ONE

Review of Linear Regression Models

As you should know,[1] the linear regression model is normally characterized with the following equation:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i \quad \{\text{or use } \beta_0 \text{ for } \alpha\}.$$

Consider this equation and try to answer the following questions:

- What does the y_i represent? The β? The x? (Which often include subscripts i—do you remember why?) The ε_i?
- How do we judge the size and direction of the β?
- How do we decide which xs are important and which are not? What are some limitations in trying to make this decision?
- Given this equation, what is the difference between prediction and explanation?
- What is this model best suited for?
- What role does the mean of y play in linear regression models?
- Can the model provide causal explanations of social phenomena?
- What are some of its limitations for studying social phenomena and causal processes?

[1] This book assumes familiarity with linear regression models that are estimated with ordinary least squares (OLS). There are many books that provide excellent overviews of the model and its assumptions (e.g., Fox, 2016; Hoffmann and Shafer, 2015; Weisberg, 2013).

Researchers often use an estimation technique known as *ordinary least squares (OLS)* to estimate this regression model. OLS seeks to minimize the following:

$$\mathrm{SSE} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 .$$

The SSE is the *sum of squared errors*, with the observed *y* and the predicted *y* (*y-hat*) utilized in the equation. In an OLS regression model[2] that includes only one explanatory variable, the slope (β_1) is estimated with the following least squares equation:

$$\hat{\beta}_1 = \frac{\sum \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)}{\sum \left(x_i - \bar{x} \right)^2 / \left(n - 1 \right)} .$$

Notice that the variance of *x* appears in the denominator, whereas the numerator is part of the formula for the covariance (cov(*x*,*y*)). Given the slope, the intercept is simply

$$\hat{\alpha} = \bar{y} - \left\{ \hat{\beta}_1 \times \bar{x} \right\}.$$

Estimation is more complicated in a multiple OLS regression model. If you recall matrix notation, you may have seen this model represented as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} .$$

The letters are bolded to represent vectors and matrices, with **Y** representing a vector of values for the outcome variable, **X** indicating a matrix of explanatory variables, and **β** representing a vector of regression coefficients, including the intercept (β_0) and slopes (β_i). The OLS regression coefficients may be estimated with the following equation:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Y} .$$

A vector of residuals is then given by

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{Y}\hat{\boldsymbol{\beta}} .$$

Often, the residuals are represented as *e* to distinguish them from the errors, *ε*. You should recall that residuals play an important role in linear regression analysis. Various types of residuals also have a key role throughout this book. Assuming a sample and that the

[2] The term "OLS regression model" is simply a shorthand way of indicating that the linear regression model is estimated with OLS. As shown in subsequent chapters, another common estimation technique is maximum likelihood estimation (MLE). Thus, an ML regression model refers to a model that is estimated using MLE.

model includes an intercept, some of the properties of the OLS residuals are (a) they sum to zero ($\Sigma\varepsilon_i = 0$), (b) they have a mean of zero ($E[\varepsilon] = 0$), and (c) they are uncorrelated with the predicted values of the outcome variable ($r(\varepsilon, \hat{y}) = 0$).

Analysts often wish to infer something about a target population from the sample. Thus, you may recall that the standard error (SE) of the slope is needed since, in conjunction with the slope, it allows estimation of the *t*-values and the *p*-values. These provide the basis for inference in linear regression modeling. The standard error of the slope in a simple OLS regression model is computed as

$$\text{SE}\left(\hat{\beta}_1\right) = \sqrt{\frac{\sum\left(y_i - \hat{y}_i\right)^2 / n-2}{\sum\left(x_i - \bar{x}\right)^2}} = \sqrt{\frac{\text{SSE}/n-2}{\text{SS}[x]}}.$$

Assuming we have a multiple OLS regression model, as shown earlier, the standard error formula requires modification:

$$\text{SE}\left(\hat{\beta}_i\right) = \sqrt{\frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(x_i - \bar{x}\right)^2\left(1 - R_i^2\right)\left(n - k - 1\right)}}.$$

Consider some of the components in this equation and how they might affect the standard errors. The matrix formulation of the standard errors is based on deriving the variance-covariance matrix of the OLS estimator. A simplified version of its computation is

$$\sigma_\varepsilon^2\left(\mathbf{X'X}\right)^{-1}, \text{with } \sigma_\varepsilon^2 \text{ estimated by } \hat{\sigma}_\varepsilon^2 = \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n-k} = \frac{\sum\varepsilon_i^2}{n-k}.$$

Note that the numerator in the right-hand-side equation is simply the SSE since $\left(y_i - \bar{y}\right) = \varepsilon_i$ or e_i. The right-hand-side equation is called the *residual variance* or the *mean squared error* (MSE). You may recognize that it provides an estimate—albeit biased, but consistent—of the variance of the errors. The square roots of the diagonal elements of the variance–covariance matrix yield the standard errors of the regression coefficients. As reviewed subsequently, several of the assumptions of the OLS regression model are related to the accuracy of the standard errors and thus the inferences that can be made to the target population.

