13

Regression Analysis

Learning Objectives

In this chapter you will

1. Understand and use bivariate and multiple linear regression analysis
2. Understand the concept of the regression line and how it relates to the regression equation
3. Understand the assumptions behind linear regression
4. Be able to correctly interpret the conceptual and practical meaning of coefficients in linear regression analysis
5. Be able to use SPSS and Excel to conduct linear regression analysis

Predicting Relationships

Emily's Case

"It was a great conference," Leo exclaimed as he slipped into the backseat of Emily's car.

Mei-Lin agreed enthusiastically as she got in the front passenger side. "This was really good. Thank you, Emily."

"My pleasure," Emily replied with a laugh as she settled behind the driver's wheel. "People liked Leo, don't you think?"

Mei-Lin turned toward the back. "I was so proud of you!"

"It's true," Emily continued, "you did a great job making the statistical analysis on the impact of the diversity training understandable. HR professionals are not usually into statistics, but I think they liked it."

As they drove back to Westlawn, they talked about the things they learned at the conference. Emily and Mei-Lin were particularly interested in a presentation where the speaker talked about the cumulative effects of training and employee education. The point was that a one-time training is not enough to make an impact on employee development. The speaker emphasized the importance of having a long-term strategic plan for training and employee education and to track the results.

"She made a good point," Mei-Lin argued. "We need to continue the diversity trainings if we really want to make an impact."

Emily smiled. "We just need to secure the resources to keep it going."

Leo leaned forward between them. "You know, it occurs to me that this 'cumulative effect' of training; if it's true, it should show up in our survey data." Neither Emily nor Mei-Lin responded, so Leo explained what he meant. "We have a question that asks how many diversity trainings the employee has attended in the past. We had quite a few who responded. I wonder if we can predict the level of cultural competence by the number of diversity trainings they attended in the past. If it's cumulative, then the level of cultural competence should be higher with the people who attended more trainings, right?"

Mei-Lin picked up on the significance of the idea first. She was writing a proposal to justify a new round of diversity training, and this looked like a useful piece of evidence.

"That's a great idea, Leo." Mei-Lin responded. "I would really like to see what that looks like. Is it easy to run that analysis?"

"Hey, he's Leo," Emily joked. They all laughed.

"Sure, I can do that pretty quickly," Leo confirmed. He was already curious. "I'll get on it tonight."

Mary's Case

Mary was a little frustrated that nobody at Health First showed much interest in her research-based approach to volunteer recruitment and retention. She needed a sounding board. Yuki, her grad-school friend who headed the research department at one of the major foundations in the area would certainly understand. She had helped Mary get started on this project. Mary sent Yuki an invitation and met her the next day at a coffee shop about halfway between their two offices.

"It's such a nice day, let's sit outside," Yuki said, holding a latte in her hand.

As soon as they sat down at a metal table outside the coffee shop, Yuki jumped right to the topic she knew was on Mary's mind. "So, how's your research on the volunteers going?"

Typical Yuki style, Mary thought. No "how's your parents?" or "how's your boyfriend?" or "how's your dog?" niceties. This was one thing she liked about Yuki. She smiled appreciatively as she responded.

"I read your qualitative research books. Thanks for loaning them to me. I want to keep them a little longer, if you don't mind."

"That's fine," replied Yuki.

Mary sipped her coffee and continued, "I've been spending a lot of time thinking about whom I should interview and what questions I should ask."

Yuki nodded and said, "As you should be."

"In the mean time, I obtained data from HR on the background of the volunteers, and I've been analyzing it." Mary told Yuki about the correlation and chi-square analyses she had conducted using the volunteer profile data.

Yuki answered with a knowing look: "Not surprising you are doing statistics. You like quantitative data. Sounds like you are getting interesting results."

Encouraged by Yuki's interest, Mary shared her thoughts on another analysis she was thinking about. "At Health First, we don't have much money to put into a major volunteer recruitment campaign, so we need to focus our resources on the most efficient ways to recruit volunteers." Mary paused to be sure Yuki was following.

"Go on," Yuki encouraged.

"We don't have very much information about the current volunteers, but we do ask when they start how many hours they are willing to put in for their volunteer work. It appears from feedback from other managers that the number of hours the volunteers actually work is pretty close to what they said they would work. So—"

Yuki leaned forward, anticipating the punch.

"—I thought, in addition to increasing the number of volunteers themselves, I should focus on volunteers who are willing to work more hours. That gives us a better return on our investment."

Yuki laughed and said, "I can't believe you use a phrase like 'return on investment' about the volunteers. You were always so opposed to the business approach in nonprofit management." She saw Mary was a little startled. "But what you say makes sense."

Yuki, looked down and stirred her latte, and continued, "So, I suppose this idea of yours means you have a new plan for your research?"

"You are right on." Mary was glad she had a friend she could talk to about research without worrying about coming across too geeky. She opened up: "Literature suggests that people who have more income tend to volunteer more. I don't know if that's the case with the volunteers in our region, but I was thinking about running a regression analysis that predicts volunteer hours with our volunteers' level of income."

"Interesting idea," Yuki nodded. "What other volunteer background information do you have? Do you know their age and gender, and anything else? You can run a multiple regression and see which volunteer background information predicts volunteer hours significantly and also has the strongest relationship with the volunteer hours."

"Brilliant idea, Yuki!" Mary exclaimed. She pursued Yuki's suggestion, and the two friends blithely slipped into geekdom, talking about regression analysis. They did not notice their drinks getting cold.

Linear Regression Analysis

In Chapter 11, we introduced a way to examine the relationship of two continuous variables. In this chapter, we will build on this idea with an analytical tool, called **linear regression** analysis that uses correlation as a basis to predict the value of one variable from the value of a second variable or the combination of several variables.

In regression analysis, the variable that the researcher intends to predict is the **dependent variable** (sometimes called **outcome variable** or **criterion variable**). Typically the notation "Y" is used to describe the dependent variable. The variable

that the analysis uses to predict the value of the dependent variable is the **independent variable** (sometimes called *predictor variables*). The notation "X" is used to describe the independent variable. Linear regression analysis provides information about the strength of the relationship between the dependent variable and independent variable. When there is only one independent variable in the regression analysis, it is called **bivariate** (or **simple**) **linear regression analysis**. When there are two or more independent variables involved in the analysis, it is called **multiple regression analysis**.

In Mary's case, she is considering using bivariate linear regression analysis to predict volunteer hours (dependent variable) with the volunteers' income level (independent variable). Yuki suggested a *multiple regression analysis* to predict volunteer hours (dependent variable) with not only the income level, but also age, gender, and other information that might be available on the volunteers (independent variables). By examining the relative strength of the relationship of each independent variable with the dependent variable, Mary can identify the kind of volunteers she needs to maximize volunteer hours.

