes occur and what factors are directly responsible for these changes. In a regression model, the researcher tries to disentangle the various potential factors that have an impact on the dependent variable, in order to provide an accurate picture of which variables are in fact most important in causing change.

In this chapter, we discuss why it is generally necessary to take into account more than just one independent variable in building a regression model. Previous chapters have focused on bivariate statistical analysis, in which we relate two variables—nominal, ordinal, or interval—to each other. This chapter introduces multivariate analysis, in which the researcher takes into account a series of independent variables within one statistical model.

The Importance of Correct Model Specifications

The most important assumption we make in regression modeling is that the model we have estimated is specified correctly. A **correctly specified regression model** is one in which the researcher has taken into account all of the relevant predictors of the dependent variable and has measured them accurately. This requirement of regression modeling is the most difficult one that researchers face. Its importance is linked both to prediction of the dependent variable and to correct estimation of regression coefficients.

Errors in Prediction

Predictions of Y in regression are based on the factors that are included in a regression model. So far, we have examined bivariate regression models, in which one independent variable is used to predict values of Y. But in the real world it is unlikely that only one variable will influence the dependent measure you are examining. Most often, it will be necessary to take into account a number of independent variables. Regression

analysis that takes into account more than one independent variable is called **multivariate regression** analysis. The regression model we have discussed so far can be extended to the multivariate case simply by adding a term for each new variable. For example, to include years of education in the model predicting number of arrests presented earlier, we would express our regression equation as follows:

$$Y_{\text{arrests}} = b_0 + b_1(\text{age}) + b_2(\text{education}) + e$$

The population model for this equation would be written as

$$Y_{\text{arrests}} = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{education}) + \epsilon$$

Sometimes, when examining multiple independent variables, researchers find it tedious to include the names of the variables in subscripts. Accordingly, they will often use a general form of the regression equation and then define each variable in a table or in their description of results. For example, the above equation could be expressed in terms of the population parameters as

Model 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

where Y = arrests
X_1 = age
X_2 = years of education

In theory, you could define all relevant predictors of Y and include them all in your regression model. This correctly specified model would also provide the most accurate predictions of Y. Conversely, a misspecified model, or one that does not include all relevant predictors, will provide **biased** predictions of Y.

Let's say, for example, that family median income is also an important predictor of arrests. In this case, the corrected population regression equation would be written as follows:

Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

where Y = arrests
X_1 = age
X_2 = years of education
X_3 = family median income

By adding this additional variable, we improve our predictions of Y over those provided by Model 1. Because we have taken into account the influence of family income on arrests, we have added to our ability to correctly predict the dependent variable. By implication, our predictions of Y will be less trustworthy when we do not include a factor that influences the dependent variable.

Sometimes, statisticians express this fact in terms of an assumption about the error term in the population regression model. The error term, ϵ, should represent only random fluctuations that are related to the outcomes (Y) that you are examining. For this reason, we also call the errors residuals, since they are in theory what is left over once you have taken into account all systematic causes of Y. However, if you fail to include an important predictor of Y as an independent variable, then by implication it moves to your error term. The error term now is not made up only of random—or what statisticians sometimes call stochastic—variation in Y, but rather includes the systematic variation that can be attributed to the excluded variable. For example, in Model 1, the effect of median family income is not taken into account, and thus the systematic relationship between median family income and number of arrests is found in the error term for that regression equation. When a model is not correctly specified, the error term will not represent only random or stochastic variation, as is required by the assumptions of regression analysis; it will be systematically related to the dependent variable.

Correctly Estimating the Effect of *b*

Failure to correctly specify a regression model may also lead the researcher to present biased estimates of the effects of specific independent variables. Suppose, for example, that a bivariate regression is defined in which number of years in prison is identified as influencing number of arrests after prison:

$$Y_{\text{rearrests}} = b_0 + b_1(\text{years in prison}) + e$$

In estimating this relationship from the data presented in Table 16.1, we find that the regression coefficient based on this model is 1.709. That is, every additional year of imprisonment produces about a 1.709 increase in our prediction of number of subsequent arrests.

Working It Out

$$b = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

$$= \frac{31.7}{18.55}$$

$$= 1.7089$$

Our model for subsequent arrests states that the only causal factor influencing arrests is years of imprisonment. This, of course, is a questionable statement, because common sense tells us that this model is not

Table 16.1 Number of Rearrests (Y) and Years Spent in Prison (X) for 20 Former Inmates

SUBJECT	REARRESTS Y	$Y_i - \bar{Y}$	YEARS SPENT IN PRISON X	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	0	−3.1	2	−1.15	1.3225	3.565
2	0	−3.1	3	−0.15	0.0225	0.465
3	1	−2.1	1	−2.15	4.6225	4.515
4	1	−2.1	2	−1.15	1.3225	2.415
5	1	−2.1	3	−0.15	0.0225	0.315
6	1	−2.1	3	−0.15	0.0225	0.315
7	2	−1.1	4	0.85	0.7225	−0.935
8	2	−1.1	2	−1.15	1.3225	1.265
9	2	−1.1	2	−1.15	1.3225	1.265
10	3	−0.1	3	−0.15	0.0225	0.015
11	3	−0.1	3	−0.15	0.0225	0.015
12	3	−0.1	3	−0.15	0.0225	0.015
13	4	0.9	3	−0.15	0.0225	−0.135
14	4	0.9	4	0.85	0.7225	0.765
15	4	0.9	4	0.85	0.7225	0.765
16	4	0.9	4	0.85	0.7225	0.765
17	5	1.9	4	0.85	0.7225	1.615
18	6	2.9	4	0.85	0.7225	2.465
19	7	3.9	5	1.85	3.4225	7.215
20	9	5.9	4	0.85	0.7225	5.015
	$\bar{Y} = 3.1$	$\bar{X} = 3.15$			$\sum_{i=1}^{N}(X_i - \bar{X})^2 = 18.55$	$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = 31.7$

Bivariate Regression Model:
Dependent Variable: Subsequent Rearrests
Independent Variable: Years in Prison
Regression Coefficient: b(years in prison) = 31.7/18.55 = 1.7089

correctly specified. There are certainly other factors that influence arrests. Some of those factors, in turn, may also be related to the number of years that an offender serves in prison. If this is true—that relevant factors related to years of imprisonment have been omitted from the model—then the regression coefficient may provide a very misleading estimate of the effect of imprisonment on arrests.

Judges, for example, are likely to impose longer prison sentences on offenders with more serious prior records. Using the sample data in Table 16.2, we can look at the correlations among these three variables (see Table 16.3). The number of prior arrests is strongly related (r = 0.63) to the length of prison term served. Prior arrests are even more strongly related to subsequent arrests (r = 0.76). This suggests, first of all, that prior record is a relevant factor that should be included if our model is to be correctly specified. But it also raises a very important concern: How do we know that our finding that "years in prison" increases reoffending is not simply a result of the fact that those who serve longer prison terms generally have more serious prior records of offending?