OLS results in the smallest value of the SSE, if some of the specific assumptions of the model discussed later are satisfied. If this is the case, the model is said to result in the best linear unbiased estimators (BLUE) (Weisberg, 2013). It is important to note that this says *best linear,* so we are concerned here with linear estimators (there are also nonlinear estimators). In any event, BLUE implies that the estimators, such as the slopes, from an OLS regression model are unbiased, efficient, and consistent. But what does it mean to say they have these qualities? Unbiasedness refers to whether the mean of the sampling distribution of a statistic equals the parameter it is meant to estimate in the population. For example, is the slope estimated from the sample a good estimate of an analogous slope in the population? Even though

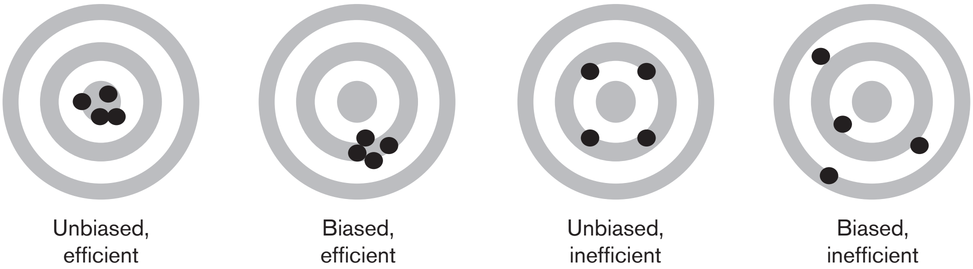

FIGURE 1.1

we rarely have more than one sample, simulation studies indicate that the mean of the sample slopes from the OLS regression model (if we could take many samples from a population), on average, equals the population slope (see Appendix B). Efficiency refers to how stable a statistic is from one sample to the next. A more efficient statistic has less variability from sample to sample; it is therefore, on average, more precise. Again, if some of the assumptions discussed later are satisfied, OLS-derived estimates are more efficient—they have a smaller sampling variance—than those that might be estimated using other techniques. Finally, consistency refers to whether the statistic converges to the population parameter as the sample size increases. Thus, it combines characteristics of both unbiasedness and efficiency.

A standard way to consider these qualities is with a target from, say, a dartboard. As shown in figure 1.1, estimators from a statistical model can be imagined as trying to hit a target in the population known as a parameter. Estimators can be unbiased and efficient, biased but efficient, unbiased but inefficient, or neither. Hopefully, it is clear why having these properties with OLS regression models is valuable.

You may recall that we wish to assess not just the slopes and standard errors, but also whether the OLS regression model provides a good "fit" to the data. This is one way of asking whether the model does a good job of predicting the outcome variable. Given your knowledge of OLS regression, what are some ways we may judge whether the model is a "good fit"? Recall that we typically examine and evaluate the R^2, adjusted R^2, and root mean squared error (RMSE). How is the R^2 value computed? Why do some analysts prefer the adjusted R^2? What is the RMSE and why is it useful?

A BRIEF INTRODUCTION TO STATA[3]

In this presentation, we use the statistical program Stata to estimate regression models (www.stata.com). Stata is a powerful and user-friendly program that has become quite popular in the social and behavioral sciences. It is more flexible and powerful than SPSS and, in

[3] There are many excellent introductions and tutorials on how to use Stata. A good place to start is Stata's YouTube channel (https://www.youtube.com/user/statacorp) and the following website: http://www.ats.ucla.edu/stat/stata/. This website also includes links to web books that demonstrate how to estimate OLS regression models with Stata. For a more thorough treatment of

my judgment, much more user-friendly than SAS or R, its major competitors. Stata's default style consists of four windows: a command window where we type the commands; a results window that shows output; a variables window that shows the variables in the data file; and a review window that keeps track of what we have entered in the command window. If we click on a line in the review window, it shows up in the command window (so we don't have to retype commands). If we click on a variable in the variables window, it shows up in the command window, so we do not have to type variable names if we do not want to.

It is always a good idea to save the Stata commands and output by opening a log file. This can be done by clicking the brown icon in the upper left-hand corner (Windows) or the upper middle portion (Mac) of Stata or by typing the following in the command window:

```
log using "regression.log"                    * the name is arbitrary
```

This saves a log file to the local drive listed at the bottom of the Stata screen. To suspend the log file, type `log off` in the command window; or to close it completely type `log close`.

It is also a good idea to learn to use *.do* files. These are similar to SPSS syntax files or R script files in that we write—and, importantly, save—commands in them and then ask Stata to execute the commands. Stata has a do-editor that is simply a notepad screen for typing commands. The Stata icon that looks like a small pad of paper opens the editor. But we can also use Notepad++, TextEdit, WordPad, Vim, or any other text-editing program that allows us to save text files. I recommend that you use the handle .do when saving these files, though. In the do-editor, clicking the *run* or *do* icon feeds the commands to Stata.

AN OLS REGRESSION MODEL IN STATA

We will now open a Stata data file and estimate an OLS regression model. This allows us to examine Stata's commands and output and provide guidance on how to test the assumptions of the model. A good source of additional instructions is the *Regression with Stata* web book found at http://www.ats.ucla.edu/stat/stata/webbooks/reg. Stata's help menu (e.g., type `help regress` in the command window) is also very useful.

To begin, open the *GSS* data file (*gss.dta*). This is a subset of data from the biennial *General Social Survey* (see www3.norc.org/GSS+Website). You may use Stata's drop-down menu to open the file. Review the content of the Variables window to become familiar with the file and its contents. A convenient command for determining the coding of variables in Stata is called `codebook`. For example, typing and entering `codebook sei` returns the label and some information about this variable, including its mean, standard deviation, and some percentiles. Other frequently used commands for examining data sets and variables include

using Stata for this purpose, see Hoffmann and Shafer (2015). For more information about using Stata to conduct quantitative research, see Long (2009) and Kohler and Kreuter (2012).