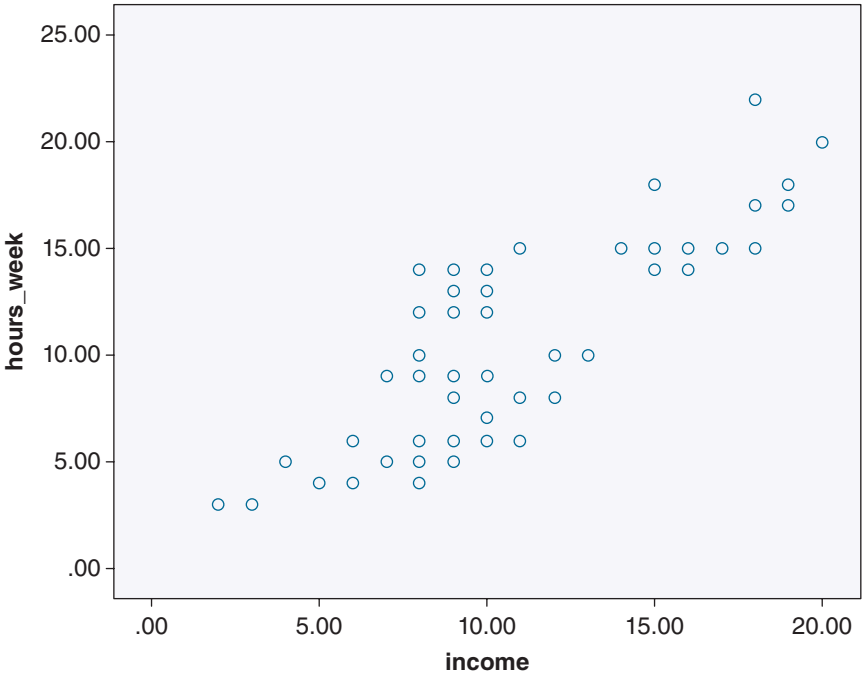
As with all statistical tests we introduced in this book, linear regression analysis is also based on a set of assumptions (Fox, 1991; Kahane, 2008), as follows:

1. **Linearity**: The relationship between the dependent variable and the independent variables are linear in nature.
2. **Normality**: The dependent variable is measured as a continuous variable and is normally distributed. The basic form of linear regression also assumes that the independent variables in the linear regression are continuous and are normally distributed. There are ways, however, to incorporate and interpret categorical independent variables in the regression analysis as **dummy variables.**
3. **Homoscedasticity**: The word *homoscedasticity* is derived from the Greek *homo* for same and *skedastickos* for dispersion (Merriam-Webster, 2012). It means having *same variance*. This assumption requires the degree of random noise in the dependent variable to remain the same regardless of the values of the independent variables.

Regression Equation and Regression Line: Basis for Prediction

Let's use Mary's example to illustrate the logic of prediction. Starting with her first question, she wants to predict volunteer hours (dependent variable) based on the volunteers' level of income (independent variable). Basically, prediction means estimating an unknown outcome based on a known outcome (Upton & Cook, 2011). For Mary to predict the unknown volunteer hours using income level information, she can use the known pattern of the relationship between volunteer hours and income level. So let's look at the pattern of the relationship between volunteer hours and income level with a scatterplot (Figure 13.1).

Figure 13.1 Scatterplot of the Volunteers' Income Level and the Volunteer Hours



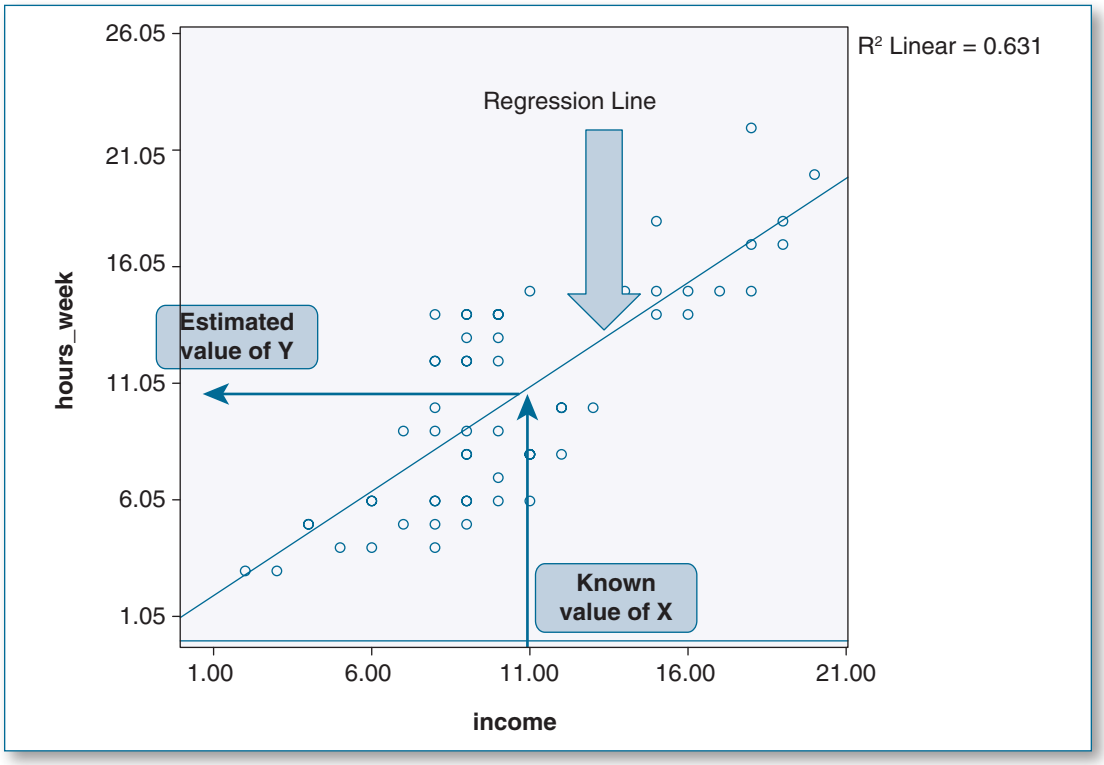
On the scatterplot, the value of the dependent variable is plotted on the Y-axis and the value of the independent variable is plotted on the X-axis. You can draw a line through the scatterplot to represent the minimum distance between the line and each one of the actual points. This is called a **regression line**. As you can see in Figure 13.2, once you identify the regression line, then you can use the line to estimate what the dependent variable (Y) would be if you know the value of the independent variable (X). A regression line is also called the **line of best fit** (Munro, 2004), because it is the line that best represents the pattern of the relationship between the dependent variable and the independent variable.

Once we identify the pattern of the relationship between the dependent variable Y and the independent variable X as a regression line, then we can describe the line with a formula. The basic equation for the regression line is as follows:

Formula 13.1 Basic Equation for the Regression Line

$$Y = a + bX$$

Figure 13.2 Scatterplot of the Volunteers' Income Level and the Volunteer Hours With Regression Line



Where

Y=the dependent variable (value on the vertical axis)

X=the independent variable (value on the horizontal axis)

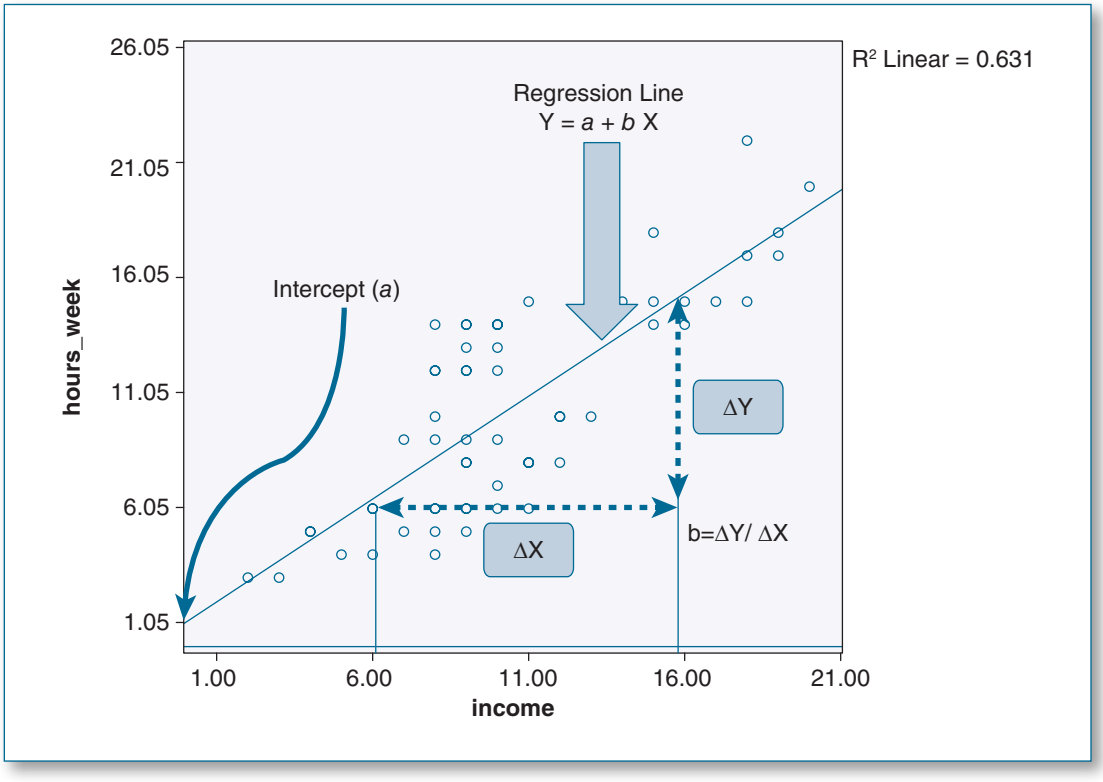
a=the point where the regression line crosses the Y axis, called the **intercept** (the value of Y when X is zero)