Table 16.2

Number of Rearrests, Years Spent in Prison, and Number of Prior Arrests for 20 Former Inmates

SUBJECT	REARRESTS	YEARS IN PRISON	PRIOR ARRESTS
1	0	2	4
2	0	3	2
3	1	1	2
4	1	2	3
5	1	3	3
6	1	3	2
7	2	4	3
8	2	2	3
9	2	2	1
10	3	3	2
11	3	3	3
12	3	3	3
13	4	3	4
14	4	4	3
15	4	4	4
16	4	4	5
17	5	4	4
18	6	4	5
19	7	5	5
20	9	4	6
	$\overline{Y} = 3.10$	$\overline{X} = 3.15$	$\overline{X} = 3.35$
	$s = 2.300$	$s = 0.9631$	$s = 1.2360$

In an ideal world, our comparisons of the impact of imprisonment would be made with subjects who were otherwise similar. That is, we would want to be sure that the offenders with longer and shorter prison sentences were comparable on other characteristics, such as the seriousness of prior records. In this case, there would be no relationship between prior arrests and length of imprisonment, and thus we would not have to be concerned with the possibility that the effect of length of imprisonment actually reflects the influence of prior arrests on reoffending.

In criminal justice, this approach is taken in the development of **randomized experiments.**[1] A randomized study of the impact of

Table 16.3

Correlation Coefficients for the Variables Years in Prison, Prior Arrests, and Subsequent Rearrests Based on Data from 20 Former Inmates

	YEARS IN PRISON	PRIOR ARRESTS
Prior Arrests	$r = 0.6280$	
Subsequent Rearrests	$r = 0.7156$	$r = 0.7616$

[1]For a discussion of experimental methods in criminal justice, see E. Babbie and M. Maxfield, *The Practice of Social Research in Criminal Justice* (Belmont, CA: Wadsworth, 1995). For a comparison of experimental and nonexperimental methods, see D. Weisburd, C. Lum, and A. Petrosino, "Does Research Design Affect Study Outcomes in Criminal Justice?" *The Annals* 578 (2001): 50–70.

length of imprisonment on reoffending would be one in which the researcher took a sample of offenders and assigned them to treatment and control conditions at random. For example, the researcher might define a sentence of 6 months as a control condition and a sentence of 1 year as an experimental condition. In this case, the researcher could examine the effects of a longer versus a shorter prison sentence on rearrests without concern about the confounding influences of other variables. In Chapter 21, we focus more directly on the analysis of experimental data. But it is important to note here that random allocation of subjects to treatment and control conditions allows the researcher to assume that other traits, such as prior record, are randomly scattered across the treatment and control conditions. Our problem in criminal justice is that it is often impractical to develop experimental research designs. For example, it is highly unlikely that judges would allow a researcher to randomly allocate prison sanctions. The same is true for many other research problems relating to crime and justice.

Fortunately for criminal justice researchers, a correctly specified regression model will take into account and control for relationships that exist among the independent variables included in the model. So, for example, the inclusion of both length of imprisonment and prior arrests in one regression model will provide regression coefficients that reflect the specific impact of each variable, once the impact of the other has been taken into account. This is illustrated in Equation 16.1, which describes the calculation of a multivariate regression coefficient in the case of two independent variables (X_1 and X_2). Equation 16.2 applies Equation 16.1 to the specific regression model including both length of imprisonment and prior arrests. The model can be described as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + e$$

where Y = subsequent rearrests
X_1 = years in prison
X_2 = prior arrests

Here we calculate the multivariate regression coefficient b_1 for length of imprisonment.

$$b_{X_1} = \left(\frac{r_{Y,X_1} - (r_{Y,X_2}r_{X_1,X_2})}{1 - r^2_{X_1,X_2}}\right)\left(\frac{s_Y}{s_{X_1}}\right) \qquad \textbf{Equation 16.1}$$

$$b_{X_1} = \left(\frac{r_{Y,YP} - (r_{Y,PA}r_{YP,PA})}{1 - r^2_{YP,PA}}\right)\left(\frac{s_Y}{s_{YP}}\right) \qquad \textbf{Equation 16.2}$$

In Equations 16.1 and 16.2, the bivariate correlations among the three measures examined, as well as the standard deviations of years in

prison and rearrests, are used to calculate the multivariate regression coefficients. The three correlations for our specific example are (1) $r_{Y,YP}$, or the correlation between subsequent rearrests and years in prison; (2) $r_{Y,PA}$, or the correlation between subsequent rearrests and prior arrests; and (3) $r_{YP,PA}$, or the correlation between years in prison and prior arrests.

What is most important to note in Equation 16.2 is that the numerator (in the first part) takes into account the product of the relationship between prior arrests and subsequent rearrests and that of prior arrests and years in prison. This relationship is subtracted from the simple correlation between years in prison and subsequent arrests. In this way, multivariate regression provides an estimate of b that takes into account that some of the impact of years in prison may be due to the fact that longer prison terms are associated with more serious prior records. This estimate is now purged of the bias that was introduced when prior record was not included in the regression model. The multivariate regression coefficient for years in prison when prior record is included in the regression model (0.936) is considerably smaller than the estimate calculated earlier in the bivariate regression (1.709).

Working It Out

$$b_{X_1} = \left(\frac{r_{Y,YP} - (r_{Y,PA}r_{YP,PA})}{1 - r^2_{YP,PA}}\right)\left(\frac{s_Y}{s_{YP}}\right)$$

$$= \left(\frac{0.7156 - (0.7616)(0.6280)}{1 - (0.6280)^2}\right)\left(\frac{2.300}{0.9631}\right)$$

$$= \left(\frac{0.2373152}{0.605616}\right)(2.388122)$$

$$= 0.9358$$

With the same information, we can calculate the multivariate regression coefficient for prior arrests. We find the value for b_2 to be 0.9593 (see working it out, page 490). The bivariate regression coefficient for prior arrests is calculated in the box on page 489 so that you can compare the results. As you can see, the value of b when we take into account years in prison (0.96) is much smaller than that in the bivariate case (1.4).

Calculation of Bivariate Regression Coefficient for Number of Rearrests (Y) and Number of Prior Arrests (X) for 20 Former Inmates

SUBJECT	REARRESTS Y	$Y_i - \bar{Y}$	PRIOR ARRESTS X	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	0	−3.1	4	0.65	0.4225	−2.015
2	0	−3.1	2	−1.35	1.8225	4.185
3	1	−2.1	2	−1.35	1.8225	2.835
4	1	−2.1	3	−0.35	0.1225	0.735
5	1	−2.1	3	−0.35	0.1225	0.735
6	1	−2.1	2	−1.35	1.8225	2.835
7	2	−1.1	3	−0.35	0.1225	0.385
8	2	−1.1	3	−0.35	0.1225	0.385
9	2	−1.1	1	−2.35	5.5225	2.585
10	3	−0.1	2	−1.35	1.8225	0.135
11	3	−0.1	3	−0.35	0.1225	0.035
12	3	−0.1	3	−0.35	0.1225	0.035
13	4	0.9	4	0.65	0.4225	0.585
14	4	0.9	3	−0.35	0.1225	−0.315
15	4	0.9	4	0.65	0.4225	0.585
16	4	0.9	5	1.65	2.7225	1.485
17	5	1.9	4	0.65	0.4225	1.235
18	6	2.9	5	1.65	2.7225	4.785
19	7	3.9	5	1.65	2.7225	6.435
20	9	5.9	6	2.65	7.0225	15.635
	$\bar{Y} = 3.10$		$\bar{X} = 3.35$		$\sum_{i=1}^{N}(X_i - \bar{X})^2 = 30.55$	$\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = 43.30$

