9 ANALYZING DATA

Multiple Variables

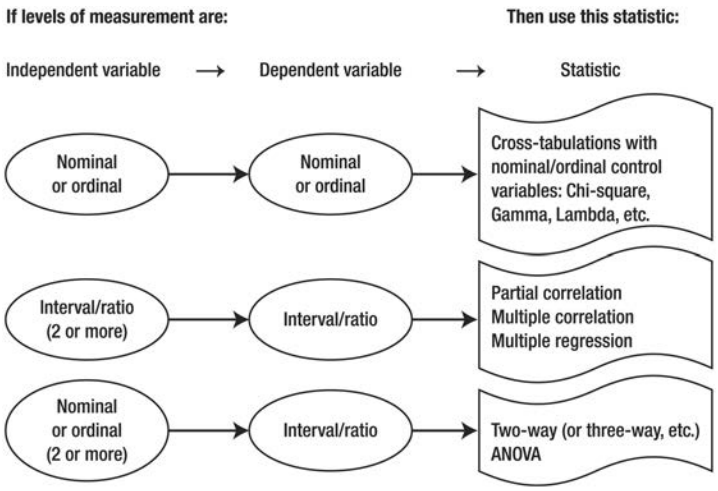
Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

—Marie Curie, Nobel laureate scientist

LEARNING GOALS

This chapter focuses on the analysis of three or more variables to answer more complex research questions. It discusses when to use various kinds of multivariate analyses and how to elaborate your findings with additional variables (see Figure 9.1). By the end of the chapter, you should be able to interpret linear multiple regression analyses and perform elaboration techniques with control variables.

Figure 9.1 Statistical Decision Steps (also see Statistical Analysis Decision Tree in Appendix)



Although it may be tempting to try to understand people's behavior and opinions with just one simple explanation or cause, in reality there are multiple reasons for those beliefs and actions. Understanding more about how your variables work together in explaining behavior and attitudes is essential for doing quality research. If, for example, you are trying to understand why students' grades vary, the number of hours studying is probably just one reasonable answer. What else might explain differences in grades? Quality of the instructor, test-taking abilities, physical and mental health at the time of doing the required work or test, difficulty of an assignment, basic intelligence levels, and the party the night before are all plausible alternative explanations.

One of the goals of research is to evaluate multiple explanations or predictors of some dependent variable or outcome. Sometimes it is also necessary to verify that the relationships we have uncovered withstand alternative explanations or are consistent for a variety of subgroups in the sample. Unless the elimination of other variables that could possibly explain the main relationship occurs, we have not established cause and effect. Recall the three elements discussed in Chapter 1 that are essential to declaring causality. Most advanced methods books and online websites describe multivariate statistics in greater depth, but for our introductory purposes, this chapter focuses on some basic techniques and concepts that are helpful for analyzing data for the first time.

ELABORATING RELATIONSHIPS: CONTROL VARIABLES

A typical question is whether the relationships you have discovered are strong enough to withstand other plausible explanations. What we need to do is something called *elaboration*, which extends our knowledge about the association to see if it continues or changes under different situations. Researchers who use experimental designs use a control group as a means of verifying the results that occurred in the experimental group. Similarly, those of us doing survey research can control for other variables during the data analysis phase.

Consider an association that is statistically significant between respondents' educational level and yearly family income (Table 9.1). Let's first take the case where these variables are ordinal measurements—education: less than high school, high school graduate, and so on; income: under $25,000; $25,000 to $39,999; and so forth. Minimally, a chi-square can be calculated for this cross-tabulation of data, along with some other statistics appropriate for ordinal measures, such as Gamma or Kendall's Tau-c. Perhaps you know, though, that women often earn less than men, and you wonder if educational level remains as good a predictor of total family income for women. The question is whether the relationship is still statistically significant when *controlling* for sex of the respondent; that is, does it still apply equally to men and women? Sex becomes the test or control variable used to elaborate the original relationship between education and income.

Table 9.1 Cross-Tabulation of Income and Education, SPSS

TOTAL FAMILY INCOME BY HIGHEST DEGREE CROSS-TABULATION

			HIGHEST DEGREE					
			Less Than High School	High School	Junior College	Bachelor	Graduate	Total
Total family income ($)	24,999 or less	Count	196	315	25	39	9	584
		% of highest degree	70.3%	40.4%	27.8%	16.7%	8.0%	39.0%
	25,000 to 39,999	Count	28	175	21	58	18	300
		% of highest degree	10.0%	22.4%	23.3%	24.8%	15.9%	20.1%
	40,000 to 59,999	Count	16	121	23	52	18	230
		% of highest degree	5.7%	15.5%	25.6%	22.2%	15.9%	15.4%
	60,000 or more	Count	39	169	21	85	68	382
		% of highest degree	14.0%	21.7%	23.3%	36.3%	60.2%	25.5%
Total		Count	279	780	90	234	113	1,496
		% of highest degree	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

CHI-SQUARE TESTS

	Value	*df*	Asymptotic Significance (two-tailed)
Pearson chi-square	264.299[a]	12	.000
N of valid cases	1,496		

[a] Zero cells (0.0%) have expected count less than 5. The minimum expected count is 13.84.