b = the **slope** of the regression line, indicating how much the Y value changes when there is a one-unit change in the value of X. It indicates the strength of the relationship between X and Y (the **regression coefficient**).

The relationship between the equation and the visual representation of the regression line is presented in Figure 13.3.

The slope of the regression line can be positive (+) or negative (–). When the slope is positive, that means the line goes up toward the upper right corner. When the slope is negative, that means the line goes down towards lower right corner. In other words,

Figure 13.3 Relationship Between the Regression Equation and the Visual Representation of Regression Line



the slope also indicates the direction of the relationship between X and Y. The intercept (a) and the slope (b) can be calculated based on the value of X and Y.

The formula for the slope (b) is:

Formula 13.2 The Formula for the Slope (b) of a Regression Line

$$b = \frac{\sum XY - (\sum X \sum Y)/n}{\sum X^2 - [(\frac{\sum X)^2}{n}]}$$

Once you have the slope (b) then you can use it to calculate the intercept (a):

Formula 13.3 The Formula for the Intercept (a) of a Regression Line

$$a = \frac{\sum Y - b\sum X}{n}$$

Of course, you don't have to calculate the slope and the intercept by hand. The SPSS and the Excel programs can do the calculation for you. In Mary's example, it turned out that the intercept (*a*) is 1.05 and the slope (*b*) is .90 (which you will see in the SPSS and Excel output). That means, in Mary's example, the regression equation is:

$$Y = 1.05 + .90\ X$$

Once Mary obtains this regression formula, she can plug in a volunteer's income level and predict how many hours this particular person is likely to volunteer. For example, with this regression equation, Mary can expect that if a volunteer reports an income level as "5" ($60,000 to $70,000), then the estimated number of volunteer hours will be 5.55 hours per week, as shown in the calculation below.

$$\begin{aligned} Y &= 1.05 + .90\ X \\ &= 1.05 + .90 * 5 \\ &= 1.05 + 4.5 \\ &= 5.55 \end{aligned}$$

Assessing the Prediction: Coefficient of Determination (R^2)

Once we identify the regression line, it is important to assess how well it predicts an outcome from the basis of a known variable. You can see from the scatterplot that the dispersion of the points will affect how accurate the estimate is likely to be. With this predictive model, we calculate a **coefficient of determination (R^2)** to measure how much of the variance in one variable is explained by variance in another.

R^2 is obtained by examining first how much the actual score in the dependent variable differs from the mean. This gives us a familiar measure of variance, with a **total sum of squares (**denoted SST). Then we measure how much the actual score in the dependent variable differs from the value estimated by the regression equation. This is called a **residual sum of squares (**denoted SSR).

The formula for R^2 is:

Formula 13.4 Calculating the Coefficient of Determination (R^2)

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

From the formula, you can see that R^2 will take a value between zero and 1. The closer the R^2 is to 1, the better the prediction. When R^2 is 1, the regression equation has a perfect prediction.

Let's unpack these concepts a little. In Mary's case, she could get an idea of how many volunteer hours to expect from her volunteers by looking at the mean. However, the actual hours put in by most of the volunteers will probably not equal the mean. The difference between the mean and the actual volunteer hours represents what we called

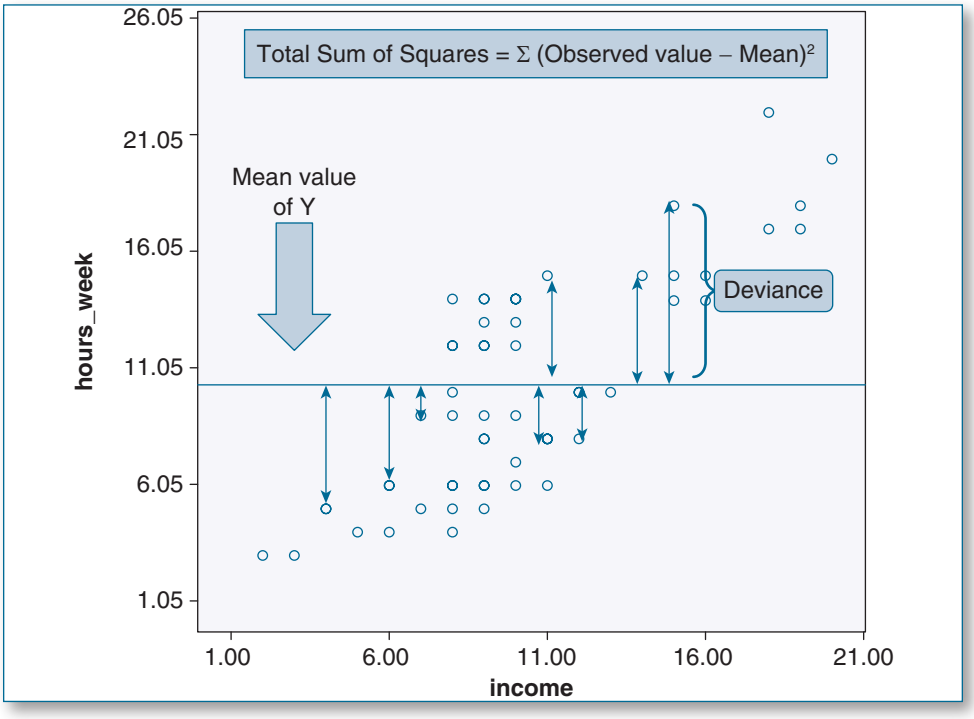
deviance in the discussion of measures of central tendency and variance in Chapter 7. Here we think of the same concept as an error in prediction when using the mean to predict. Remember earlier that we used a sum of squares to measure variance, because otherwise the sum of the plus and minus differences from the mean cancel each other out and always add up to zero. We use the same procedure here. This produces a total sum of squares (SST), as represented in the following formula (Formula 13.5) and illustrated in Figure 13.4:

Formula 13.5 Calculating the Total Sum of Squares (SST)

$$SST = \Sigma (\text{Observed value} - \text{Mean})^2$$

The regression equation identifies the line that minimizes the distance between the line and the observed values. The regression line offers a more sophisticated approach for prediction than just using the mean, but the prediction still does not perfectly match the observed values. There is still some inaccuracy. The differences are referred to as the **residuals,** or the **error in prediction**. Similar to deviance, summing up the residuals will result in a zero value because the directions of the differences cancel out. Therefore, we square the residuals before we add them all up to capture the overall error in prediction in the regression equation. This total is referred

Figure 13.4 Visual Representation of the Total Sum of Squares (SST)



to as the sum of residuals (SSR), as represented in the following Formula 13.6 and illustrated in Figure 13.5:

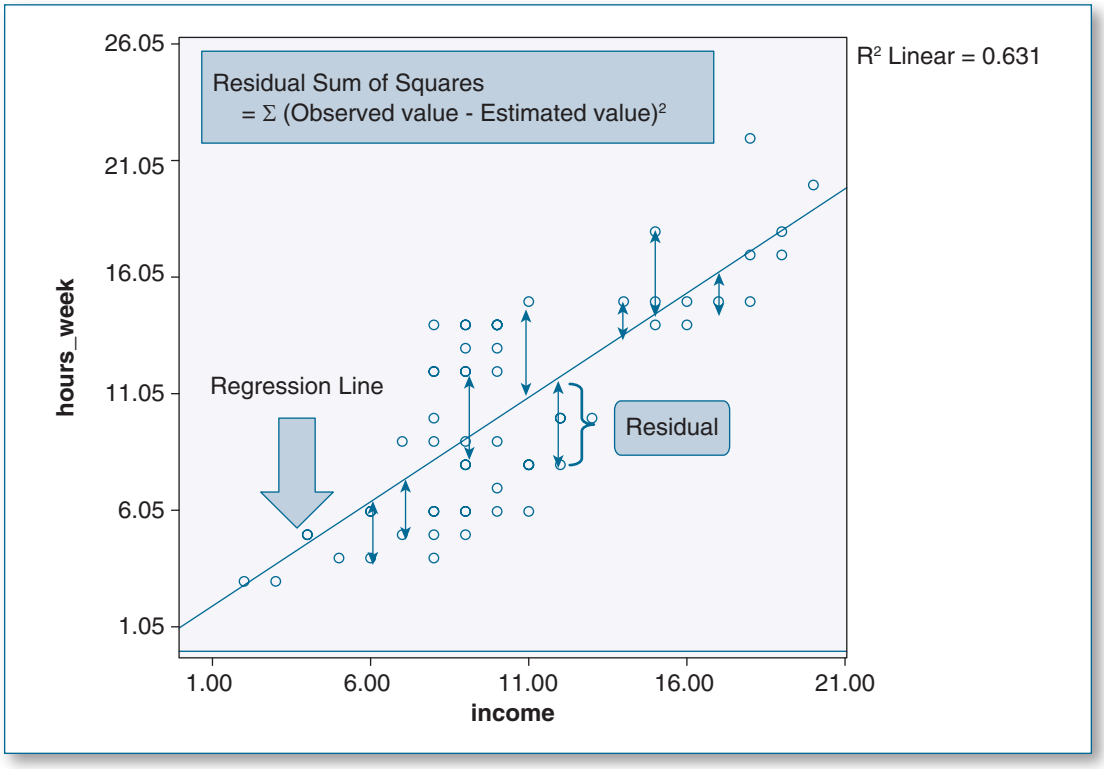
Formula 13.6 Calculating the Total Sum of Residuals (SSR)

$$SSR = \Sigma \text{ (Observed value} - \text{predicted value)}^2$$

The assessment on how well the regression equation predicts one outcome from another can be determined by calculating R^2. In the formula for R^2, we see that the quotient for SSR/SST—the sum of squares residual (SSR) over the sum of squares (SST)—will equal 1 if SSR and SST are exactly the same, meaning R^2 (=1–SSR/SST) will be zero. This result would indicate that the prediction using the regression equation is no different from the prediction using the mean and did not improve the prediction. When the SSR is smaller than the SST, then SSR/SST will be less than 1, and R^2 (=1–SSR/SST) will be greater than zero, meaning the prediction using the regression line is incrementally better than the prediction using the mean. When R^2 is closer to 1, the prediction is better (Field, 2009).

R^2 can also be explained as a measure of association between the dependent variable and the independent variable. It indicates the proportion of variance explained in

Figure 13.5 Visual Representation of the Sum of Squares (SSR)



the dependent variable by variance in the independent variable. When R^2 is zero, that means none of the variance is shared between the two variables. They are unrelated. When R^2 is 1—which would only be possible if the sum of residuals (SSR) equaled zero—then 100% of the variance is shared. This would mean that an exact prediction of the value of one variable would be possible by knowing the value of the other. Intermediate values for R^2 provide a good measure of the degree of the relationship between the independent and dependent variables (Pedhazur, 1997).

The null hypothesis for R^2 would state that there is no relationship between the independent and dependent variables. We can test the null hypothesis by calculating an F-statistic, as we did with ANOVA in Chapter 10. If the result of the test is significant, with the *p*-value below .05, then we reject the null hypothesis that R^2 is zero and accept the research hypothesis that R^2 is significantly different from zero, and there is a relationship between the independent and dependent variables in the population (Cohen, 2010).

Assessing Individual Predictors: Regression Coefficient (*b*)

In the regression equation, the independent variable X that we use to predict the value of Y has a coefficient (*b*). In the bivariate regression analysis, where there is only one independent variable X, the value of *b* represents the slope of the regression line. It indicates the change in the dependent variable Y, when there is a one-unit change in the independent variable X. When the regression coefficient *b* is zero, then a unit change in the value of the independent variable X results in no change in the dependent variable Y. In the regression analysis, we can conduct a t-test to test the null hypothesis that the regression coefficient *b* is zero. If the result of the t-test is significant, with a *p*-value below .05, then we reject the null hypothesis that *b* is zero and accept the research hypothesis that *b* is significantly different from zero. This means the independent variable X significantly contributes to the value of the dependent variable.

Running Bivariate Regression Using Software Programs

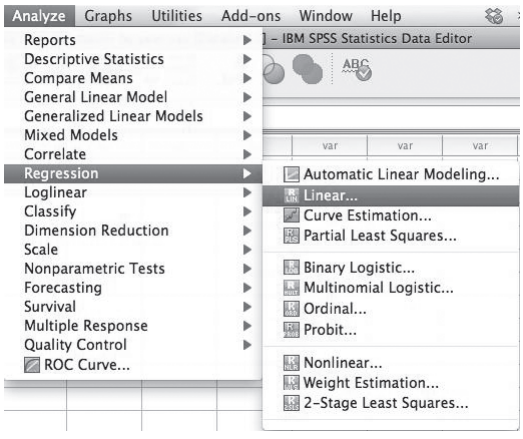
Let's take a look at Mary's case to see if she can predict volunteer hours by volunteer income level, using a bivariate regression analysis. We will go through the procedure in SPSS and then in Excel.

Running Bivariate Regression Using SPSS

The following steps outline how Mary will examine the relationship of volunteer hours to level of income with a bivariate regression analysis in SPSS:

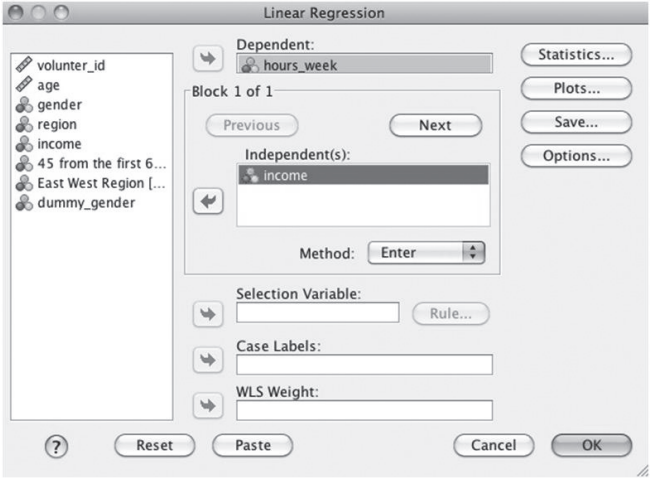
1. Open **Mary_Volunteer_profile.sav**
2. Click **Analyze→Regression→Linear.**

Figure 13.6 Menu Selection for Linear Regression



3. Move the variable **hours_week** into the Dependent Variable box.
4. Move the variable **income** into the Independent Variable box.
5. Click **statistics** and check **Descriptives** box.
6. Click **OK**.