Bivariate Regression Model:

Dependent Variable: Subsequent Rearrests
Independent Variable: Prior Arrests
Regression Coefficient: b(Prior Arrests) = 43.30/30.55 = 1.4173

Working It Out

$$b_{X_2} = \left(\frac{r_{Y,PA} - (r_{Y,YP}r_{YP,PA})}{1 - r^2_{YP,PA}}\right)\left(\frac{s_Y}{s_{PA}}\right)$$

$$= \left(\frac{0.7616 - (0.7156)(0.6280)}{1 - (0.6280)^2}\right)\left(\frac{2.300}{1.2360}\right)$$

$$= \left(\frac{0.3122032}{0.605616}\right)(1.8608)$$

$$= 0.9593$$

The fact that the results are different when we examine the effects of years in prison and prior arrests in the multivariate regression model shows that the bivariate regression coefficients were indeed biased. In both cases, the estimate of the effect of *b* provided by the bivariate regression coefficient was much too high. These differences also reflect a difference in interpretation between the multivariate regression coefficient and the bivariate regression coefficient. In the bivariate case, the regression coefficient represents the estimated change in *Y* that is produced by a one-unit change in *X*. In the multivariate case, *b* represents the estimated change in *Y* associated with a one-unit change in *X* when *all other independent variables in the model are held constant.* Holding prior arrests constant leads to a reduction in the impact of years in prison. Holding years in prison constant leads to a reduction in the estimate of the effect of prior arrests. These differences may be seen as the bias introduced by misspecifying the regression model through the exclusion of prior arrests.

We can also identify this bias in terms of assumptions related to the error term in regression. It is assumed not only that the errors in the regression are stochastic, but also that there is no specific systematic relationship between the error term and the independent variables included in the regression. If there is such a relationship, the regression coefficient will be biased. While this may seem like a new concept, it is really a restatement of what you learned above.

Let's use our model predicting rearrest as a substantive example. We saw that if we estimated the regression coefficient for years in prison without taking into account prior arrests, the regression coefficient would be biased—in this case, overestimated. What happens in theory to the error term in this case? As we discussed earlier in the chapter, when we exclude an independent variable, the effect of that variable moves to

the error term. In our case, the population model including both independent variables may be stated as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where Y = subsequent rearrests
X_1 = years in prison
X_2 = prior arrests

When we take into account only one independent variable, the model includes only the term X_1:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

where Y = subsequent rearrests
X_1 = years in prison

In the latter model, number of prior arrests is included by implication in the error term. But what does this mean regarding the relationship in this model between the error term and years in prison? Our sample data suggest that the number of prior arrests is related to years in prison (as was shown in Table 16.3). By implication, since number of prior arrests is now found in the error term, the error term can now be assumed to be related to years in prison as well. Accordingly, if we leave prior arrests out of our equation, then we violate the assumption that there is no systematic relationship between the error term and the independent varibles in the equation.

By looking at bias in terms of the error term, we can also specify when excluding an independent variable will not lead to bias in our estimates of the regression coefficients of other variables. If the excluded variable is unrelated to other variables included in the regression, it will not cause bias in estimates of *b* for those specific variables. This is the case because when there is no systematic relationship between the excluded variable and the included variable of interest, its exclusion does not lead to a systematic relationship between the error term and the variable of interest.

For example, if years in prison and prior arrests were not systematically related (e.g., the correlation between the variables was 0), it would not matter whether we took into account prior arrests in estimating the regression coefficient for years in prison.[2] In this case, the exclusion of prior arrests would not lead to a systematic relationship between the error term

[2]It is important to note that bias can be caused by a nonlinear relationship between the excluded and the included variable. The assumption is that there is no systematic relationship of any form.

and years in prison, because there is no systematic relationship between prior arrests and years in prison. However, it is important to remember that the exclusion of prior arrests will still cause bias in our estimate of *Y*. In this situation, we continue to violate the assumption that the error term is stochastic. It now includes a systematic predictor of *Y*, prior arrests.

Comparing Regression Coefficients Within a Single Model: The Standardized Regression Coefficient

A multivariate regression model allows us to specify the impact of a specific independent variable while holding constant the impact of other independent variables. This is a very important advantage of multivariate regression analysis over bivariate regression analysis. However, when we include multiple variables in the same model, it is natural to want to compare the impact of the different variables examined. For example, in our case, does years in prison have a stronger effect on subsequent rearrests than number of prior arrests does? Or does number of prior arrests have a stronger effect than years in prison? The ordinary regression coefficient b does not allow us to answer this question, since it reports the effect of a variable in its original units of measurement. Accordingly, the regression coefficient for years in prison reports the predicted change in subsequent rearrests for each year change in years in prison. The regression coefficient for number of prior arrests reports the predicted change in subsequent rearrests for each change in number of prior arrests. Though the interpretation of the regression coefficients in these cases is straightforward, we cannot directly compare them.

Another statistic, called the **standardized regression coefficient** or **Beta,** allows us to make direct comparisons. Beta weights take the regression coefficients in an equation and standardize them according to the ratio of the standard deviation of the variable examined to the standard deviation of the dependent variable. Beta is expressed mathematically in Equation 16.3:

$$\text{Beta} = b\left(\frac{s_X}{s_Y}\right) \qquad \textbf{Equation 16.3}$$

The interpretation of the standardized coefficient is similar to that of b (the unstandardized coefficient), except that we change the units. We interpret Beta as the expected amount of change in the standard deviation of the dependent variable, given a one-unit change in the standard deviation of the independent variable.

For years in prison in our example, we take the regression coefficient of 0.9358 and multiply it by the ratio of the standard deviation of years in prison (0.9631) and subsequent rearrests (2.3000). The result is 0.3919,

which tells us that an increase of one standard deviation in years in prison is expected to result in an increase of 0.392 standard deviation in rearrests.

Working It Out

$$\text{Beta} = b\left(\frac{s_X}{s_Y}\right)$$

$$= 0.9358\left(\frac{0.9631}{2.3000}\right)$$

$$= 0.3919$$

For prior arrests, we begin with our regression coefficient of 0.9593. Again, we standardize our estimate by taking the ratio of the standard deviation of prior arrests (1.2360) and subsequent rearrests (2.3000). Our estimate of Beta here is 0.5155, which indicates that an increase of one standard deviation in prior arrests is expected to result in an increase of 0.516 standard deviation in rearrests.