SYMMETRIC MEASURES

		Value	Approximate Significance
Ordinal measures	Kendall's Tau-c	0.295	.000
	Gamma	0.457	.000
N of valid cases		1,496	

As the chi-square and Gamma indicate, there is a significant relationship, and those with less education tend to be in households with less family income: 70.3 percent of those with less than a high school degree earn under $25,000 yearly, compared with only 8 percent of those with graduate degrees. As it turns out in this sample, 60.2 percent of people with graduate degrees earn over $60,000 a year. Now let's see how well this relationship holds up when introducing a third variable: respondents' sex, which has two categories or values, male and female (see Table 9.2).

Partialling

Two tables are developed, one showing the relationship between education and family income for males and one for females, along with two sets of chi-squares and Gammas. There are now three variables involved in the multivariate analysis. When you introduce a control variable, it is referred to as first-order *partialling*. You can continue to add multiple variables, called second-order, third-order, and so on, for more elaborate models, but interpretation can get complex at that point, especially if each of those variables has numerous values or categories. The original bivariate association is the *zero-order* relationship.

Notice in the new partial tables that all the statistics remain significant, and the ordinal measures of strength (Gamma and Kendall's Tau-c) are about the same, although slightly stronger for men. You have replicated your original finding and can now conclude that there is a strong relationship between education and income, because it holds up even when controlling for sex. It is unlikely that being male or female solely affects total family income level; education appears equally or more important. For example, 65.6 percent of men with less than a high school education tend to have under $25,000 a year in total family income; 74 percent of women with less than a high school education earn under $25,000 a year. This pattern is replicated across the other education and income categories. You now might want to test the original zero-order relationship again with a new control variable to see if it is sustained for other conditions.

Spurious and Antecedent Relationships

What would it mean when you introduced a control variable and the relationship did not hold up? If the original relationship disappears or becomes less strong—it is statistically significant for neither the men nor the women, or the correlation coefficients decline substantially—you would claim the original bivariate relationship was *spurious*. You would verify this by creating a cross-tabulation for sex and education and another one for sex and income and see if indeed sex was *antecedent* to or came

Table 9.2 Cross-Tabulation of Income and Education, Controlling for Sex, SPSS

TOTAL FAMILY INCOME BY HIGHEST DEGREE
BY RESPONDENT'S SEX CROSS-TABULATION

Respondent's Sex				HIGHEST DEGREE					
				Less Than HS	High School	Junior College	Bachelor	Graduate	Total
Male	Total family income	24,999 or less	Count	82	109	11	17	5	224
			% of highest degree	65.6%	35.9%	29.7%	15.7%	7.5%	34.9%
		25,000 to 39,999	Count	22	78	6	26	11	143
			% of highest degree	17.6%	25.7%	16.2%	24.1%	16.4%	22.3%
		40,000 to 59,999	Count	9	62	10	27	10	118
			% of highest degree	7.2%	20.4%	27.0%	25.0%	14.9%	18.4%
		60,000 or more	Count	12	55	10	38	41	156
			% of highest degree	9.6%	18.1%	27.0%	35.2%	61.2%	24.3%
	Total		Count	125	304	37	108	67	641
			% of highest degree	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Female	Total family income	24,999 or less	Count	114	206	14	22	4	360
			% of highest degree	74.0%	43.3%	26.4%	17.5%	8.7%	42.1%
		25,000 to 39,999	Count	6	97	15	32	7	157
			% of highest degree	3.9%	20.4%	28.3%	25.4%	15.2%	18.4%
		40,000 to 59,999	Count	7	59	13	25	8	112
			% of highest degree	4.5%	12.4%	24.5%	19.8%	17.4%	13.1%
		60,000 or more	Count	27	114	11	47	27	226
			% of highest degree	17.5%	23.9%	20.8%	37.3%	58.7%	26.4%
	Total		Count	154	476	53	126	46	855
			% of highest degree	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

CHI-SQUARE TESTS

Respondent's Sex		Value	*df*	Asymptotic Significance (two-tailed)
Male	Pearson chi-square	136.855	12	.000
	N of valid cases	641		
Female	Pearson chi-square	145.194	12	.000
	N of valid cases	855		

SYMMETRIC MEASURES

Respondent's Sex			Value	Approximate Significance
Male	Ordinal measures	Kendall's Tau-c	0.341	.000
		Gamma	0.491	.000
	N of valid cases		641	
Female	Ordinal measures	Kendall's Tau-c	0.261	.000
		Gamma	0.433	.000
	N of valid cases		855	

before the dependent variable of income, and if sex explained the dependent variable of education. If so, then the original relationship had the illusion of an association only because the independent (X) and dependent (Y) variables were both related to a third antecedent variable (Z) for which you controlled.

Here is a classic example of a spurious relationship: The number of fire engines that show up at a fire (independent variable) is significantly related to the amount of the fire damage (dependent variable). In other words, you might conclude that if only one or two fire trucks appeared, the damage would have been less. Here you can see the problem of making a cause-and-effect conclusion on the basis of just a significant correlation. Clearly, there is need for an alternative explanation, and perhaps size of the fire is one such control or test variable. Sure enough, the original relationship disappears when you demonstrate that fire size (independent variable) is highly correlated to the number of fire trucks that are called (dependent) and that fire size (independent) is highly related to the amount of damage (dependent). Fire size is the more direct reason for the amount of damage, not the number of fire engines!