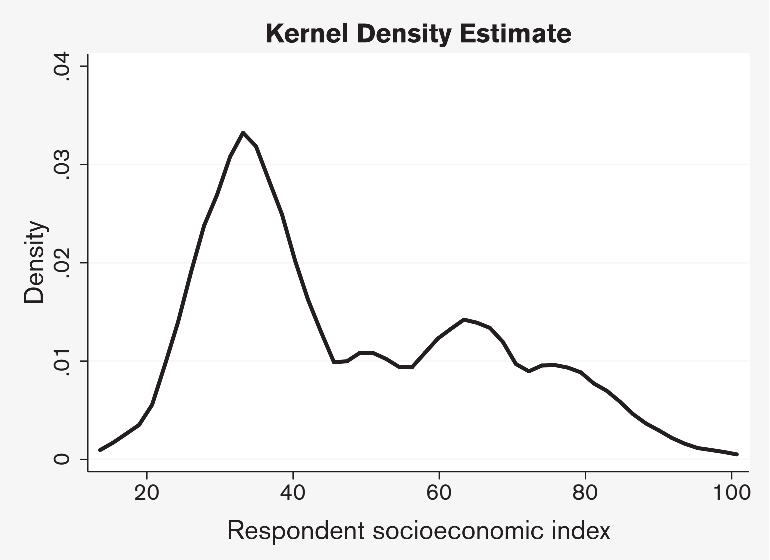

FIGURE 1.2

describe, table, tabulate, summarize, graph box (boxplot), graph dotplot (dot plot), stem (stem-and-leaf plot), hist (histogram), and kdensity (kernel density plot) (see the Chapter Resources at the end of this chapter). Stata's help menu provides detailed descriptions of each. As shown later, several of these come in handy when we wish to examine residuals and predicted values from regression models.

Before estimating an OLS regression model, let's check the distribution of *sei* with a *kernel density graph* (which is also called a *smoothed histogram*). The Stata command that appears below opens a new window that provides the graph in figure 1.2. If *sei* follows a normal distribution, it should look like a bell-shaped curve. Although it appears to be normally distributed until it hits about 50, it has a rather long tail that is suggestive of positive skewness. We investigate some implications of this skewness later.

```
kdensity sei
```

To estimate an OLS regression model in Stata, we may use the regress command.[4] The Stata code in Example 1.1 estimates an OLS regression model that predicts *sei* based on sex (the variable is labeled *female*). The term beta that follows the comma requests that Stata

[4] If you wish to see detailed information about how this command operates, including its many subcommands and technical details about how it estimates the model, select the **[R] regress** link from the regress help page. Assuming that Stata is found locally on your computer, this brings up the section on the regress command in the Stata reference manual.

furnish *standardized regression coefficients*, or *beta weights*, as part of the output. You may recall that beta weights are based on the following equation:

$$\hat{\beta}_k^S = \hat{\beta}_k \times \frac{\sigma_{x_k}}{\sigma_y}.$$

Whereas unstandardized regression coefficients (the `Coef.` column in Stata) are interpreted in the original units of the explanatory and outcome variables, beta weights are interpreted in terms of *z*-scores. Of course, the *z*-scores of the variables must be interpretable, which is not always the case (think of a categorical variable like *female*).

Example 1.1

```
regress sei female, beta
```

```
      Source |       SS           df       MS      Number of obs   =     2,780
-------------+----------------------------------   F(1, 2778)      =      5.56
       Model |  2002.36916         1  2002.36916   Prob > F        =    0.0184
    Residual |  1000216.71     2,778   360.04921   R-squared       =    0.0020
-------------+----------------------------------   Adj R-squared   =    0.0016
       Total |  1002219.08     2,779  360.640186   Root MSE        =    18.975

------------------------------------------------------------------------------
         sei |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      female |  -1.704067     .722596    -2.36   0.018                -.0446983
       _cons |   48.79006    .5330803    91.52   0.000                        .
------------------------------------------------------------------------------
```

The results should look familiar. There is an analysis of variance (ANOVA) table in the top-left panel, some model fit statistics in the top-right panel, and a coefficients table in the bottom panel. For instance, the R^2 for this model is 0.002, which could be computed from the ANOVA table using the regression (Model) sum of squares and the total sum of squares (SS(*sei*)): 2,002/1,002,219 = 0.002. Recall that the R^2 is the squared value of the correlation between the predicted values and the observed values of the outcome variable. The *F*-value is computed as *MSReg*/*MSResid* or 2,002/360 = 5.56, with degrees of freedom equal to *k* and $\{n - k - 1\}$. The adjusted R^2 and the RMSE[5]—two useful fit statistics—are also provided.

The output presents coefficients (including one for the *intercept* or *constant*), standard errors, *t*-values, *p*-values, and, as we requested, beta weights. Recall that the unstandardized regression coefficient for a binary variable like *female* is simply the difference in the expected means of the outcome variable for the two groups. Moreover, the intercept is the predicted mean for the reference group if the binary variable is coded as {0, 1}. Because *female* is coded

[5] Recall that the RMSE is simply $\sqrt{\text{MSE}}$ or the estimate σ_e. It may also be characterized as the standard deviation of the residuals.

as {0 = male and 1 = female}, the model predicts that mean *sei* among males is 48.79 and mean *sei* among females is 48.79 – 1.70 = 47.09. The *p*-value of 0.018 indicates that, assuming we were to draw many samples from the target population, we would expect to find a slope of –1.70 or one farther from zero about 18 times out of every 1,000 samples.[6]

The beta weight is not useful in this situation because a one *z*-score shift in *female* makes little sense. Perhaps it will become more useful as we include other explanatory variables. In the next example, add years of education, race/ethnicity (labeled *nonwhite*, with 0 = white and 1 = nonwhite), and parents' socioeconomic status (*pasei*) to the model.

The results shown in Example 1.2 suggest that one or more of the variables added to the model may explain the association between *female* and socioeconomic status (or does it?—note the sample sizes of the two models). And we now see that education, nonwhite, and parents' status are statistically significant predictors of socioeconomic status. Whether they are important predictors or have a causal impact is another matter, however.