Figure 13.7 Input Variables for Linear Regression in SPSS



There will be multiple tables in the output, including the descriptive statistics of the variables in the analysis. The table labeled *Model Summary* (Figure 13.8) includes information about R, R square (R^2), and Adjusted R Square.

Figure 13.8 Bivariate Linear Regression Model Summary Output From SPSS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.795[a]	.631	.625	2.84580

a. Predictors: (Constant), income
b. Dependent Variable: hours_week

R is the square root of R^2. We introduced R in Chapter 11 as the Pearson product moment correlation coefficient, indicating the strength and the direction of the linear relationship between the dependent variable (volunteer hours) and the independent variable (income level). In Mary's data, volunteer hours and volunteer income level are positively correlated, and the strength of the relationship is strong at .795.

R-Square (R^2) in Mary's analysis is .631, which suggests that volunteer income level explains 63.1% of the variance of their volunteer hours. This indicates that the relationship between volunteer income level and volunteer hours is moderately strong.

Adjusted R-Square (R^2) adjusts the value of R^2 when the sample size is small, because an estimate of R^2 obtained when the sample size is small tends to be higher than the actual R^2 in the population. The rule of thumb is to report adjusted R^2 when it substantially differs from R^2 (Green & Salkind, 2010). In this analysis, the difference is very small (adjusted $R^2 = .625$). Therefore, Mary can report the unadjusted R^2.

The SPSS output table labeled *ANOVA* (Figure 13.9) provides the results of a test of significance for R and R^2 using the F-statistic. In this analysis, the *p*-value is well below .05 ($p < .001$). Therefore, Mary can conclude that the R and R^2 between volunteer hours and the volunteer's income level is statistically significant (different than zero).

Figure 13.9 Bivariate Linear Regression ANOVA Output From SPSS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	804.215	1	804.215	99.303	.000[a]
	Residual	469.718	58	8.099		
	Total	1273.933	59			

a. Predictors: (Constant), income
b. Dependent Variable: hours_week

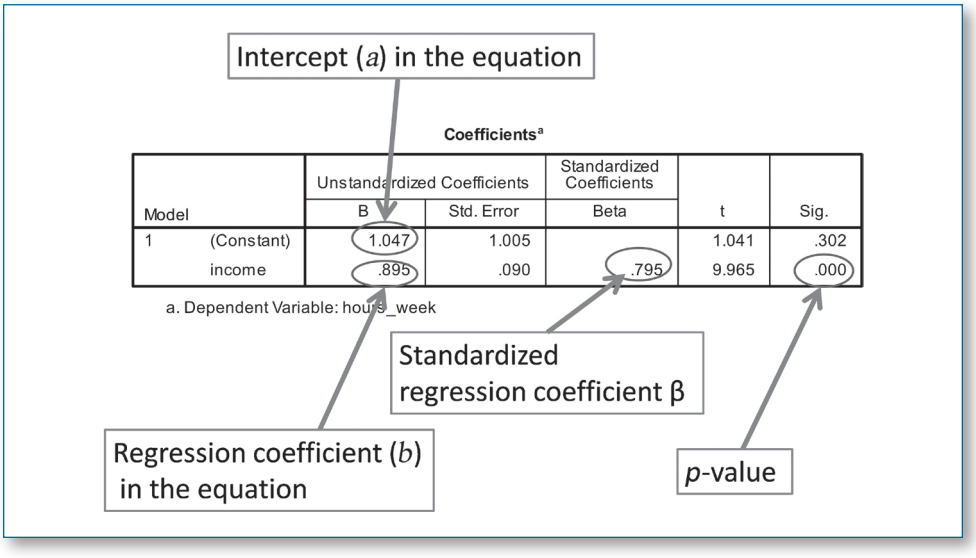
The table in the SPSS output labeled *Coefficients* (Figure 13.10) provides information that is useful for understanding the regression equation. Under the column marked *Unstandardized Coefficient* and sub-column *B*, the numerical value on the first row, labeled *(Constant)*, is the value for the intercept (*a*) in the regression equation. The numerical value on the second row, labeled as *Income* in this case (representing the independent variable), is the value for the slope (*b*) for the regression equation. Based on these results, Mary can report the following regression equation, predicting volunteer hours based on level of income.

$$Y \text{ (Volunteer hours)} = 1.05 + .895X \text{ (income level)}$$

Taking these values for the slope and intercept in the resulting regression equation, we can make the following statement: According to the intercept, when income is zero, the average number of hours will be 1.05, and according to the slope, for each additional unit change in the income level (by defined income categories), the volunteer hours (per week) will increase by .895 hours. Notice in the table that the *p*-value is repeated here ($p < .001$).

Under the column *Standardized Coefficient* and the sub-column *Beta*, the value shown in the second row indicates the slope (*b*) when the independent and dependent variables are converted into scores that have a mean of zero and a standard deviation of 1 (scores with these properties are called **z-scores**). This standardized regression coefficient β (Beta) is useful when making comparisons of the relationship between the variables when the units of measurement are different. We will discuss this concept further below in the section on multiple regression.

Figure 13.10 Bivariate Regression Coefficients SPSS Output



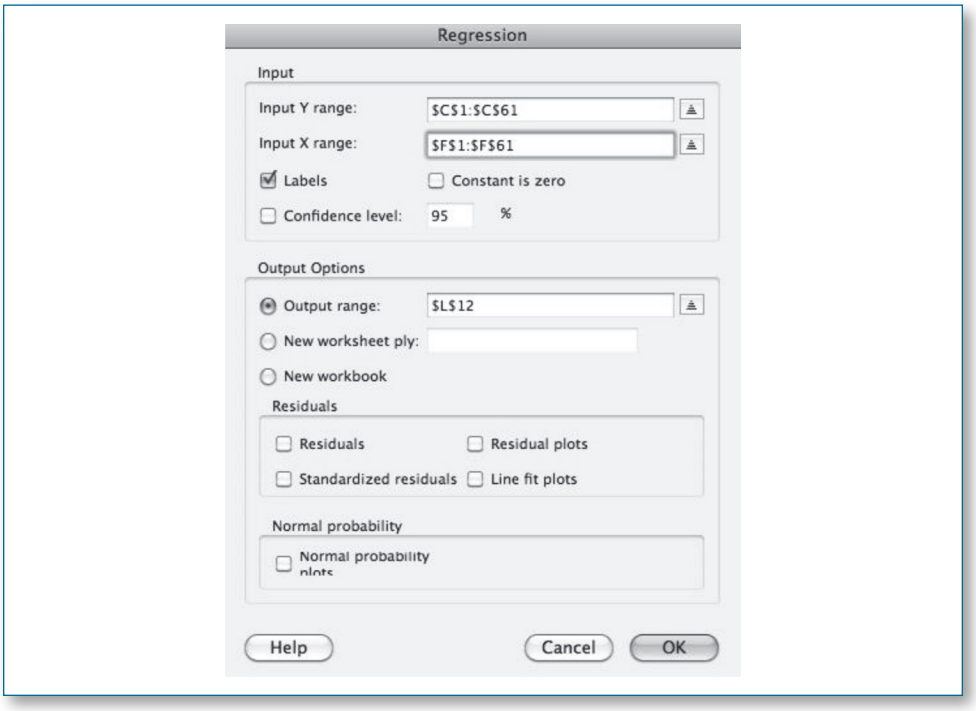
Running Bivariate Regression Using Excel

The bivariate regression analysis can be conducted using Excel with the following steps:

1. Open the Data Analysis window and choose *regression*.
2. Click in the *Input Y Range* to activate.
3. Highlight cells C1 through C61.
4. Once you are finished highlighting these cells, C1:C61 will appear in the *Input Y Range* box.
5. Click on *Input X Range* to activities.
6. Highlight cells F1 through F61.
7. Again, after highlighting, they should appear in the box.
8. Be sure and click *Labels*.
9. Specify your output range and click OK.