Specification

Sometimes when you control for a third variable, you notice that the original relationship disappears only for some of the values of the test variable. Let's say that an

elaboration of the education and income model showed that the original relationship applied only to the men in the sample and not to the women. You would notice this by seeing that the statistical tests remained significant for the men but were no longer statistically significant for the women or they were weakened. In such cases, you have *specification* because you determined very specific situations in which the relationship holds and those in which it does not.

Suppressing Relationships

Elaboration is also useful when you find no relationship or a very weak one between your independent and dependent variables. Don't give up and think that you found little. Elaborate on this nonfinding and introduce a control variable to see if for some reason this is *suppressing* or holding back the original zero-order relationship. Imagine that we found a weak relationship between parents' educational level and their adult children's (the respondents') income, despite reports about greater income for those growing up in higher educated families. You introduce respondents' education (Z) as a test control variable and see whether the weak relationship between parents' education (X) and income (Y) continues for each level of respondents' education (Z_1, Z_2, Z_3, etc.). If a relationship now appears in one or more of the respondents' education categories, we can conclude that education was suppressing the original relationship (see Box 9.1).

BOX 9.1
THE PROCESS OF ELABORATION

Zero-order relationship: X (independent variable) → *Y* (dependent variable)
First-order partialling: Z (control/test variable with three values or categories, Z_1, Z_2, and Z_3)

1. Antecedent $Z \rightarrow X \rightarrow Y$ original relationship elaborated

2. Intervening $X \rightarrow Z \rightarrow Y$ relationship works through the test variable

3. Spurious $X \rightarrow Y$ relationship is significant
   for Z_1, $X \rightarrow Y$ relationship disappears/weakens
   for Z_2, $X \rightarrow Y$ relationship disappears/weakens
   for Z_3, $X \rightarrow Y$ relationship disappears/weakens
   because of antecedent $Z \rightarrow X$
   $Z \rightarrow Y$

BOX 9.1 CONTINUED

4. Specification	for Z_1, X → *Y* relationship disappears/weakens for Z_2, X → *Y* relationship remains (replicated) for Z_3, X → *Y* relationship remains (replicated) or any other combination where the relationship disappears, remains the same (replicated), or is reversed, in some but not all of the categories of the test variable
5. Suppressor	X → *Y* relationship is weak or none for Z_1, X → *Y* relationship appears for Z_2, X → *Y* relationship appears for Z_3, X → *Y* relationship appears because the control variable weakened or suppressed the original relationship

Intervening Variables

Now imagine that you find that a strong relationship occurs within each level of respondents' education. In other words, for those with less than high school education, there is now a strong relationship between parents' education and income; among those who are high school graduates, there is also a strong relationship between parents' education and income; and so on. What you have shown is that respondents' education was suppressing the original relationship because it intervenes between parents' educational level and income. Respondents' education is an *intervening* variable that was mediating and keeping the original relationship weak, since parents' education is only an indirect influence on respondents' income. Parents' education is highly correlated with respondents' education, which in turn predicts income.

Partial Correlations

When variables are interval/ratio measures, controlling for a test variable can be done using partial Pearson r correlation (partial Tau and partial Gamma are also available for ordinal measures). A *partial correlation* takes into account the relationship between *X* and *Y* when the control variable *Z* is introduced. It calculates the Pearson r correlation between *X* and *Y*, *X* and *Z*, and *Y* and *Z* and then recalculates the original zero-order bivariate relationship between *X* and *Y* after the effect of *Z* has been partially removed. For example, a Pearson r correlation of 0.084 was calculated for seniority on the job and current salary. That very low coefficient seemed to indicate that there was no statistically significant relationship between these two variables, somewhat counterintuitive to what you might think. After all, the longer people work someplace, the more raises they get and their salaries increase.

However, it's also possible that those who worked at the company the longest started at lower salaries than those who began more recently, given the realities of inflation. Starting salary might be an antecedent variable that could elaborate on the original relationship found. A partial correlation was calculated, removing the relationship between starting salary and current salary, and sure enough the original relationship increases to 0.214 (see Table 9.3). Starting salary (the control variable *Z*) was suppressing the original relationship between seniority and salary. Even though it is statistically significant, it remains somewhat low in strength, suggesting that there are other reasons besides seniority to explain current salary.

There are numerous ways to introduce additional variables for elaboration and to test various scenarios. For example, some categories of a control variable measuring political views (such as "Moderates") may be suppressing the original relationship, while other categories do not (such as "Liberals" and "Conservatives"). Or an original relationship might be a strong positive one, but after controlling for a third variable, it becomes a strong negative one in some categories of the control variable but remains positive in the others. The important point to remember is that in order to establish any kind of cause-and-effect relationship you must be able to rule out alternative plausible explanations; the process of elaboration is a good way of doing this and a good way of uncovering other kinds of relationships among your variables.

Table 9.3 Example of Partial Correlation, SPSS

CORRELATIONS

			Seniority Months Since Hire
	Current Salary	Pearson Correlation	0.084
		Significance (two-tailed)	.067
		N	474

Partial Correlation	Coefficients
Controlling for . . .	Beginning salary
	seniority
Current salary	0.2138
	$p = .000$

MULTIPLE RELATIONSHIPS

Using control variables is one kind of multivariate analysis, primarily focused on providing more elaborate understanding of initial relationships. Another important type is an analysis of additional independent variables that might contribute to a better and more complete understanding of the dependent variable or outcome. When these variables are interval/ratio measures or ordinal measures with equal-appearing intervals—instead of the nominal and ordinal ones better suited for cross-tabulations and the elaboration analyses just described—then some fairly sophisticated statistical procedures can be used.