Example 1.2

```
regress sei female educate nonwhite pasei, beta
```

```
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(4, 2226)      =    305.49
       Model |  293540.757         4  73385.1893   Prob > F        =    0.0000
    Residual |  534724.324     2,226  240.217576   R-squared       =    0.3544
-------------+----------------------------------   Adj R-squared   =    0.3532
       Total |  828265.081     2,230  371.419319   Root MSE        =    15.499

------------------------------------------------------------------------------
         sei |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      female |  -.2358993   .6584269    -0.36   0.720                -.0061131
     educate |    3.72279   .1232077    30.22   0.000                 .5563569
    nonwhite |  -2.860583   .9294002    -3.08   0.002                -.0528386
       pasei |   .0765712   .0191531     4.00   0.000                 .0740424
       _cons |  -4.804311   1.677362    -2.86   0.004                        .
------------------------------------------------------------------------------
```

The R^2 increased from 0.002 to 0.353, which appears to be quite a jump. Stata's `test` command provides a multiple partial (nested) *F*-test to determine if the addition of these variables leads to a statistically significant increase in the R^2. Simply type `test` and then list the additional explanatory variables that have been added to produce the second model. The result of this test with the three additional variables is an *F*-value of 406.4 (3, 2226 *df*) and

[6] Consider this statement carefully: we assume what would happen if we were to draw many samples and compute a slope for each. As suggested earlier, this rarely happens in real-world applications. Given this situation, as well as other limitations discussed by statisticians, the *p*-value and its inferential framework are often criticized (e.g., Gelman, 2015; Hubbard and Lindsay, 2008). Although this is an important issue, we do not explore it here.

a *p*-value of less than 0.0001. Given the different sample sizes, do you recommend using the nested *F*-test approach for comparing the models? How would you estimate the effect of *female* in this model?[7]

Some interpretations from the model in Example 1.2 include the following:

- Adjusting for the effects of sex, race/ethnicity, and parents' *sei*, each 1-year increase in education is associated with a 3.72 unit increase in socioeconomic status.
- Adjusting for the effects of sex, race/ethnicity, and parents' *sei*, each one *z*-score increase in education is associated with a 0.556 *z*-score increase in socioeconomic status.
- Adjusting for the effects of sex, education, and race/ethnicity, each one-unit increase in parents' *sei* score is associated with a 0.077 unit increase in socioeconomic status.

It is useful to graph the results of regression models in some way. This provides a more informed view of the association between explanatory variables and the outcome variable than simply interpreting slope coefficients and considering *p*-values to judge effect sizes. For instance, figure 1.3 provides a visual depiction of the linear association between years of education and *sei* as predicted by the regression model. Stata's `margins` and `marginsplot` post-estimation commands are used to "adjust" the other variables by setting them at particular levels, including placing *pasei* at its mean. The vertical bars are 95% confidence intervals (CIs). What does the graph suggest?

```
qui margins, at(educate=(10 12 14 16 18 20)
  female=0 nonwhite=0) atmeans              * qui suppresses the output
marginsplot
```

Although it should be obvious, note that the graph (by design) shows a linear association. This is because the OLS regression model assumes a linear, or straight line, relationship between education and socioeconomic status (although this assumption can be relaxed). If we know little about their association, then relying on a linear relationship seems

[7] Consider the different sample sizes. Why did the sample size decrease? What variables affected this decrease? How can you ensure that the same sample is used in both analyses? An essential task in regression and other statistical modeling efforts is to always look closely at the sample size and whether it changes from one model to the next. As a hint about how to approach this issue, look up Stata's `e(sample)` option. Use it to estimate the effect of *female* in the reduced sample. Did we reach the wrong initial conclusion regarding the association between *female* and *sei*? Why or why not? Although we do not cover missing data and their implications in this chapter (but see Appendix C), it is an essential topic for statistical analysts.

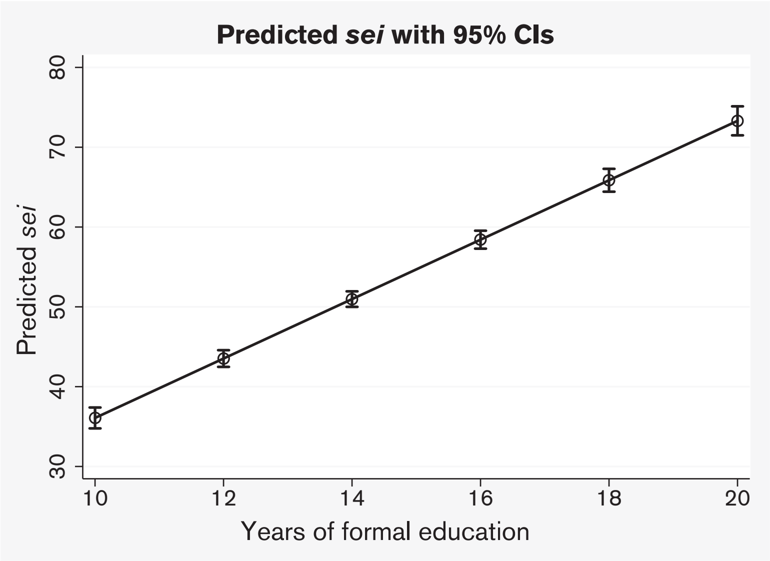

FIGURE 1.3

reasonable. But it is important to keep in mind that many associations are not linear. Think about what this means given how popular linear regression is in many scientific disciplines.

CHECKING THE ASSUMPTIONS OF THE OLS REGRESSION MODEL

Stata provides many convenient tools for checking the assumptions of OLS regression models. Recall that some of these assumptions are important because, if they are satisfied, we can be confident that the OLS regression coefficients are BLUE. We now briefly examine the assumptions and learn about some ways to examine them using Stata. For a comprehensive review, see Fox (2016).