Working It Out

$$\text{Beta} = b\left(\frac{s_X}{s_Y}\right)$$

$$= 0.9593\left(\frac{1.2360}{2.3000}\right)$$

$$= 0.5155$$

In our example, the Beta weight for prior arrests is larger than that for years in prison. According to this estimate, the number of prior arrests has a greater impact on subsequent rearrests than the number of years in prison does. The standardized regression coefficient thus provides us with an answer to our original question regarding which of the independent variables examined has the most influence on the dependent variable. As you can see, the standardized regression coefficient is a useful tool for comparing the effects of variables measured differently within a

single regression model. However, because standardized regression coefficients are based on the standard deviations of observed samples, they are generally considered inappropriate for making comparisons across samples.

Correctly Specifying the Regression Model

The previous section illustrated the importance of a correctly specified regression model. If a regression model is not correctly specified, then the predictions that are made and the coefficients that are estimated may provide misleading results. This raises important theoretical as well as practical questions for criminal justice research.

In criminal justice research, we can seldom say with assurance that the models we develop include all relevant predictors of the dependent variables examined. The problem is often that our theories are not powerful enough to clearly define the factors that influence criminal justice questions. Criminal justice is still a young science, and our theories for explaining crime and justice issues often are not well specified. This fact has important implications for the use of criminal justice research in developing public policy. When our predictions are weak, they do not form a solid basis on which to inform criminal justice policies.[3]

One implication of our failure to develop strongly predictive models in criminal justice is that our estimates of variable effects likely include some degree of bias. We have stressed in this chapter the importance of controlling for relevant predictors in regression modeling. The cost of leaving out important causes is not just weaker prediction but also estimates of variable effects that include potentially spurious components. This fact should make you cautious in reporting regression analyses and critical in evaluating the research of others. Just because regression coefficients are reported to the fifth decimal place on a computer printout does not mean that the estimates so obtained are solid ones.

The fact that regression models often include some degree of misspecification, however, should not lead you to conclude that the regression approach is not useful for criminal justice researchers. As in any

[3]Mark Moore of Harvard University has argued, for example, that legal and ethical dilemmas make it difficult to base criminal justice policies about crime control on models that still include a substantial degree of statistical error. See M. Moore, "Purblind Justice: Normative Issues in the Use of Prediction in the Criminal Justice System," in A. Blumstein, J. Cohen, A. Roth, and C. A. Visher (eds.), *Criminal Careers and "Career Criminals,"* Vol. 2 (Washington, DC: National Academy Press, 1986).

science, the task is to continue to build on the knowledge that is presently available. The researcher's task in developing regression models is to improve on models that were developed before. With each improvement, the results we gain provide a more solid basis for making decisions about criminal justice theory and policy. This, of course, makes the practical task of defining the correct model for the problem you are examining extremely important. How then should you begin?

Defining Relevant Independent Variables

Importantly, model specification does not begin with your data. Rather, it starts with theory and a visit to the library or other information systems. To build a regression model, you should first identify what is already known about the dependent variable you have chosen to study. If your interest, for example, is in the factors that influence involvement in criminality, you will need to carefully research what others have said and found regarding the causes of criminality. Your regression model should take into account the main theories and perspectives that have been raised by others.

If you do not take prior research and theory into account, then those reviewing your work will argue that your predictions and your estimates of variable effects are biased in one way or another. Just as the exclusion of prior record from our example led to a misleading estimate of its impact on length of imprisonment, so too the exclusion of relevant causal factors in other models may lead to bias. The only way to refute this potential criticism is to include such variables in your regression model.

Taking into account the theories and perspectives of others is the first step in building a correctly specified regression model. However, in most research we seek to add something new to existing knowledge. In regression modeling, this usually involves the addition of new variables. Sometimes, such new variables are drawn from an innovative change in theory. Other times, they involve improvements in measurement. Often, the finding that these new or transformed variables have an independent impact above and beyond those of variables traditionally examined by researchers leads to important advances in criminal justice theory and policy.

Taking into Account Ordinal- and Nominal-Scale Measures in a Multivariate Regression

Until now, we have assumed that ordinary least squares regression analysis requires an interval level of measurement, both for the dependent and for the independent variables. However, criminal justice researchers will sometimes use this approach with ordinal-level dependent variables when there are a number of categories and there is good reason to assume that the intervals for the categories are generally similar. In practice, you should be cautious in using OLS regression when your

dependent variable is not interval level. When the assumptions of OLS regression cannot be met in the case of ordinal dependent variables, you should use ordinal regression (see Chapter 19). As will be explained in Chapter 18, the use of OLS regression in the case of a binary dependent variable is inappropriate.

What about the inclusion of non–interval-level independent variables? Such variables often are important in explaining criminal justice outcomes. If we are required to include all relevant causes of *Y* in order to correctly specify our model, how can we exclude ordinal- and nominal-level measures? Fortunately, we do not have to. In multivariate regression, it is acceptable to include ordinal- and nominal-level independent variables as long as there is at least one interval-level independent variable also included in the analysis.

But even though you can include ordinal- and nominal-level variables, you need to take into account the specific interpretation used by regression analysis for interpreting the effects of one variable on another. Including an ordinal-level measure in a multivariate regression is relatively straightforward. This is done in Table 16.4, which presents a standard SPSS printout for a regression analysis. The data used are drawn from a national sample of police officers developed by the Police Foundation.[4] The dependent variable in this analysis is hours worked per week. There are two independent variables. One, years with the department, is measured at the interval level. The second, level of education, is on an ordinal scale with eight levels, ranging from some high school to doctoral

Table 16.4 SPSS Printout for Regression Analysis of the Police Officer Example

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	44.968	.749		60.031	.000
	YEARS WITH DEPARTMENT	−7.354E-02	.026	−.092	−2.816	.005
	LEVEL OF EDUCATION	.456	.173	.086	2.636	.009

a Dependent Variable: HOURS PER WEEK WORKED

[4]David Weisburd et al., *The Abuse of Authority: A National Study of Police Officers' Attitudes* (Washington, DC: The Police Foundation, 2001).

degree. We can see that both of these variables have a significant impact on hours worked per week—the observed significance levels ("Sig." in the SPSS table) are less than the conventionally applied significance level of 0.05 we would likely use in this case. This result, as in most statistical packages, is calculated for a two-tailed test of statistical significance (generally the default option). For years with the department, we can see that the impact is negative. When we control for the impact of level of education, each year with the department is associated with an average decrease of about 0.074 hours in number of hours worked each week.

But what is the meaning of the effect of level of education? Here, what we have is not an interval scale but a group of ordered categories. For the regression, this ordinal-level scale is treated simply as an interval-level scale. It is assumed that the categories must be roughly similar in value, or that each level increase in that scale is related in a linear manner to the dependent variable. Thus, our interpretation of this regression coefficient is that for every one-level increase in education level, there is, on average, a 0.456 increase in the number of hours worked (once we have taken into account years with the department). In this case, the standardized regression coefficient is very useful. It appears from the size of the coefficients that the overall effect of years with the department is much less than that of level of education. However, the standardized regression coefficients (represented by Beta) show that the difference between the two variables is not large.

This example illustrates how we can include an ordinal-level variable in a multivariate regression. The inclusion of an ordinal variable is straightforward, and its interpretation follows that of an interval-level independent variable. But when we include a nominal-level variable, we have to adjust our interpretation of the regression coefficient.