Two-Way ANOVA

Let's say you are interested in the differences in mean income among several racial/ethnic groups. As you recall, for comparing means among three or more nominal or ordinal categories, use one-way analysis of variance (ANOVA). If the probability of obtaining the F-value is less than .05, you declare that a relationship exists between race/ethnicity and income and specify which groups obtain higher average salaries. Yet this seems too simple an explanation, so you introduce a third variable.

You could control for sex by rerunning the ANOVA for men and then another one for women, but another technique would be to introduce sex as a third variable (sometimes called a factor or covariate) and perform a *two-way analysis of variance*, or three-way or more, depending on how many new factors are included. In addition, you might want to see if there is another factor affecting the outcome that is a result of the possible combined impact of the two (or more) independent variables, not just their individual effects. This is what's called an *interactive effect*. For example, race/ethnicity might affect income, and sex might also have a main effect on income, but the combined impact of race/ethnicity and sex might be even stronger. The computer calculates the F-values for the main effects of each of the independent variables, in this example sex and race/ethnicity, along with the F-value of any interactive effects, sex by race/ethnicity. It is interpretable similarly to the F calculated in a one-way ANOVA. If F is statistically significant, the ANOVA would suggest that, for example, women of color receive lower salaries than white women, and so on. Details for doing multivariate and factorial ANOVA are available in advanced statistics books and online.

Multiple R

Another very common multivariate technique is *multiple correlation*. This is based on the Pearson r correlation coefficient and essentially looks at the combined effects of two or more independent variables on the dependent variable. These variables

should be interval/ratio measures, dichotomies, or ordinal measures with equal-appearing intervals, and assume a linear relationship between the independent and dependent variables. Rather than partialling out the impact of a third variable to see what remains of the original relationship, multiple correlation includes the effects of a third or fourth, or more, variable on the dependent variable. It is represented as the capital letter *R* and, similar to r, is a PRE (proportional reduction in error) statistic when squared. However, unlike bivariate r, multiple R cannot be negative because it represents the combined impact of two or more independent variables, so direction is not given by the coefficient.

What multiple R^2 tells us is the proportion of the variation in the dependent variable that can be explained by the combined effect of the independent variables. Let's say you found that current salary depends on three explanations: starting salary, seniority, and age of employee. Each of these independent variables is correlated with current salary, but you want to know their combined impact. The result is not simply the addition of the bivariate Pearson r correlations for each of these variables and the dependent one, because some variables, like age and seniority, for example, are correlated with one another, and this overlap needs to be taken into account mathematically.

Let's say the statistics program calculates a very strong positive multiple R correlation coefficient of 0.897 and an R^2 of 0.805. This tells you that approximately 80.5 percent of the variation in current salaries in your sample is accounted for by a combination of starting salary, seniority, and age. Which ones contribute more or less cannot be determined just by looking at the multiple correlation; however, as discussed in the next section, there is a way of getting this information. R^2 also tells you that if you wanted to predict salaries, you would reduce your errors in predicting those figures by around 80 percent knowing three things: what the starting salaries were, how long the employees worked for the company, and ages of the workers.

Multiple Linear Regression

Uncovering which independent variables are contributing more or less to the explanations and predictions of the dependent variable is accomplished by a widely used technique called *multiple linear regression*. It is based on the idea of a straight line that has the formula $Y = a + bX$, where

- Y is the value of the predicted dependent variable, sometimes called the criterion and in some formulas represented as Y' to indicate Y-predicted.
- X is the value of the independent variable or predictor.
- a is the constant or the value of Y when X is unknown, that is, zero; it is the point on the Y axis where the line crosses when X is zero.

- *b* is the slope or angle of the line and, because not all the independent variables are contributing equally to explaining the dependent variable, *b* represents the unstandardized weight by which you adjust the value of *X*. For each unit of *X*, *Y* is predicted to go up or down by the amount of *b*.

The Regression Line. Pearson r correlations and regressions assume linearity. If you were to construct a scatterplot, and the correlation was a perfect 1.0, then all the points representing a score on *X* and a score on *Y* would fall into a straight line. Most of the time, of course, a correlation is not perfect, so the points are more likely not to form a straight line. Therefore, through the scatterplot, a best-fitting line is drawn, around which the points are the closest. This is similar to the idea of a mean being the point around which the scores in a distribution are the closest. This best-fitting line is called the *regression line*, which predicts the values of *Y*, the outcome variable, when you know the values of *X*, the independent variables. Linear regression analysis calculates the constant (*a*), the coefficient weights for each independent variable (*b*), and the overall multiple correlation (R). Preferably, low intercorrelations exist among the independent variables in order to find out the unique impact of each of the predictors. This is the formula for a multiple regression line:

$$Y' = a + bX_1 + bX_2 + bX_3 + bX_4 \ldots + bX_n$$

Methods for Entering Variables. One method for generating a regression equation is to enter the independent variables in the order you believe—on the basis of theory or prior research—they contribute to explaining the dependent variable. Another is to use what is termed *stepwise* multiple regression, which lets the statistics program enter the variables, either individually or in blocks or groups of variables, in the order of their correlation to the dependent variable. It selects the strongest predictor first, then the next, and so on, one at a time from the entire list of independent variables (or from within each block) until a variable is not statistically significant to enter the equation, usually set at the .05 level of significance. Another rule of thumb is to be sure that each new variable entered contributes at least 1 percent to the overall explanation of the variance of the dependent variable.