1. Independence

To be precise, this assumption is that *the errors from one observation are independent of the errors in other observations* (cov(ε_i, ε_j) = 0). However, this typically is not the case when the sampling strategy was not randomly driven and there is nesting or clustering among the observations. It is quite common in the social and behavioral sciences given how often surveys are conducted that use clustered or multistage sampling designs. If not taken into account, clustering tends to lead to underestimated OLS standard errors (see Chapter 9). Stata has a host of routines for dealing with complex sampling designs and clustering. In fact, there is a `cluster` subcommand (this means it appears after the main command, usually following a comma)

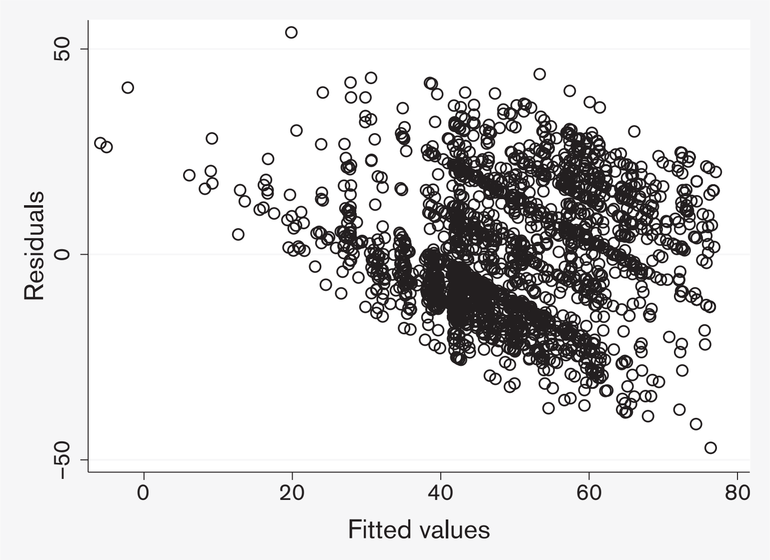

FIGURE 1.4

that may be used to adjust the standard errors for clustering: `regress y x`$_1$ `x`$_2$, cluster(*cluster variable*). We do not pursue this topic any further here, but see Chapters 7 and 8.

2. Homoscedasticity ("Same Scatter")

This involves the assumption that the *variance of the error terms is constant for all combinations of the x*. If this is violated—in other words, if there are *heteroscedastic errors*—then the OLS standard errors are inefficient. Recall that the standard error equation shown earlier included in its denominator the sums of squares for the *x* variable. This value increases, perhaps substantially, in the presence of heteroscedasticity, thus leading to smaller standard errors, on average. How might it affect one's conclusions about the model's coefficients?

There are several options available to test for heteroscedasticity, some of which examine the residuals from the model. For example, a simple way to build a residual-by-predicted plot is with Stata's `rvfplot` (*residuals vs. fitted plot*) post-estimation command (type `help regress postestimation` for a thorough description of other options). Example 1.3 shows the Stata code to execute this command and figure 1.4 furnishes the graph. Recall that we look for a funnel shape in this graph as evidence of heteroscedasticity (other patterns might also provide evidence). Although no funnel exists, there is a peculiar pattern. Some analysts prefer to graph the *studentized residuals* versus the standardized predicted values. The second part of Example 1.3 provides a set of commands for computing and graphing these predictions. Notice that we use a Stata command called `egen` to create a variable that consists of the standardized values (*z*-scores) of the predicted values.

Example 1.3

```
rvfplot                                        * after Example 1.2
predict rstudent, rstudent               * studentized residuals
predict pred, xb                               * predicted values
egen zpred = std(pred)              * z-scores of predicted values
twoway scatter rstudent zpred ||
  lowess rstudent zpred
```

The last command in Example 1.3 creates a scatter plot and overlays a *lowess* (locally weighted regression) fit line. The graph (not shown) identifies the same peculiar pattern. Although it might not be clearly indicative of heteroscedasticity, this is an unusual situation that should be examined further.

Stata also offers several numeric tests for assessing heteroscedasticity. These include `hettest` (the Breusch–Pagan test [Breusch and Pagan, 1979]) and White's (1980) test, which is based on the information matrix (`imtest`).

Example 1.4

```
hettest                                        * after Example 1.2

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of sei

         chi2(1)      =    22.77
         Prob > chi2  =   0.0000


estat imtest, white

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      60.79     12    0.0000
            Skewness |     172.28      4    0.0000
            Kurtosis |      39.23      1    0.0000
---------------------+-----------------------------
               Total |     272.30     17    0.0000
---------------------------------------------------
```

Both tests shown in Example 1.4 indicate that there is heteroscedasticity in the model (the null hypotheses are that the errors are homoscedastic). These tests are quite sensitive to violations of the normality of the residuals, so we should reexamine this issue before deciding what to do.

We may also estimate Glejser's (1969) test by saving the residuals, taking their absolute values, and reestimating the model but predicting these values rather than *sei*. Example 1.5 furnishes this test.

Example 1.5

** after Example 1.3*

```
gen absresid = abs(rstudent)
regress absresid female educate nonwhite pasei
```

```
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(4, 2226)      =     12.16
       Model |   14.794001         4  3.69850024   Prob > F        =    0.0000
    Residual |   677.18008     2,226  .304213873   R-squared       =    0.0214
-------------+----------------------------------   Adj R-squared   =    0.0196
       Total |  691.974081     2,230  .310302279   Root MSE        =    .55156

------------------------------------------------------------------------------
    absresid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.0024566   .0234312    -0.10   0.917    -.0484059    .0434927
     educate |   .0241652   .0043845     5.51   0.000      .015567    .0327635
    nonwhite |   .0088975   .0330742     0.27   0.788     -.055962    .0737571
       pasei |    .001275   .0006816     1.87   0.062    -.0000616    .0026116
       _cons |   .4418138   .0596917     7.40   0.000     .3247565     .558871
------------------------------------------------------------------------------
```

We can now conclude with some certainty that education is involved in a heteroscedasticity issue since it appears to have a substantial association with this form of the residuals.