Table 16.5 SPSS Printout for Regression Analysis with an Interval-Level and Nominal-Level Variable

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	48.550	.977		49.672	.000
	YEARS WITH DEPARTMENT	−7.737E-02	.026	−.097	−2.943	.003
	RESPONDENT GENDER	−1.669	.803	−.068	−2.077	.038

a Dependent Variable: HOURS PER WEEK WORKED

Table 16.5 reports the results of a regression with a single interval-level variable and a binary nominal-level variable. We use the same dependent variable and data examined in our prior example. To make our example easier, we include only two measures in predicting number of hours worked. Again, we have an interval-level measure, years with the department. A binary independent variable, gender, is also included.

In regression analysis, a binary nominal-level independent variable is generally called a **dummy variable**. Our first problem is to give numbers to this dummy variable. Multivariate regression analysis does not recognize qualitative categories. By convention, we give one category a value of 0 and the other a value of 1. It is generally good practice to give the category with the largest number of cases a value of 0 because, as we will illustrate in a moment, that category becomes the reference category. Since this sample included many more men than women, we assigned men the value 0 and women the value 1.

Again, we can see that both variables have a statistically significant impact on hours worked per week. The observed significance level for years with the department is 0.003, and that for gender is 0.038. But how can we interpret the dummy variable regression coefficient of −1.669? One way to gain a better understanding of the interpretation of dummy variable regression coefficients is to see how they affect our regression equation. Let's begin by writing out the regression equation for our example:

$$Y = b_0 + b_1X_1 + b_2X_2$$

where Y = hours worked per week
X_1 = years with the department
X_2 = gender of officer

As a second step, let's insert the coefficients gained in our regression analysis:

$$Y = 48.550 + (-0.077)X_1 + (-1.669)X_2$$

or

$$Y = 48.550 - 0.077X_1 - 1.669X_2$$

What happens if we try to write out the regression equations for men and women separately? For men, the regression equation is

$$Y = 48.550 - 0.077X_1 - 1.669(0)$$

or

$$Y = 48.550 - 0.077X_1$$

Because men are coded as 0, the second term of the equation falls out. But what about for women? The second term in the equation is a constant because all of the women have a value of 1. If we write it out, we have the following result:

$$Y = 48.550 - 0.077X_1 - 1.669(1) \text{ or}$$

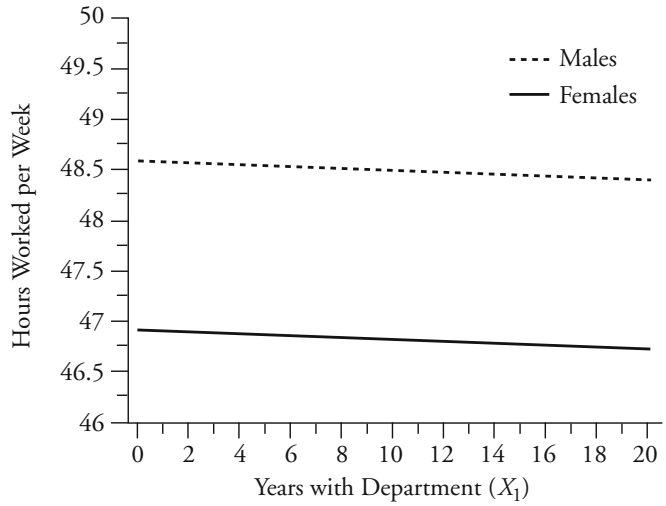
$$Y = 48.550 - 0.077X_1 - 1.669$$

We can simplify this formula even more, because the two constants at the beginning and the end of the equation can be added together:

$$Y = 46.881 - 0.077X_1$$

What then is the difference between the regression equations for men and women? In both cases, the slope of the regression line is given by the term $-0.077X_1$. The difference between the two equations lies in the *Y*-intercept, as illustrated in Figure 16.1. As you can see, men and women have parallel regression lines. However, the women's line intersects the *Y*-axis about 1.7 hours lower than the men's line. This provides us with the interpretation of our coefficient. Women police officers, on

Figure 16.1 *Regression Lines for Men and Women Police Officers*



average, work about 1.669 hours a week less than men police officers, taking into account years with the department.

This example also suggests why it is generally recommended that you place the category with the largest number of cases as the 0 category of a binary dummy variable. The category men, in this case, is the reference category, meaning that the coefficient for gender gives us the estimate of the female category in reference to the male category. We want our reference category to be as stable as possible, and a large number of cases makes this category more stable.

But how can we assess the impact of a nominal variable that has multiple categories? In fact, multiple-category nominal variables create a good deal more complexity for the researcher than do ordinal or binary nominal variables. In this case, you must create a separate variable for each category in your analysis. For example, the Police Foundation study divided the United States into four regions: North Central, Northeast, South, and West. In practice, you would need to create a separate variable for each of these regions. In other words, you would define a variable North Central, which you would code 1 for all those officers in the North Central region and 0 for all other officers. You would repeat this process for each of the other regional categories.

As with the binary independent variable, you must choose one of the categories to be a reference category. In this case, however, the reference category is excluded from the regression. Again, it is generally recommended that you choose as the reference category the category with the largest number of cases.[5] In our example, the largest number of officers is drawn from the South. Suppose that we include only one interval-level variable in our equation: years with the department. Table 16.6 presents the results from an analysis in which years with the department and region are used to predict number of hours worked.

In this example, we included in the regression a single interval-level variable and three region measures: North Central, Northeast, and West. While South is not included as a variable, it is in fact the reference

[5]There may be times when you want to choose a category that does not include the largest number of cases as the reference. For example, if you wanted to compare a series of treatments to a no-treatment, or control, condition, it would make sense to have the control condition as the excluded category, even if it did not include the largest N. However, if the excluded category has a small number of cases, it may lead to instability in the regression estimates.

Table 16.6 SPSS Printout for Regression Analysis with Multiple-Category Nominal Variable

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	47.654	.470		101.435	.000
	YEARS WITH DEPARTMENT	−6.138E-02	.026	−.077	−2.352	.019
	NORTH CENTRAL	−2.335	.610	−.141	−3.825	.000
	NORTHEAST	−1.758	.573	−.114	−3.067	.002
	WEST	−.846	.616	−.050	−1.372	.170

a Dependent Variable: HOURS PER WEEK WORKED

category. If we again write out our regression equation, we can see why this is the case:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

where Y = hours worked per week
X_1 = years with department
X_2 = North Central
X_3 = Northeast
X_4 = West

Using the results for this model presented in Table 16.6, we can write the results in equation form as follows:

$$Y = 47.654 - 0.061X_1 - 2.335X_2 - 1.758X_3 - 0.846X_4$$

In this case, we can also write out a separate regression equation for each of the four regions. Since the South is our reference category, those from the South are coded 0 on the three included variables. Thus, our equation is simply the Y-intercept and the variable years with the department. For officers from the North Central region, the equation includes the Y-intercept, b_1X_1, and b_2X_2. The other parameters are set to 0, since those in the North Central region have 0 values on X_3 and X_4. Similarly, for both the Northeast and the West, only

one of the three dummy variables is included. For each equation, we can once again add the constant for the dummy variable to the *Y*-intercept:

Officers from the South:

$$Y = 47.654 - 0.061X_1$$

Officers from the North Central:

$$Y = 47.654 - 0.061X_1 - 2.335X_2 \quad \text{or} \quad Y = 45.319 - 0.061X_1$$

Officers from the Northeast:

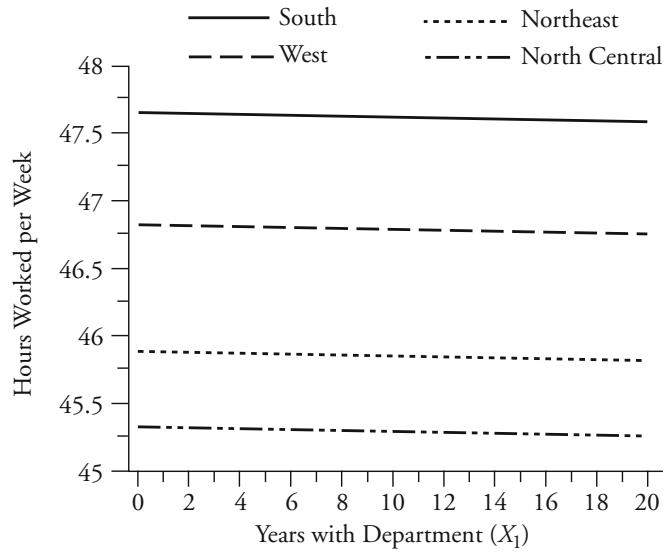
$$Y = 47.654 - 0.061X_1 - 1.758X_3 \quad \text{or} \quad Y = 45.896 - 0.061X_1$$

Officers from the West:

$$Y = 47.654 - 0.061X_1 - 0.846X_4 \quad \text{or} \quad Y = 46.808 - 0.061X_1$$

Once again, we can gain a better conceptual understanding of our results if we plot them, as in Figure 16.2. In this case, each of the included categories is found to have a *Y*-intercept lower than that of the excluded

Figure 16.2 *Plot of Hours Worked and Years with Department, by Region*



category, the South. This means that, on average, officers work fewer hours in all of the other regions. The least number of hours worked per week is found in the North Central region. Here, officers work, on average, about 2.335 hours less than they do in the South, once we have taken into account years with the department. In the Northeast, officers work about 1.758 hours less and in the West about 0.846 hour less a week.

Are these differences statistically significant? It is important to note that the significance statistic reported for each coefficient tells us only whether the category is significantly different from the reference category. This is one reason it is so important to be clear about the definition of the reference category. In our example, the North Central and Northeast regions are significantly different from the South, using a 5% significance threshold and a two-tailed significance test (the default option in SPSS). The West, however, is not significantly different from the South, using this threshold.

If you wanted to determine whether region overall as a variable had a statistically significant impact on hours worked per week, you would have to run an additional significance test based on the *F*-test for the regression model, introduced in Chapter 15. The *F*-test for multiple-category dummy variables in regression compares the R^2 statistic gained with the dummy variables included in the regression to the R^2 statistic gained without those variables. In practice, you must run two separate regressions, although most computer programs now provide this statistic directly. First, you calculate the regression without the new dummy variable categories (referred to as the reduced model) and identify its R^2. In our case, the regression without the dummy variables produces an R^2 of only 0.008. You then compute the regression with the dummy variables, as we did earlier (referred to as the full model). In this case, R^2 is 0.023. The *F*-test formula is presented in Equation 16.4.

$$F = \frac{(R^2_{fm} - R^2_{rm})/(k_{fm} - k_{rm})}{(1 - R^2_{fm})/(N - k_{fm} - 1)} \qquad \textbf{Equation 16.4}$$

To apply Equation 16.4 to our example, we first subtract the R^2 of the reduced model (R^2_{rm}) from the R^2 of the full model (R^2_{fm}) and then divide this quantity by the number of variables in the full model (k_{fm}) minus the number of variables in the reduced model (k_{rm}), which is 3. The denominator is found by subtracting the R^2 of the full model from 1, and then dividing this quantity by $N - k_{\text{fm}} - 1$. For this sample, N is 923 and k_{fm} is 4. Our final result is $F = 4.55$. Looking at the *F*-distribution (see

Appendix 5) with 3 and 918 degrees of freedom, we can see that our result is statistically significant at the 0.05 level.

Working It Out

$$F = \frac{(R^2_{fm} - R^2_{rm})/(k_{fm} - k_{rm})}{(1 - R^2_{fm})/(N - k_{fm} - 1)}$$

$$= \frac{(0.023 - 0.008)/(4 - 1)}{(1 - 0.023)/(923 - 4 - 1)}$$

$$= \frac{(0.015/3)}{(0.977/918)} = \frac{0.005}{0.0011}$$

$$= 4.5455$$

One final question we might ask is whether we can use the standardized regression coefficient to compare dummy variables to ordinal- and interval-level measures. In general, statisticians discourage such use of standardized regression coefficients, since they are based on standard deviations and the standard deviation is not an appropriate statistic for a nominal-level variable. Additionally, for a nominal-level variable, the standardized regression coefficient refers only to the difference between the reference category and the dummy variable category examined. This may sometimes make sense in the case of a binary dummy variable, since we can say that one category is Beta standard deviations higher or lower on the dependent variable. But it can be extremely misleading in the case of multi-category nominal-level variables, such as region in our example. The size of the standardized regression coefficient, like the size of the coefficient itself, will depend on which category is excluded. In general, you should exercise caution when interpreting standardized regression coefficients for dummy variables in a multivariate regression analysis.

Chapter Summary

In a **bivariate regression model,** there is only one independent variable, and it must be an interval-level measure. Importantly, the researcher can rarely be sure that the change observed in the dependent variable is due to one independent variable alone. And if variables that have an impact on the dependent measure are excluded, the predictions

of *Y* gained in a regression will be biased. If the excluded variables are related to the included factor, then the estimate of *b* for the included factor will also be biased. **Randomized experiments,** which scatter different traits at random, offer a solution to the latter problem, but they are often impractical in criminal justice research. A statistical solution that enables us to correct for both types of bias is to create a **multivariate regression model**.

In a multivariate regression model, there may be several independent variables, only one of which needs to be interval level. Such a model considers the effect of each independent variable, while holding all the other variables constant. A binary nominal-level variable included in a regression model is called a **dummy variable**. Regression coefficients measured using different scales may be compared with a **standardized regression coefficient (Beta)**. A regression model is **correctly specified** if the researcher has taken into account and correctly measured all of the relevant predictors of the dependent variable. Existing literature and prior research are suitable places to start.