Beta Coefficients. The result is an equation that can be used to predict values of the dependent variable. The information provided in the regression analysis includes the *b* coefficients for each of the independent variables and the overall multiple R correlation and its corresponding R^2. Assuming the variables are measured using different units as they typically are (such as pounds of weight, inches of height, or scores on a test), then the *b* weights are transformed into standardized units for comparison purposes. These are called *Beta (β) coefficients* or weights and essentially are interpreted

like correlation coefficients: Those farthest away from zero are the strongest, and the plus or minus sign indicates direction. If all the variables are measured in the same units, or if you are comparing a particular variable in one regression equation with the same one in another equation, then the unstandardized *b* coefficients are easily interpreted. With the Beta coefficients and the R^2, you have the most relevant information you need to arrive at some conclusions and possibly make some predictions.

A multiple regression provides a description of possible influences on the dependent variable. It is a statistical technique designed especially for explanatory and predictive research hypotheses. Recall from Chapter 1 that the goals of some research are to predict or to explain, not just to describe. Multiple regression is a robust statistical procedure that allows some flexibility in the kinds of variables used. Ideally, your measures are interval/ratio ones that have a linear relationship between the independent and dependent variables. If your measures are not interval/ratio, then determine whether any ordinal measures have equal-appearing intervals (such as Likert-type scales, or if income is in categories with equal ranges) or there are dichotomous variables with two categories.

However, you cannot use nominal measures with three or more categories. Typically, you have to create dichotomous or *dummy variables* for nominal variables, such as religion. Because there is no order to the religions listed in your questionnaire, you might recode your data into Religious, Not Religious; or Protestant, Not Protestant; or Christian, Not Christian; or Muslim, Not Muslim; or any other possible dichotomy, depending on the goals of your research. The same applies to such nominal demographic variables as race/ethnicity, sexual orientation, political party affiliation, and other variables with several nominal categories.

Let's hypothesize that there is a relationship between students' grades in college (CUM GPA) and their high school grades (HS GPA), total scores on a standardized test (TOTSAT), and their participation in extracurricular activities (Extracurricular), where a score of 1 = high involvement and 3 = low involvement. This might be a model used to predict admission to a university, where it is assumed that high grades, high test scores, and high involvement in activities represent the kind of well-rounded people an admissions office would like to recruit to its university.

The first step is to decide how you want to enter the independent variables into the regression analysis. One method is to list them in the order you want them to be entered, based on some theory or model. The order can make a difference: Given that high school grades and standardized test scores are likely to be correlated, if you enter high school grades first, then test scores will contribute less to the overall final prediction results. On the other hand, if you enter test scores first, then grades might get suppressed. Or you can let the computer program (such as SPSS or other data analysis software) assist you by selecting a stepwise method for entering the variables. With stepwise, the program calculates the unique contributions of the independent variables to the dependent variable and lists them according to their order of strength. Those that are not statistically significant,

based on a t-test and alpha level of .05, are not entered. Think of it as requiring an ID to get into the party: If you are not of "strength," you have to stay outside!

Stepwise multiple regression is selected in Tables 9.4a to 9.4d, and the results are reported in several ways. Although more advanced statistics books and websites go into detail about all the components of a regression equation and output, for our purposes, the information from the SPSS software most needed to interpret them is

Table 9.4 Example of Multiple Regression, SPSS

a **MODEL SUMMARY**

Model	R	R Square	Adjusted R Square	Standard Error of the Estimate
1	0.400[a]	0.160	0.159	0.56952
2	0.425[b]	0.180	0.179	0.56282
3	0.440[c]	0.194	0.192	0.55842

[a] Predictors: (Constant), HS GPA.

[b] Predictors: (Constant), HS GPA, TOTSAT.

[c] Predictors: (Constant), HS GPA, TOTSAT, Extracurricular.

b **ANOVA[a]**

Model		Sum of Squares	*df*	Mean Square	F	Significance
1	Regression	69.427	1	69.427	214.046	.000[b]
	Residual	364.579	1,124	0.324		
	Total	434.006	1,125			
2	Regression	78.279	2	39.139	123.559	.000[c]
	Residual	355.727	1,123	0.317		
	Total	434.006	1,125			
3	Regression	84.125	3	28.042	89.925	.000[d]
	Residual	349.881	1,122	0.312		
	Total	434.006	1,125			

[a] Dependent variable: CUM GPA.

[b] Predictors: (Constant), HS GPA.

[c] Predictors: (Constant), HS GPA, TOTSAT.