But what do we do in this situation? Some experts suggest using weighted least squares (WLS) estimation (Chatterjee and Hadi, 2006, Chapter 7). Since we have evidence that education is implicated in the problem, we may wish to explore further using a WLS model with some transformation of education as a weighting factor. It is much simpler, though, to rely on the Huber–White sandwich estimator to compute robust standard errors (Fox, 2016, Chapter 12). This is a simple maneuver in Stata. By adding the subcommand `robust` after the `regress` command, Stata provides the robust standard errors.

Example 1.6

```
regress sei female educate nonwhite pasei, robust
```

```
Linear regression                               Number of obs     =      2,231
                                                F(4, 2226)        =     277.88
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3544
                                                Root MSE          =     15.499

------------------------------------------------------------------------------
             |               Robust
         sei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.2358993   .6587073    -0.36   0.720    -1.527644    1.055846
     educate |    3.72279   .1338505    27.81   0.000     3.460305    3.985274
    nonwhite |  -2.860583   .9346099    -3.06   0.002    -4.693381   -1.027784
       pasei |   .0765712    .020297     3.77   0.000     .0367683    .1163742
       _cons |  -4.804311   1.713414    -2.80   0.005    -8.164367   -1.444255
------------------------------------------------------------------------------
```

Notice that the standard errors in the model in Example 1.6 are generally larger than in the model displayed in Example 1.2. Since we are most concerned with education, it is interesting that its coefficient's standard error increased from about 0.123 to 0.134, even though its *p*-value indicates that it remains statistically significant. Thus, even in the presence of heteroscedasticity, education appears to be a relevant predictor of *sei* in these data (although other evidence is needed to substantiate this conclusion).

Another popular option for attenuating the effects of heteroscedasticity is to estimate bootstrapped standard errors (consider `help bootstrap`). This is a resampling technique that takes repeated samples, with replacement, from the sample data set and calculates the model for each of these samples. It then uses this information to estimate a sampling distribution of the coefficients from the model, including the standard errors (Guan, 2003). This usually results in estimates that are affected less by heteroscedacity or other problematic issues.

However, it is critical to remember that issues such as heteroscedasticity do not necessarily signal that there is something wrong with the data or the model. It is important to think carefully about *why* there is heteroscedasticity or other seemingly nettlesome issues. Considering why these issues occur may lead to a better understanding of the associations that interest us. Moreover, there might be additional issues directly or indirectly involving heteroscedasticity, so we will explore some more topics after reviewing the other assumptions—as well as a few other characteristics—of the OLS regression model.

3. Autocorrelation

This assumption states that there is *no autocorrelation among the errors*. In other words, they are not correlated based on time or space. Autocorrelation is related to the independence assumption: when errors are correlated, they are not independent. Thus, the main consequence of violating this assumption is underestimated standard errors. Graphs of the residuals across time (or space in spatial regression models) are typically used as a diagnostic tool. Stata also offers a Durbin–Watson test that is invoked after a regression model (`estat dwatson`) and a whole range of models designed to adjust time-series and longitudinal models for the presence of autocorrelation. Two useful regression models are Prais–Winsten and Cochran–Orcutt regression (Wooldridge, 2010), which are implemented in Stata with the command `prais`. There are also tools available for spatial autocorrelation: errors that are correlated across space. Chapter 8 reviews regression models that are appropriate for longitudinal data, which often experience autocorrelation.

4. Collinearity

There is no *perfect collinearity* among the predictors. The model cannot be estimated when this occurs. It should be noted, though, that problems can arise with standard errors and regression coefficients even when there is high collinearity. Recall that the standard error formula shown earlier included the tolerance $\left(1 - R_i^2\right)$ in the denominator. This, you may remember, is based on regressing each explanatory variable on all the other explanatory variables. If there is perfect collinearity, then at least one of these R^2 values is 1, and the tolerance is 0. Thus, the standard error for this particular coefficient cannot be estimated. However, even very small tolerance values may lead to odd and unstable results.

The Stata post-estimation command `vif` provides *variance inflation factors* (VIFs) for an OLS regression model. For example, if we follow Example 1.2 by typing `vif`, Stata returns the next set of results.

```
vif                                              * after example 1.2
```

```
    Variable |       VIF       1/VIF  
-------------+----------------------
       pasei |      1.18    0.845523
     educate |      1.17    0.855440
    nonwhite |      1.02    0.984092
      female |      1.00    0.996199
-------------+----------------------
    Mean VIF |      1.09
```

The values in the table show no evidence of multicollinearity. But what might demonstrate evidence of a potential problem? Fox (2016) suggests examining the square root of the VIFs and cautions that the estimates are substantially affected when this value is between 2 and 3. However, some researchers point out that other aspects of the model are just as consequential as collinearity for getting stable results (O'Brien, 2007). In any event, there are some downloadable commands in Stata that provide other collinearity diagnostics, such as condition indices, that are called `collin` and `coldiag2`. Typing and entering `findit coldiag2` in the command line searches for locations that have this command and provide instructions about how to download it.

Assumptions 1–4 are the most consequential for the OLS regression model in terms of the earlier discussion of unbiasedness, efficiency, and consistency. According to the Gauss–Markov theorem, when these assumptions are satisfied, the OLS estimates offer the BLUE among the class of linear estimators (see Lindgren, 1993, 510). No other linear estimator has lower bias, or is more precise, on average, than OLS estimates. Nevertheless, there are some additional assumptions that are important for this model.