Your window should look similar to the Figure 13.11 below.

Figure 13.11 Input Variables for Bivariate Regression in Excel



The output from Excel appears below in Figure 13.12. Note that the *p*-value in this case is reported in scientific notation as a very small value, which we can interpret as in the SPSS output as $p < .001$.

Figure 13.12 Bivariate Regression Output From Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79453464
R Square	0.63128529
Adjusted R Square	0.62492814
Standard Error	2.84580138
Observations	60

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	804.215374	804.2154	99.30319	3.5329E-14
Residual	58	469.7179594	8.098586		
Total	59	1273.933333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.04653671	1.004848223	1.041487	0.301971	-0.96488553	3.057959	-0.96488553	3.0579589
income	0.89473248	0.089786616	9.965099	3.53E-14	0.715005039	1.07446	0.71500504	1.0744599

Multiple Regression

Multiple regression is an extension of bivariate regression. Rather than having only one independent variable in the regression equation, multiple regression includes more than one independent variable in the equation. By incorporating more than one independent variable in the analysis, multiple regression predicts the dependent variable taking multiple factors into account. It also examines the effect of each independent variable on the dependent variable while holding the effect of other variables constant. In other words, multiple regression identifies the unique contribution of the individual independent variables, while controlling for the effects of other independent variables.

The regression equation remains essentially the same for multiple regression, appearing as follows (the subscripts identify additional variables):

Formula 13.7 Equation for Multiple Regression

$$Y = a + b_1X_1 + b_2X_2 + \ldots. + b_iX_i$$

In conducting multiple regression analysis, it is important to think carefully about what independent variables should be included in the analysis (Allison, 1999). An effort should be made to include all relevant independent variables in explaining the dependent variable, and there should be a good theoretical basis for the inclusion of each variable. Additional independent variables should explain differences in the dependent variable that the other independent variables do not. All the independent variables included in the analysis, in combination, should predict the dependent variable better than any one of the independent variables alone.

Multicollinearity

In addition to the assumptions for the linear regression analysis noted earlier, in multiple regression analysis, there is one more important assumption that needs to be met. In multiple regression, independent variables included in the analysis should not have a strong linear relationship to each other. When there is a strong relationship among the independent variables it is referred to as **multicollinearity**. When there is multicollinearity, the two independent variables already share much of the information about the dependent variable and the analysis will not be able to distinguish the effects of one over the other (Allison, 1999; Norusis, 2009).

One way to examine if there is multicollinearity among the independent variables is to run correlations of all pairs of independent variables. When the correlation is high (rule of thumb is above .8), there is a likelihood that you have multicollinearity. SPSS will conduct a diagnosis for multicollinearity by computing what is called a **variance inflation factor (VIF)**. The general rule of thumb is when any VIF is greater than 10 there is a multicollinearity problem (Stevens, 2009). (Some researchers suggest using 5 to be conservative.) If SPSS indicates there is a multicollinearity problem, examine the direct correlation between each pair of independent variables and take out one from a pair that has a high correlation.

Using Dummy Variables in the Multiple Regression

As previously mentioned, a basic premise of linear regression analysis is that the variables are continuous. Yet, there are research questions that hypothesize categorical variables—such as race, gender, political party affiliation—may affect the variance in the dependent variable. Including a categorical variable in the analysis may make the prediction of the dependent variable more accurate. Linear regression analysis allows the inclusion of categorical independent variables as *dummy variables*.

Dummy variables take a value of 0 or 1. The value 0 indicates the absence of the attributes of the category, and the value 1 indicates the presence of the attribute of the category. For example, gender has two attributes, male and female. As a dummy variable, male could be designated as 0, and female as 1. In the regression equation, then, the coefficient for the dummy variable would indicate how the female attribute (1) has an effect on the dependent variable in contrast, or in reference, to the male attribute (0). The category designated as 0 in the dummy variable is called the **reference group**.

In Mary's case, she is considering a second analysis that examines multiple volunteer characteristics to predict volunteer hours (dependent variable), including income level (independent variable 1), age (independent variable 2), and gender (independent variable 3). Notice that gender is a categorical variable. In this case, gender can be added as a dummy variable to the regression equation as follows:

Formula 13.8 Equation for Multiple Regression With Categorical Gender Variable

$$\text{Y (Volunteer hours)} = a + b_1X_1 \text{ (income level)} + b_2X_2 \text{ (age)} + b_3X_3 \text{ (gender)}$$

In interpreting the regression coefficients in this equation, the value of *a* indicates the intercept, or mean volunteer time for male volunteers (the reference

group), when (hypothetically) the volunteer has no income and is zero years old. In other words, the intercept represents the value of the dependent variable when the values of all the independent variables are zero. We expect a relationship between volunteer time (dependent variable) and income level (independent variable 1) indicated by b_1 and the relationship between volunteer time (dependent variable) and age (independent variable 2) indicated by b_2. We expect these relationships to be the same for both male and female volunteers (independent variable 3). The coefficient b_3 indicates the mean difference in the dependent variable between the group coded as 1 (female) and the reference group (male). When the regression coefficient for the dummy variable *gender* is significant, it means the difference in the mean volunteer time between male and female volunteers is significantly different from zero.

Creating a dummy variable for a categorical variable with more than two attributes is more complicated. For example, if Mary wanted to include a *region* variable to indicate the part of town in which the volunteers live, she would have four categorical attributes (or groupings): North, South, East, and West. Including region in a regression analysis would require three dummy variables as follows:

Dummy Variable 1 (North): North = 1, Other region designation = 0

Dummy Variable 2 (South): South =1, Other region designation = 0

Dummy Variable 3 (East): East =1, Other region designation = 0

Notice that we only need to define three of the four regions. In this case, *West* is designated as the reference group, with a value of 0, for all three of the created dummy variables. With a categorical variable like this with multiple attributes, all the dummy variables need to be entered as a block. The coding in this example is summarized in Table 13.1.

If Mary adds these dummy variables in the regression analysis, the equation will appear as follows:

Table 13.1 Dummy Variable Coding

	Dummy Variable 1 (North)	Dummy Variable 2 (South)	Dummy variable 3 (East)
North	1	0	0
South	0	1	0
East	0	0	1
West	0	0	0

Formula 13.9 Equation for Multiple Regression With Categorical Gender Variable and Dummy Coded Region Variable

$$Y \text{ (Volunteer hours)} = a + b_1X_1 \text{ (income level)} + b_2X_2 \text{ (age)} + b_3X_3 \text{ (gender)} + b_4X_4 \text{ (North)} + b_5X_5 \text{ (South)} + b_6X_6 \text{ (East)}$$

The interpretation of the regression coefficient is similar to the case described above with a dummy variable for gender. When the regression coefficient (b_4, b_5, b_6) for the dummy variable is significant, it means the difference in the mean volunteer time between the region represented by the dummy variable (North, South, East respectively) and the reference group (West) is significantly different from zero.

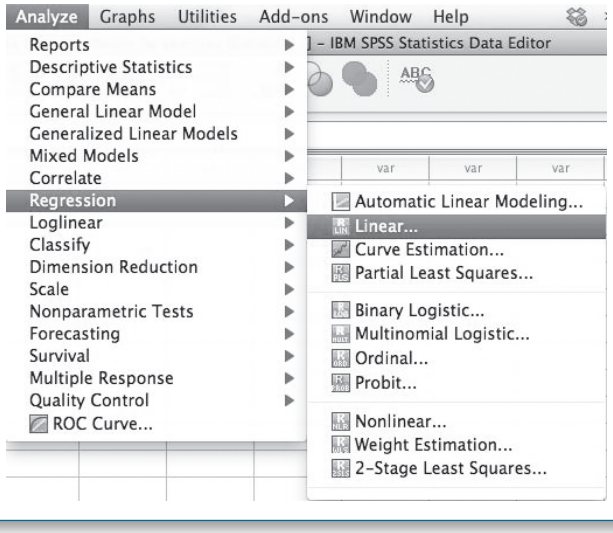
Running Multiple Regression Using Software Programs

Now let's look at Mary's case to see if she can predict volunteer hours better with multiple independent variables in a multiple regression analysis, including volunteer income level, age, and gender. We will go through the procedure in SPSS and Excel.