[d] Predictors: (Constant), HS GPA, TOTSAT, Extracurricular.

c **COEFFICIENTS[a]**

Model		UNSTANDARDIZED COEFFICIENTS *b*	Standard Error	STANDARDIZED COEFFICIENTS Beta	t	Significance
1	(Constant)	1.443	0.108		13.420	.000
	HS GPA	0.456	0.031	0.400	14.630	.000
2	(Constant)	0.852	0.154		5.526	.000
	HS GPA	0.426	0.031	0.374	13.595	.000
	TOTSAT	6.042E−04	0.000	0.145	5.286	.000
3	(Constant)	1.410	0.200		7.049	.000
	HS GPA	0.362	0.034	0.317	10.499	.000
	TOTSAT	5.343E−04	0.000	0.128	4.665	.000
	Extracurricular	−0.129	0.030	−0.131	−4.330	.000

[a] Dependent variable: CUM GPA.

d **EXCLUDED VARIABLES[a]**

Model		Beta in	t	Significance	Partial Correlation	COLLINEARITY STATISTICS Tolerance
1	TOTSAT	0.145[b]	5.286	.000	0.156	0.967
	Extracurricular	−0.151[b]	−4.991	.000	−0.147	0.796
2	Extracurricular	−0.131[c]	−4.330	.000	−0.128	0.780

[a] Dependent variable: CUM GPA.

[b] Predictors in the Model: (Constant), HS GPA.

[c] Predictors in the Model: (Constant), HS GPA, TOTSAT.

discussed here. (Although other statistic software packages may have different formats and information, the explanations here should help in interpreting findings in those programs and as presented in published research like Box 9.2.) The section numbered Table 9.4a is the overall summary of the regression model that has been calculated. Because it is stepwise, it presents the model in steps. Model 1 shows "HS GPA" entered first (according to footnote a). Its correlation (R) with the dependent variable is 0.400, signifying a moderate to strong relationship. This one variable explains 16 percent of the variance in grades, according to R square (R^2). It is also at this point the same as the bivariate Pearson r between high school grades and college grades.

BOX 9.2
MULTIPLE REGRESSION ANALYSIS IN AN ACADEMIC ARTICLE

Teenagers everywhere appear to be tied to their cell phones and other media, often at the same time. Yang and Zhu (2016) decided to research predictors of media multitasking among Chinese adolescents. They hypothesized that media-oriented families would predict more multitasking. Family media orientation was measured using the following items (Yang and Zhu 2016: 433): "Can you see the television when your seat is in front of the computer?" and "Does your family do housework or other tasks with the TV on?" Responses were made on a 4-point scale from "very often" to "never." Children's media ownership was measured using the following items: "Do you have a mobile phone?" "Is your mobile phone a smartphone?" and "Do you have a television in your bedroom?" Media multitasking is measured with a Media Use Questionnaire and calculates how much multitasking occurs during a typical media-consumption hour.

Here are the results of one linear regression analysis:

REGRESSION OF FAMILY MEDIA ENVIRONMENT ON MEDIA MULTITASKING

Predictor Variables	*b*	Standard Error	β	t
Owns smartphone	0.57	0.29	.11	1.99*
TV in bedroom	0.62	0.23	.16	2.69**
Computer placement	0.33	0.22	.10	1.54
Housework with TV on	0.32	0.10	.20	3.38**
$R^2 = 0.14$				

n = 310

* $p < .05$, ** $p < .01$

The Beta (β) coefficients for owning a smartphone, having a TV in the bedroom, and family members doing housework and other tasks with the television on are statistically different from zero (based on the t-tests). Together, all four variables studied explain 14 percent (R^2) of the adolescents' multitasking scores, with television in the bedroom being the strongest predictor and placement of the computer not having a statistically significant impact on multitasking. How much would these variables explain media multitasking in other cultures beside China?

Then the computer program looks for the second strongest predictor or independent variable. Model 2 finds "TOTSAT" and, as footnote b reminds you, at this step, it is high school grades and total SAT scores *combined* that correlate 0.425. R^2 tells

you that 18 percent of the variance in CUM GPA in this sample can be explained by respondents' high school grade point averages and total test scores, although including test scores increased variance explained by only 2 percent. This small increase in explanation probably occurred because grades and test scores are themselves related, thereby illustrating the partial effects of a new variable. If high school grades were not included in the list of independent variables, perhaps test scores would have correlated even higher with college GPA.

The statistics program keeps calculating and tries a third time, and model 3 shows the arrival of "Extracurricular" to the equation. Multiple R increases to 0.440, and over 19 percent of the explanation of college grades is determined by these three variables together.

Table 9.4b tests the significance of the models that have been generated. Using ANOVA, an F-value is calculated for the regression equation at each step. As can be seen, the final model 3 is a statistically significant regression equation ($p < .000$). The footnotes also remind you of the dependent variable and the independent or predictor variables at each step.

By this point, the overall R^2 is known, but you still want to figure out how each of the independent variables is contributing to this R^2 in explaining or predicting the criterion or dependent variable. Each of them is not equally related or predictive of college grades, especially because there is some interrelationship among the independent variables themselves (called *multicollinearity*). The increase in R^2 at each step suggests that a good deal of the relationship is carried by the first variable, high school grade point average. Nonetheless, the other variables still contribute something, but they need to be given less value, or carry less weight. Because high school grades, test scores, and extracurricular participation are measured using different units (grades go from 0.0 to 4.0, combined test scores range from 400 to 1,600, and extracurricular is measured on a 3-point scale from 1 to 3), the standardized Beta coefficients are used to compare the influence of each variable instead of *b*, the unstandardized ones.