5. Error Distribution

The errors are normally distributed with a mean of zero and constant variance. This statement has three parts. The first part is also known as the *normality assumption*. If the second part ($E(\varepsilon_i) = 0$) is violated, the intercept is biased. The last part of this statement is simply another way of stating Assumption 2. Thus, the key issue here is whether the errors follow a normal distribution. If there is non-normality, we may get misleading regression coefficients and standard errors. Nevertheless, some studies point out that even when the errors are not normally distributed, but large samples are utilized, OLS regression models yield unbiased and

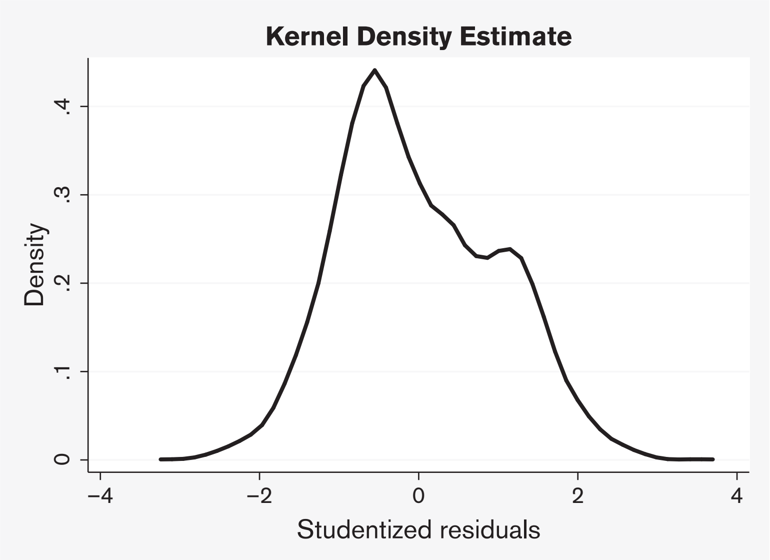

FIGURE 1.5

efficient coefficients (Lumley et al., 2002), with normally distributed intercepts and slopes (Weisberg, 2013). However, the degree of non-normality can make a difference (see Appendix B).

Saving the residuals and checking their distribution with a kernel density plot, a histogram with a normal curve overlaid, or with *q–q* plots and *p–p* plots, is the most straightforward method of testing this assumption. Stata has a `qnorm` and a `pnorm` command available for these. The `pnorm` graph is sensitive to non-normality in the middle range of data and the `qnorm` graph is sensitive to non-normality near the tails. For instance, after estimating the model in Example 1.2, save the studentized residuals and subject them to a kernel density plot, a `qnorm` plot, and a `pnorm` plot.

Figures 1.5 and 1.6 provide only a small kernel (pun intended) of evidence that there is non-normality in the tails. But figures 1.5 and 1.7 suggest there is a modest amount of non-normality near the middle of the distribution. Consider the variation from normality in the 0–2 range of the kernel density plot.

```
predict rstudent, rstudent                    * after Example 1.2
kdensity rstudent                             * try it with and without, normal
```

We examine some issues involving distributional problems later in the chapter.

```
qnorm rstudent
```

```
pnorm rstudent
```