Running Multiple Regression Using SPSS

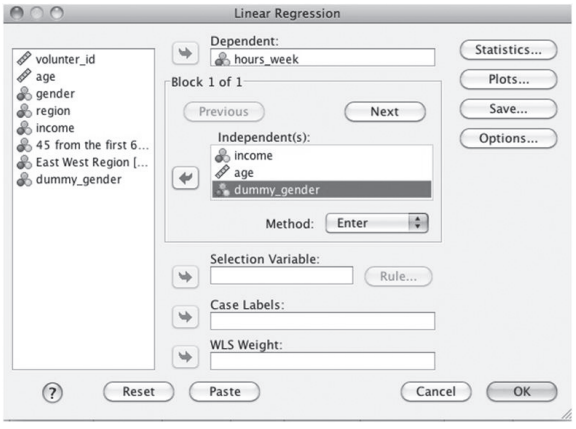
1. Open **Mary_Volunteer_profile.sav**
2. Click **Analyze→Regression→Linear.**

Figure 13.13 Menu Selections for Linear Regression



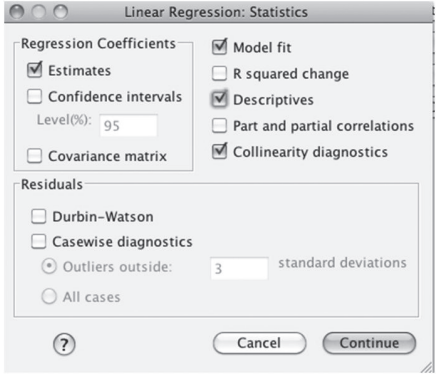
3. Move the variable **hours_week** into the Dependent Variable box.
4. Move the variable **income, age, dummy_gender** into the Independent Variable(s) box.

Figure 13.14 Input Variables for Multiple Regression in SPSS



5. Click **Statistics**. The **Estimates** and **Model Fit** should already be selected as a default.
6. Click **Collinearity diagnostics**. (You can also click **Descriptives** if you want to have the descriptive statistics.)

Figure 13.15 Statistics Options for Linear Regression in SPSS



7. Click **Continue**.
8. Click **OK**.

Just as in the bivariate regression output in SPSS, the table labeled *Model Summary* (Figure 13.16) includes information about R, R square (R^2), and Adjusted R Square. In this

case, with multiple regressions, all three R values indicate the degree to which *the linear combination of the independent variables* in the regression analysis predicts the dependent variable. We will explain the idea of *linear combination* in the discussion below.

In the multiple regression, the value of R is different than in the bivariate regression. Here, it represents the Pearson product moment correlation coefficient between

Figure 13.16 Multiple Regression Model Summary SPSS Output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.799[a]	.639	.619	2.86718

a. Predictors: (Constant), dummy_gender, income, age

the observed value of the dependent variable and the predicted value of the dependent variable using the regression equation. **R** for multiple regression is referred to as **Multiple R** (Field, 2009). The characteristics of the metric are the same, with a range from 0 to 1, a larger value indicating a larger correlation and 1 representing an equation that perfectly predicts the observed value of the dependent variable. Multiple R is an indicator of how well the overall regression equation predicts the observed data. In the current multiple regression analysis for Mary, the result of .799 indicates that the linear combination of the three independent variables (income, age, and gender) strongly predicts the actual dependent variable.

R Square (R^2) indicates the proportion of variance that can be explained in the dependent variable by the linear combination of the independent variables. The values of R^2 also range from 0 to 1. Mary's analysis suggests that the linear combination of volunteers' income, age, and gender explains 63.9% of the variance in volunteer hours. Note that this is a slight increase from the bivariate model, which was 63.1%.

Typically, anytime more variables are added to the regression equation, the value of R^2 increases. As a note of caution, adding variables haphazardly to increase the explanation of the variance in the dependent variable is not a good research practice. As noted earlier, each independent variable should be added with a purpose that comes from the research question and the theory. Sometimes there is a tendency to treat multiple regression analysis like making soup; the cook will add a bunch of leftovers just because they are there and need to be used. This kind of arbitrary, nontheoretical approach can produce misleading results (Baltagi, 2011).

Adjusted R Square (R^2), as noted for the bivariate regression analysis, adjusts the value of R^2 to more accurately represent the population of interest when the sample size is small. Also when there are a large number of independent variables included in

the multiple regression equation, it tends to produce a higher estimation of the **R^2** in the population, and therefore, Adjusted R Square adjusts the value. In Mary's analysis, the adjusted R^2 is 61.0%—more conservative than the unadjusted R^2 of 63.1%. It is different enough from the unadjusted R^2 to be worth reporting.

The table labeled *ANOVA* in the SPSS output (Figure 13.17) provides the results of a test of significance for R and R square using the F-statistic. In this analysis, the *p*-value is well below .05 ($p < .001$), and therefore, Mary can conclude that R, R^2, and Adjusted R^2 for the multiple regression she conducted predicting volunteer hours based on the linear combination of income, age, and gender is statistically significant.

The information in the table labeled *Coefficients* in the SPSS output (Figure 13.18) can be interpreted in the same way as we discussed in the bivariate regression section above. It provides information that is useful for understanding the regression equation.

Figure 13.17 Multiple Regression ANOVA SPSS Output

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	405.285	2	202.642	34.519	.000[a]
	Residual	105.667	18	5.870		
	Total	510.952	20			

a. Predictors: (Constant), age, income
b. Dependent Variable: hours_week

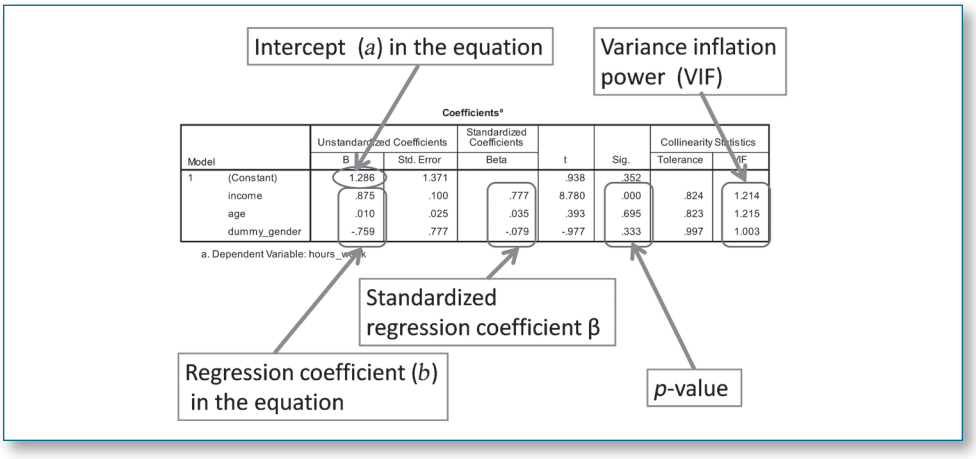
Again, under the column marked *Unstandardized Coefficient* and sub-column *B* is the value for the intercept (*a*) in the regression equation on the first row, labeled (*Constant*). The numbers below it in the same column are the values for the regression coefficients for income, age, and gender. Based on these results, the regression equation that predicts volunteer hours based on the linear combination of income, age, and gender is as follows:

Formula 13.10 Regression Equation That Predicts Volunteer Hours

$$Y \text{ (volunteer hours)} = 1.29 + .88X_1 \text{ (income)} + .01X_2 \text{ (age)} + (-.76)X_3 \text{ (gender\_reference male)}$$

This result indicates, first, that the intercept is 1.29 hours when all independent variables have a value of zero. Then, moving through the equation, holding volunteer age and gender constant, the volunteer hours (per week) increase by .88 hours for each additional increase in the income level. The *p*-value for this coefficient is statistically significant ($p < .001$), meaning that volunteer income is a significant predictor of volunteer hours. Holding income and gender constant, the volunteer hours increase by only .01 hours (per week), according to the equation, and this coefficient is not statistically significant ($p = .695$). Volunteer age is not a significant predictor of the volunteer hours. Finally, the regression coefficient for the gender dummy variable, with male as the reference group, is –.76, which means that holding volunteer income level and age constant,

Figure 13.18 Multiple Regression Coefficients SPSS Output



female volunteers put in an average of .76 hours less (per week) than male volunteers. However, the *p*-value for gender is also not statistically significant ($p = .333$).