Table 9.4c provides the Beta coefficients for all the independent variables entered at each step in the creation of the model, along with a t-test value and its significance level that tests whether the Beta coefficient is different from zero. As the table shows, at step or model 3, high school GPA has a coefficient of 0.317, total SAT has 0.128, and extracurricular has a Beta weight of −0.131. Similar to other correlation coefficients, the negative sign tells you that college grades and participation in extracurricular activities are inversely associated; in this case, those who score low on the extracurricular measure (that is, who selected 1 on the questionnaire where 1 = highly involved in activities) are more likely to score higher on the college GPA measure. However, here it is an inverse association because of the scoring; if 3 was originally written to be highly involved in extracurricular activities, then the Beta weight would

have been 0.131. The interpretation is that those highly involved in extracurricular activities tend to have higher GPAs. High school grades and test scores have positive relationships with college GPA.

In addition, Table 9.4d also shows you which variables have been excluded at each step and what their Beta weights would be *if* they are entered in the next step. Notice how at step one under Excluded Variables, the "Beta In" for TOTSAT is 0.145, and sure enough, at model 2 of the coefficients table (Table 9.4c), it enters with a Beta weight of 0.145. Should there be some that never enter the models, as happens with stepwise technique, these would be listed at the last step under "Excluded Variables" with the Beta coefficients they would have had if they had been entered into the model. Sometimes this information can tell you which variables may have been better predictors but were kept out of the model due to intercorrelation with the other variables.

Putting all the findings into words, this regression analysis tells you that around 19 percent of the variation in college cumulative GPA among students in this sample can be explained by knowing their high school grades, SAT scores, and extracurricular involvement. Those in the sample who have higher grades in high school, have higher SAT scores, and are more involved in extracurricular activities tend to have higher college grade point averages. If SAT scores had a negative or inverse coefficient (which they don't), you would conclude that lower SAT scores are associated with higher college grades, and higher SATs are related to lower grades.

The actual equation for the regression line uses the *b* weights and looks like this: $Y' = 1.410 + (0.362 \times X_1) + (0.000534 \times X_2)—(0.129 \times X_3)$, where Y' is the predicted college GPA (dependent variable), X_1 is high school GPA (independent variable 1), X_2 is total SAT scores (independent variable 2), and X_3 is the extracurricular involvement score (independent variable 3). Knowing nothing else, your predicted college GPA would be the value of the constant, 1.410, but your actual prediction is adjusted because you do have some information on the independent variables. High school GPA is giving you stronger assistance in predicting college GPA, compared to both total SAT scores and extracurricular participation, but all three together help out in explaining something about why people get different grades in college. Assuming you are similar to the characteristics of the sample completing the questionnaires that led to the creation of this regression model, you could insert your high school grades, total SAT scores, and estimated participation in high school activities (on a scale of 1 to 3 where 1 = high) and calculate a predicted college grade point average for yourself.

Normally, you don't use a regression equation to calculate information about specific individuals, although many universities use such a model for admissions to predict potential success in college for individual applicants, and many businesses perform regression analyses to estimate future sales about a specific product.

Regression analysis does not tell you about any one particular respondent, since the statistics are based on aggregated data. Mostly what you do with regression is construct a profile of characteristics related to the dependent variable from past data and use that to explain what already exists or to predict subsequent outcomes.

The regression does not tell you why any of the findings are so. You can't tell from these statistics why participation in extracurricular activities improves your chances for higher grades in college. Unless you have other data from the questionnaire that can address this query, all you can do when you write up your results is offer some possible explanations based on theory or previous research. Or you can give some speculative interpretations and suggest that maybe those who are more involved in nonacademic pursuits are well-rounded people who exhibit commitment to, interest in, and curiosity about a wide range of issues. Perhaps curiosity and involvement translate into good study habits leading to better grades. Further research would investigate these new hypotheses and research questions, thus illustrating the continuous cycle of inductive and deductive social science research, as discussed in Chapter 1.

There are a variety of other widely used multivariate techniques, such as path analysis, factor analysis, MANOVA, and logistic regression analysis, that are beyond the scope of this introductory book. But the underlying concept is similar: How do two or more independent variables work together in assisting you in making sense of the variation that exists in the dependent variable? How can we account for differences in the dependent variable, knowing two or more independent variables? Similar to other statistical procedures, values are calculated and the probability of obtaining those values by chance is reported. If the probability is fairly small that chance played a role, then you are more confident in declaring results that deserve a big announcement at a press conference or at least a Facebook status posting!

The research journey nears an end. You have completed data analysis and are now ready to put your findings into words and tell others what you have discovered. How to interpret your findings and write a report is the focus of the last chapter.

REVIEW: WHAT DO THESE KEY TERMS MEAN?

Antecedent
Beta coefficients
Dummy variables
Elaboration
Interactive effects
Intervening
Multicollinearity
Multiple correlation
Multiple linear regression
Partial correlation
Partialling
Regression line
R square
Specification
Spurious
Stepwise
Suppression
Two-way ANOVA
Zero-order

TEST YOURSELF

Here are the results of a multiple regression analysis predicting which people have the happiest outlook on life (where 1 = very happy, 2 = pretty happy, and 3 = not too happy). Variables used include sex (1 = male and 2 = female), age (18 to 90), and years of school completed (number of years from 0 to 20).

MODEL SUMMARY

Model	R	R Square	Adjusted R Square	Standard Error of the Estimate
1	0.161[a]	0.026	0.023	0.606

[a] Predictors: (Constant), respondent's sex, age of respondent, highest year of school completed.