Key Terms

biased Describing a statistic when its estimate of a population parameter does not center on the true value. In regression analysis, the omission of relevant independent variables will lead to bias in the estimate of *Y*. When relevant independent variables are omitted and those measures are related to an independent variable included in regression analysis, then the estimate of the effect of that variable will also be biased.

correctly specified regression model A regression model in which the researcher has taken into account all of the relevant predictors of the dependent variable and has measured them correctly.

dummy variable A binary nominal-level variable that is included in a multivariate regression model.

multivariate regression A technique for predicting change in a dependent variable, using more than one independent variable.

randomized experiment A type of study in which the effect of one variable can be examined in isolation through random allocation of subjects to treatment and control, or comparison, groups.

standardized regression coefficient (Beta) Weighted or standardized estimate of *b* that takes into account the standard deviation of the independent and the dependent variables. The standardized regression coefficient is used to compare the effects of independent variables measured on different scales in a multivariate regression analysis.

Symbols and Formulas

k	Number of independent variables in the overall regression model
r_{Y,X_1}	Correlation coefficient for Y and X_1
r_{Y,X_2}	Correlation coefficient for Y and X_2
r_{X_1,X_2}	Correlation coefficient for X_1 and X_2
s_Y	Standard deviation for Y
s_{X_1}	Standard deviation for X_1
R^2_{fm}	R^2 obtained for the full regression model
R^2_{rm}	R^2 obtained for the reduced regression model
k_{fm}	Number of independent variables in the full regression model
k_{rm}	Number of independent variables in the reduced regression model

To calculate a multivariate regression coefficient for two independent variables:

$$b_{x_1} = \left(\frac{r_{Y,X_1} - (r_{Y,X_2} r_{X_1,X_2})}{1 - r^2_{X_1,X_2}}\right)\left(\frac{s_Y}{s_{X_1}}\right)$$

and

$$b_{x_2} = \left(\frac{r_{Y,} - (r_{Y,X_1,} r_{X_1,X_2})}{1 - r^2_{X_1,X_2}}\right)\left(\frac{s_Y}{s_{X_2}}\right)$$

A sample multivariate regression model with three independent variables:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$$

A population multivariate regression model with three independent variables:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \epsilon$$

To calculate the standardized coefficient (Beta):

$$Beta = b\left(\frac{s_X}{s_Y}\right)$$

To calculate an *F*-test on a subset of variables in a regression model:

$$F = \frac{(R^2_{fm} - R^2_{rm})/(k_{fm} - k_{rm})}{(1 - R^2_{fm})/(N - k_{fm} - 1)}$$

Exercises

16.1 Consider the following regression model, which purports to predict the length of sentence given to convicted thieves:

$$Y = b_0 + bX + e$$

where Y = length of sentence

X = number of prior sentences

a. List the variables you might wish to include in a more comprehensive model. Include a brief statement about why each additional variable should be included.

b. Present your model in equation form.

16.2 In an article in the newspaper, a researcher claims that low self-esteem is the cause of crime. Upon closer inspection of the results in the paper, you learn that the researcher has computed a bivariate model using self-reported theft as the dependent variable (Y) and self-esteem as the one independent variable (X).

a. List the variables you might wish to include in a more comprehensive model. Include a brief statement about why each additional variable should be included.

b. Present your model in equation form.

16.3 A researcher has built a multivariate regression model to predict the effect of prior offenses and years of education on the length of sentence received by 100 convicted burglars. He feeds the data into a computer package and obtains the following printout:

Dependent Variable (Y): Length of Sentence (months)

Independent Variable (X_1): Number of Prior Offenses

Independent Variable (X_2): Years of Education

F sig = 0.018

R Square = 0.16

X_1: b = +0.4 Sig t = 0.023

X_2: b = −0.3 Sig t = 0.310

Evaluate the results, taking care to explain the meaning of each of the statistics produced by the computer.

16.4 An analysis of the predictors of physical violence at school produced the following results:

Independent Variable	*b*	Beta
Age (Years)	0.21	0.05
Sex (Female = 1, Male = 0)	−3.78	0.07
Race (White = 1, Non-white = 0)	−1.34	0.06
Number of Friends Arrested	1.96	0.33
Number of Times Attacked by Others	3.19	0.24
Number of Times Hit by Parents	2.05	0.27

Explain what each regression coefficient (*b*) and standardized regression coefficient (Beta) means in plain English.

16.5 Danny has obtained figures on the amount of drugs seized per month at a seaport over the course of two years. He wishes to explain variations in the amount of drugs seized per month and runs a regression analysis to check the effect of his independent variable—the total number of customs officers on duty for each month—on the quantity of drugs seized. The resulting regression coefficient is +4.02. Danny is worried, however, that his bivariate model might not be correctly specified, and he decides to add another variable—the number of ships that arrive at the port each month. He calculates the correlations between the three pairs of variables, and the results are as follows:

Y (drugs seized), X_1 (customs officers): +0.55

Y (drugs seized), X_2 (ships arriving): +0.60

X_1 (customs officers), X_2 (ships arriving): +0.80

The standard deviations for the three variables are 20 kg (quantity of drugs seized per month), 1.6 (number of customs officers on duty), and 22.5 (number of ships arriving).

a. Calculate the regression coefficient for customs officers.

b. Calculate the regression coefficient for ships arriving.

c. How do you account for the difference between your answer to part a and the regression coefficient of +4.02 that Danny obtained earlier?

16.6 A study of prison violence examined the effects of two independent variables—percent of inmates sentenced for a violent crime (X_1) and average amount of space per inmate (X_2)—on the average number of violent acts per day (Y). All variables were measured for a random

selection of cell blocks in three prisons. The researcher reported the following results:

$r_{Y,X_1} = 0.20$

$r_{Y,X_2} = 0.20$

$r_{X_1,X_2} = 0.20$

$s_Y = 0.35$

$s_{X_1} = 10.52$

$s_{X_2} = 2.64$

a. Calculate the regression coefficients for the effects of X_1 and X_2 on Y. Explain what these coefficients mean in plain English.

b. Calculate the standardized regression coefficients for the effects of X_1 and X_2 on Y. Explain what these coefficients mean in plain English.

c. Which one of the variables has the largest effect on prison violence? Explain why.

16.7 A study of recidivism classified offenders by type of punishment received: prison, jail, probation, fine, or community service. A researcher interested in the effects of these different punishments analyzes data on a sample of 967 offenders. She computes two regression models. In the first, she includes variables for age, sex, race, number of prior arrests, severity of the last conviction offense, and length of punishment. The R^2 for this model is 0.27. In the second model, she adds four dummy variables for jail, probation, fine, and community service, using prison as the reference category. The R^2 for this model is 0.35. Explain whether the type of punishment had an effect on recidivism (assume a 5% significance level).

16.8 A public opinion poll of 471 randomly selected adult respondents asked about their views on the treatment of offenders by the courts. Expecting race/ethnicity to be related to views about the courts, a researcher classifies respondents as African American, Hispanic, and white. To test for the effect of race/ethnicity, he computes one regression using information about the age, sex, income, and education of the respondents and finds the R^2 for this model to be 0.11. In a second regression, he adds two dummy variables for African American and Hispanic, using white as the reference category. The R^2 for this second model is 0.16. Explain whether the race/ethnicity of the respondent had a statistically significant effect on views about the courts (assume a 5% significance level).