As with the bivariate regression analysis, the values in the Coefficients table under the column Standardized Coefficient and sub-column *Beta* is the regression coefficient when the independent and dependent variables are converted to a z-score. In the multiple regression, this standardized regression coefficient Beta (β) is useful, because it allows you to compare the relative strength of each independent variable's relationship with the dependent variable. In this case, the regression coefficients (*b*) provide you with information on how much change can be expected with a one-unit change in each independent variable, but they don't tell you the relative strength of the relationship between the dependent variable and each of the independent variables. With the Beta values here, we can see in Mary's analysis that income (.777) has the strongest relationship with volunteer hours, compared to age (.035) and gender (–.079). Besides, the Beta for age and gender are not statistically significant.

In the same table, the information under the column *Collinearity Statistics* and sub-column *VIF* indicates if there is multicollinearity among the independent variables. In this current analysis, all VIF is lower than 5, and therefore, Mary can be assured that there is no multicollinearity problem in her analysis. If any of the VIF is higher than 5, Mary needs to check the correlation of that particular variable with other independent variables and eliminate one of the variables with high correlation.

Running Multiple Regression Using Excel

Running multiple regression analysis in Excel follows the same procedure that we used for bivariate regression. The only key difference is that the multiple independent variables need to be placed in columns that are contiguous to each other. Therefore, move the independent variables to be used in the analysis in columns next to each other or copy a column so that it is contiguous to your other independent variable(s). When the

Input X range box is activated, then highlight all of the columns of the independent variables (again, they must be contiguous).

Mary's Case

Mary ran her finger over the output she obtained from her multiple regression analysis to predict volunteer hours by income level, age, and gender. She talked herself through it.

"OK, R square is about .64, so this regression equation accounts for about 64% of the variances, and it is significant. That's not too bad. So it looks like this is a good regression equation model to predict volunteer hours."

She then directed her attention to the regression coefficients.

"Hmm ... so age and gender are not significant. That means age may not be a good predictor for volunteer hours. And there may not be a generalizable difference between male and female volunteers. Still, the coefficient for the gender dummy variable is a fairly strong negative value, which suggests that female volunteers put in less volunteer time when income level and age are constant. That's interesting."

Mary decided to check the descriptive statistics, comparing actual volunteer time for male and female volunteers, and indeed, female volunteers had a lower average. It may not be generalizable, she thought, but it was interesting, because it went against the assumption she had heard that women put in more time.

The regression coefficient for income level was more startling.

"Wow, it's .875 and significant. Income level clearly matters more than anything else. Does that mean I should try to target higher-income volunteers?"

Somehow this conclusion did not sit well with her. Although the data definitely suggested this relationship of income and volunteer hours, she wondered if there might be other factors that were not captured in the volunteer background information—something that coincided with higher income.

"I still think I need to interview volunteers and get their perspectives."

Mary shut down SPSS and opened the list of volunteers she had marked for interviews.

Brief Comment on Other Types of Regression Analyses

In this chapter, we introduced two types of regression analysis that can be used when the dependent variable is continuous. There is another type of regression analysis called **logistic regression**, which can be used to predict the outcome when the dependent variable is dichotomous. To learn more about logistic regression, see Kleinbaum and Klein (2011), and Menard (2008).

Another variation of regression analyses that is commonly used by public and nonprofit mangers or policymakers is **time series analysis**, which is useful when observing trends and making forecasts based on past observations at equally spaced time intervals. To learn more about time series analysis, see Brockwell and Davis (2002), and Ostrom (1990).

When you evaluate a policy or program, you have multiple observation points before and after an intervention, resulting in a time series that looks like the following

notation, where O indicates observations and X indicates the implementation of the policy or program:

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad X \quad O_5 \quad O_6 \quad O_7 \quad O_8$$

With this design, you can use an **interrupted time series** analysis. To learn more about interrupted time series analysis, see McDowall (1980).

Chapter Summary

This chapter introduced bivariate and multiple linear regression analyses. Linear regression analysis identifies a regression equation that allows a researcher to predict the scores of the dependent variable based on the scores of one or more independent variables. It also provides information on the strength of the relationship between the dependent variable and the independent variables.

Review and Discussion Questions and Exercises

1. Based on the Emily's case description at the beginning of the chapter, run a bivariate regression analysis to answer Leo's question. Write a regression equation based on the result you obtained.
2. Are there any other independent variables that are appropriate to include in the analysis you conducted in (1) above? Conduct a multiple regression analysis and report the result.
3. Create dummy variables for *region* in Mary's data and conduct multiple regression analysis with the dummy variables. (See Appendix A for the instructions on how to recode the variable in SPSS to create dummy variables.)
4. Describe the importance of the multicollinearity assumption in linear regression.
5. Describe Total Sum of Squares and Residual Sum of Squares and how it relates to Coefficient of Determination.
6. Describe the difference between standardized regression coefficient β and the unstandardized regression coefficient *b*.
7. When is it appropriate to report adjusted R^2?

References

Allison, P. D. (1999). *Multiple regression: A primer.* Thousand Oaks, CA: Pine Forge Press.
Baltagi, B. H. (2011). *Econometrics* (5th ed.). New York, NY: Springer.
Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting*. New York, NY: Springer.
Cohen, J. (2010). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York, NY: Routledge.
Field, A. P. (2009). *Discovering statistics using SPSS: (And sex and drugs and rock 'n' roll)*. Thousand Oaks, CA: Sage.
Fox, J. (1991). *Regression diagnostics*. Newbury Park, CA: Sage.

Green, S. B., & Salkind, N. J. (2010). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
Homoscedasticity. (n.d.). In *Merriam-Webster's online dictionary* (11th ed.). Retrieved from http://www.m-w.com/dictionary/homoscedasticity
Kahane, L. H. (2008). *Regression basics* (2nd ed.). Thousand Oaks, CA: Sage.
Kleinbaum, D. G., & Klein, M. (2011). *Logistic regression: A self-learning text* (3rd ed.). New York, NY: Springer.
McDowall, D. (1980). *Interrupted time series analysis* (Vol. 21). Beverly Hills, CA: Sage.
Menard, S. (2008). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.
Munro, B. H. (2004). *Statistical methods for health care research* (5th ed.). Philadelphia, PA: Lippincott, Williams, & Wilkins.
Norusis, M. J. (2009). *PASW statistics 18 statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.
Ostrom, C. W. (1990). *Time series analysis: Regression techniques*. Newbury Park, CA: Sage.
Pedhazur, E.J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.
Stevens, J. P. (2009). *Applied multivariate statistcs for the social sciences* (5th ed.). New York, NY: Routledge.
Upton, G. J. G., & Cook, I. (2011). *A dictionary of statistics* (2nd ed.). New York, NY: Oxford University Press.

Key Terms

Student Study Site

Visit the Student Study Site at **www.sagepub.com/nishishiba1e** for these additional learning tools:

- Data sets to accompany the exercises in the chapter
- Result write-ups