COEFFICIENTS[a]

		UNSTANDARDIZED COEFFICIENTS		STANDARDIZED COEFFICIENTS		
Model		*b*	Standard Error	Beta	t	Significance
1	(Constant)	2.255	0.111		20.370	.000
	Age of respondent	−0.002	0.001	−0.067	−2.394	.017
	Highest year of school completed	−0.024	0.007	−0.114	−3.494	.000
	Respondent's sex	0.045	0.033	0.036	1.359	.174

[a] Dependent variable: General happiness.

1. List the significant independent variables in order of strength. Do not include any that are statistically insignificant.
2. Interpret R and R^2 for a statistical audience.
3. Using the Beta coefficients, put into words (for a general audience) a profile of those respondents who tend to be the happiest.

INTERPRET: WHAT DO THESE REAL EXAMPLES TELL US?

1. Riggle et al. (2010) studied same-sex legal relationships in order to see if previous research suggesting that married adults experience less psychological stress and higher levels of well-being also applied to lesbian and gay relationships. Comparing

those lesbians and gay men in committed relationships with those in legally recognized same-sex relationships (domestic partnerships, civil unions, and civil marriages), the researchers reported the following results for a measure of stress:

	STRESS		
Independent Variable	*B*	Standard Error	β
Sex	0.481	0.189	0.077*
Education	0.286	0.083	−0.104**
Parent	0.092	0.198	0.014
Relationship length	0.038	0.12	−0.098**
Relationship status	0.675	0.215	−0.095*
Adjusted R^2 = 0.046			

Note: The following categories were used: sex: 0 = male, 1 = female; education: 1 = less than high school degree, 5 = PhD or professional degree; parent: 0 = do not have children, 1 = have children; relationship status: 0 = in a committed relationship, 1 = in a legally recognized relationship. Relationship length was measured in years. Stress is measured on a 5-point Likert-type scale where lower scores mean less stress experienced in the previous month.

* $p < .01$, ** $p < .001$

a. State the hypothesis being tested. Why is regression appropriate to use?
b. What do the Beta coefficients and significance levels tell you?
c. Explain R^2 and how it can be interpreted.
d. Put into words what the study has discovered.

2. Consider the hypothesis that there is a relationship between sex of respondents and whether or not they graduate from college. Table 9.5 shows a cross-tabulation of sex and graduation for a sample of 555 students at a university.
   a. Put into words what the table says about the relationship between graduation rates and sex.
   b. What do the chi-square and significance level indicate?
      However, you feel there may be something else happening to explain graduation rates and wonder if introducing a control variable could elaborate your findings. Because other studies have shown that high school grades are a good predictor of college success, you control for them. Table 9.6 shows the results.
   c. Explain what each subtable says in words.
   d. What do the chi-square statistics tell you?
   e. What do you conclude about the control variable and the original relationship?

Table 9.5 Cross-Tabulation

GRADUATE BY SEX CROSS-TABULATION					
			SEX		
			F	M	Total
Graduate	No	Count	103	101	204
		% within sex	32.2%	43.0%	36.8%
	Yes	Count	217	134	351
		% within sex	67.8%	57.0%	63.2%
Total		Count	320	235	555
		% within sex	100.0%	100.0%	100.0%

Chi-square = 6.79, *df* = 1, *p* = .009

Table 9.6 Cross-Tabulation of Graduate by Sex, Controlling for Grades

				SEX		
HS Grades				F	M	Total
B or lower	Graduate	No	Count	43	59	102
			% within sex	40.2%	45.4%	43.0%
		Yes	Count	64	71	135
			% within sex	59.8%	54.6%	57.0%
	Total		Count	107	130	237
			% within sex	100.0%	100.0%	100.0%
B+ or higher	Graduate	No	Count	60	42	102
			% within sex	28.2%	40.0%	32.1%
		Yes	Count	153	63	216
			% within sex	71.8%	60.0%	67.9%
	Total		Count	213	105	318
			% within sex	100.0%	100.0%	100.0%

For the high school grades B or lower subtable: chi-square = 0.647, *df* = 1, *p* = .421

For the high school grades B+ or higher subtable: chi-square = 4.518, *df* = 1, *p* = .034

CONSULT: WHAT COULD BE DONE?

CNN calls you to be on one of its news shows to discuss election results. The interviewer is planning to ask you about what seems to be a difference in the income of people who voted Democratic, Republican, or for other parties. You are the expert and are going to be asked if this is so.

1. How can you be sure that income is related to political party choice? Describe the steps you would take to verify that relationship. What other variables might be relevant?
2. You will also be asked for whom people might vote the next time around. Just focusing on Democratic or Republican, what variables would you use, and what would you do statistically to predict voting choice?

DECIDE: WHAT DO YOU DO NEXT?

For your study on how people develop and maintain diverse friendships, especially on social media, respond to the following items:

1. Imagine you have found some relationship between race/ethnicity and number of friends. Suggest several test or control variables you could use to elaborate on the original relationship. Be sure to include at least one antecedent and one intervening variable.
2. Look over your hypotheses and develop one to make it more suitable for a multiple regression. Which variables would you include? How would you recode any variables for a regression?
3. If you have actual data, run a regression, and interpret the results in words. Also, illustrate the process of elaboration, and show how you control for a third variable with a series of crosstabs.