Computer Exercises

In Chapter 15, we explored the basic features of the regression commands in SPSS and Stata in the computation of a bivariate regression model. To compute a multivariate regression model, we simply add additional independent variable names to the list of independent variables on the command line. The following exercises illustrate some of the additional features of the regression command. Please see the appropriate SPSS (Chapter_16.sps) or Stata (Chapter_16.do) syntax file for specific examples.

SPSS

Standardized Regression Coefficients (Betas)

The standardized regression coefficients (Betas) are part of the standard output for SPSS's linear regression command. In the table of results presenting the coefficients, the standardized coefficients are located in the column following those presenting the values for the regression coefficients (*b*) and the standard errors of *b*. Nothing else is required to obtain the standardized coefficients.

F-Test for a Subset of Variables

The computation of an *F*-test for a subset of variables requires a little planning in setting up a multivariate linear regression model. When thinking about your regression model and a test of one or more subsets of variables, you will need to enter these independent variables on separate /METHOD = ENTER lines. In general, if we have one subset that we are interested in, we would use the following syntax:

```
REGRESSION
  /STATISTICS COEFF R ANOVA CHANGE
  /DEPENDENT dep_var_name
  /METHOD = ENTER list_of_variables_NOT_in_subset
  /METHOD = ENTER list_of_variables_in_subset.
```

The trick here is to keep track of all the independent variables in your regression model and determine whether they belong to the first or the second group—the second group would be the subset of interest. For example, suppose we had an interest in looking at whether demographic characteristics of offenders affected punishment severity. In this case, we would then list the demographic characteristics (however measured) in the second block (i.e., /METHOD = ENTER line).

We have also added the /STATISTICS option line to the REGRESSION command. The reason for this is to force SPSS to compute the *F*-test on the subset of variables and to simultaneously report all of the other results in a linear regression analysis that it usually reports. Specifically, the items on the /STATISTICS line request the coefficient table (COEFF), model summary (R),

ANOVA table (ANOVA), and change in R^2 when the second block of variables is added to the regression model (CHANGE). The *F*-test on the subset of variables is produced with the CHANGE option.

The output from running this command is nearly identical to what you have viewed previously. The major difference is that there will be two major rows of results for all of the tables viewed in the output before—one row will be labeled Model 1 and the other row Model 2. In other words, there will be a row for the "reduced" model (Model 1 in SPSS) that contains only those variables included in the first block of variables and a second row for the "full" model that includes all variables (Model 2 in SPSS).

The *F*-test for the subset of variables can be found in the "Model Summary" table of results under the columns labeled "Change Statistics." For Model 2, the *F*-statistic for the subset of variables appears in the column labeled "F Change." The value for the numerator degrees of freedom (df_1) will appear in the next column to the right and will equal the number of independent variables included in the subset. The value for the denominator degrees of freedom (df_2) appears in the next column, providing you with all the information you need to test whether the subset of variables makes a statistically significant contribution to the overall regression model.

Since the description of the various pieces may be confusing, we encourage you to open and run the accompanying SPSS syntax file for this chapter (Chapter_16.sps).

Residual Plot

It is also possible with the regression command to analyze residuals in ways ranging from simple to complex. Perhaps the most straightforward way of analyzing residuals is graphically, through the use of a residual plot that SPSS can produce. There are many different kinds of residual plots that SPSS could create—we highlight only one simple example here. A histogram of the residuals from a regression analysis with a normal curve overlaid on the histogram is obtained as follows:

```
REGRESSION
/DEPENDENT dep_var_name
/METHOD=ENTER list_of_indep_vars
/RESIDUALS HISTOGRAM(ZRESID).
```

where the /RESIDUALS line will request a plot of residuals—the HISTOGRAM(ZRESID) option specifies a histogram of what are known as "standardized residuals." A histogram of the residuals with the overlaid normal curve will give you some idea of how closely the residuals approximate a normal distribution (which is what is to be expected). If the residuals do not resemble a normal distribution, this is often an indication of a problem with the regression model, such as one or more relevant independent variables having been omitted from the analysis.

Stata

Standardized Regression Coefficients (Betas)

To request the standardized regression coefficients (Betas) in a multivariate linear regression model in Stata, you will need to add the option **b** to a **regress** command:

regress dep_var_name indep_var_names, **b**

The last column of output in the coefficient table will then report the standardized coefficients.

F-Test for a Subset of Variables

In contrast to the cumbersome syntax in SPSS for testing a subset of variables, the syntax required in Stata involves two steps: (1) estimate the full regression model with the regress command and (2) test the subset of variables using the **testparm** command. The form of the **testparm** command is simply

testparm list_of_subset_variables

The output from running this command is an *F*-test on the set of variables listed on the **testparm** command line.

Residual Plot

We illustrated at the end of Chapter 15 the process for computing residuals from a linear regression analysis. To create a histogram of the residuals, we would use the **histogram** command (discussed at the end of Chapter 3), and if we wanted to overlay a normal curve, we use the **normal** option:

regress dep_var_name indep_var_names

predict RES_1, **r**

histogram RES_1, **normal**

Problems

1. Enter the data from Table 16.2. Run the regression command to reproduce the unstandardized and standardized regression coefficients presented in this chapter.

   a. Compute two bivariate regression models, using years in prison as the independent variable in one regression and prior arrests as the independent variable in the second regression. Generate a histogram of the residuals for each regression model. What does the pattern of results in this plot suggest to you about the distribution of error terms?

   b. Compute the multivariate model, and generate a histogram of the residuals for this regression model. How has the pattern of error terms changed relative to the two histograms produced in part a?

   Open the NYS data file (nys_1.sav, nys_1_student.sav, or nys_1.dta) to do Exercises 2 through 5.

2. Compute a multivariate regression model, using number of times the student hit other students as the dependent variable. From the variables included in the data file, select at least five independent variables that you think have some relationship to hitting other students.

   a. Explain what each regression coefficient (b) and standardized regression coefficient (Beta) means in plain English.

   b. Generate a histogram of the residuals for this regression model. What does the pattern of results in this plot suggest to you about the distribution of error terms?

3. Compute a multivariate regression model, using number of times something worth $5 or less has been stolen as the dependent variable. From the variables included in the data file, select at least five independent variables that you think have some relationship to stealing something worth $5 or less.

   a. Explain what each regression coefficient (b) and standardized regression coefficient (Beta) means in plain English.

   b. Generate a histogram of the residuals for this regression model. What does the pattern of results in this plot suggest to you about the distribution of error terms?

4. Compute a multivariate regression model, using number of times the student cheated on exams as the dependent variable. From the variables included in the data file, select at least five independent variables that you think have some relationship to cheating on exams.

   a. Explain what each regression coefficient (b) and standardized regression coefficient (Beta) means in plain English.

   b. Generate a histogram of the residuals for this regression model. What does the pattern of results in this plot suggest to you about the distribution of error terms?

5. Compute a multivariate regression model, using number of times drunk as the dependent variable. Use age, sex, race, employment status, hours spent studying per week, grade point average, and number of friends who use alcohol as the independent variables.

   a. Use an *F*-test to test whether demographic characteristics—age, sex, and race—affect drinking behavior.

   b. Use an *F*-test to test whether academic characteristics—hours spent studying per week and grade point average—affect drinking behavior.