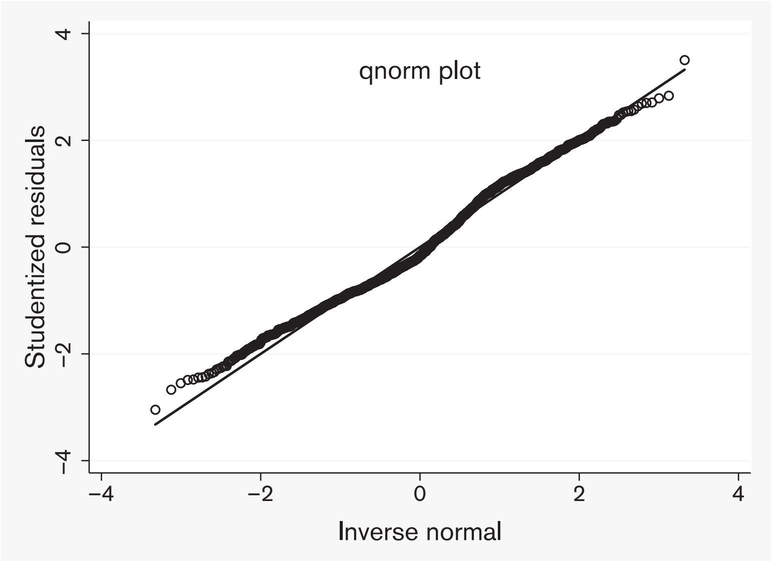

FIGURE 1.6

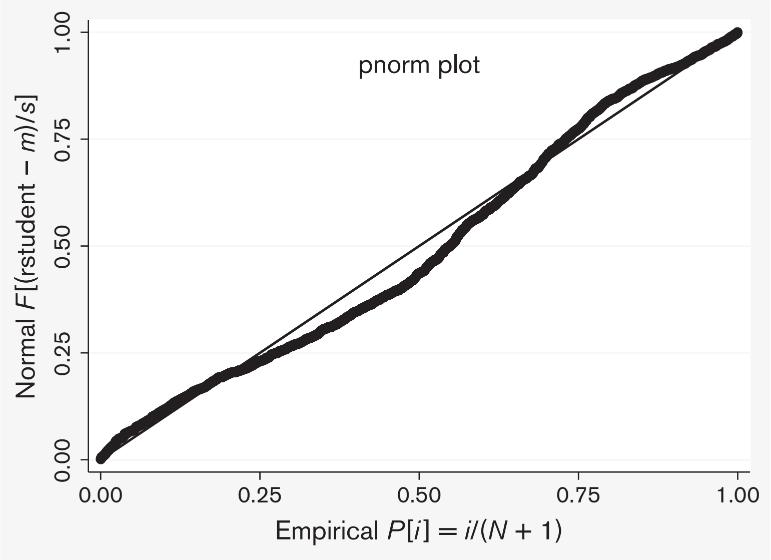

FIGURE 1.7

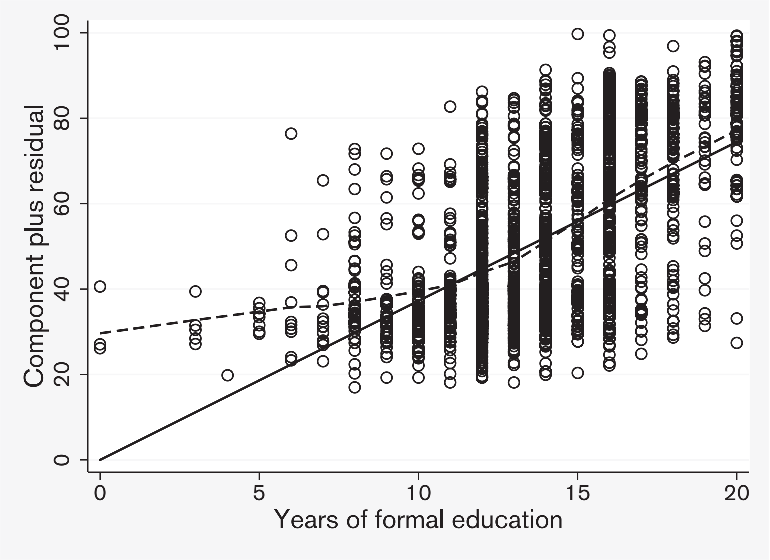

FIGURE 1.8

6. Linearity

The mean value of y for each specific combination of the x is a linear function of the x. A simple way to understand this assumption is to consider that researchers should use linear estimators, like OLS, to estimate linear associations. If the association is not linear, then they should use a nonlinear approach or else there is a risk of obtaining misleading predictions.[8] For example, imagine if an *x* variable and *y* variable have a U-shaped relationship: the slope from an OLS regression model is zero, even though there is certainly an association between them.

In Stata, we may test this assumption by estimating partial residual plots or residual-by-predicted plots and looking for nonlinear patterns. For example, use the post-estimation command below to view a partial residual plot given in figure 1.8. This is also called a *component plus residual plot*. Adding a lowess line makes it easier to detect nonlinearities in the associations. It appears that there are two linear associations between the residuals and education (above and below 10 or 11 years). These might be a concern or suggest some modifications to the model. We consider this issue in a bit more detail later.

```
cprplot educate, lowess                          * after Example 1.2
```

[8] One way to modify the OLS regression model to allow for a particular type of nonlinear association is presented later in the chapter.

7. Specification

For each explanatory variable, the correlation with the error term is zero. If not, then there is specification error in the model. This is also known as *misspecification bias. Omitted variable bias* is one type: an explanatory variable that should be included in the model is not. *Endogeneity bias* is another type. Recall that a property of the residuals is that they are uncorrelated with the predicted values of the outcome variable $(r(\varepsilon,\hat{y})=0)$, yet the predicted values are a function of the explanatory variables. Therefore, the most consequential omitted variables are those that are correlated with one or more of the explanatory variables. Using the wrong functional form—such as when linear term is used rather than a more appropriate nonlinear term (see the linearity assumption)—is another type. Misspecification bias usually results in incorrect standard errors because it can cause the independence assumption to be violated (recall that the errors are assumed independent). The slopes are also affected when misspecification bias is present. Stata has several approaches for examining specification problems, although none is a substitute for a good theory.

The most straightforward test for OLS regression models is Stata's `linktest` command. This command is based on the notion that if a regression model is properly specified, we should not be able to find any additional explanatory variables that are statistically significant except by chance (Pregibon, 1980). The test creates two new variables: the predicted values, denoted *_hat*, and squared predicted values, denoted *_hatsq*. The model is then reestimated using these two variables as predictors. The first, *_hat*, should be statistically significant since it is the predicted value. On the other hand, *_hatsq* should not be significant, because if our model is specified correctly, the squared predictions should not have explanatory power. That is, we would not expect *_hatsq* to be a significant predictor if our model is specified correctly. So, we should assess the *p*-value for *_hatsq*.

Example 1.7

* *after Example 1.2*

```
linktest
```

```
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(2, 2228)      =    667.86
       Model |  310443.029         2  155221.514   Prob > F        =    0.0000
    Residual |  517822.052     2,228  232.415643   R-squared       =    0.3748
-------------+----------------------------------   Adj R-squared   =    0.3742
       Total |  828265.081     2,230  371.419319   Root MSE        =    15.245

------------------------------------------------------------------------------
         sei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |  -.2378828   .1478597    -1.61   0.108    -.5278399    .0520744
      _hatsq |   .0128182   .0015031     8.53   0.000     .0098706    .0157659
       _cons |   28.19505   3.595472     7.84   0.000     21.14422    35.24587
------------------------------------------------------------------------------
```

The results shown in Example 1.7 suggest that there is an omitted variable or some other specification problem. An alternative, but similar, approach is Ramsey's (1969) RESET (*regression specification error test*), which is implemented in Stata using the `ovtest` command.

Example 1.7 (Continued)

```
ovtest
```

```
Ramsey RESET test using powers of the fitted values of sei
       Ho:  model has no omitted variables
                F(3, 2223) =     32.54
                  Prob > F =      0.0000
```

This test has as its null hypothesis that there is no specification error. It is clearly rejected. A third type of test that we do not discuss here, but is available in Stata, is called a Hausman (1978) test. It is more generally applicable than the other two tests. Unfortunately, these tests do not tell us if the misspecification bias is due to omitted variables or because we have the wrong functional form for one or more features of the model. To test for problems with functional forms, we are better off examining partial residual plots and the distribution of the residuals. Theory remains the most important diagnostic tool for understanding this assumption, though.

8. Measurement Error

We assume that y and the x are measured without error. When both are measured with error—a common situation in the social sciences—the regression coefficients are often underestimated and standard errors are incorrect. For example, suppose that only the explanatory variable is measured with error. Here we have x, as we observe it, made up of a true score plus error $\{x^*_{1i} = x_{1i} + v_i\}$. This can be represented with the following:

$$y_i = \alpha + \beta_1 x^*_{1i} + \varepsilon_i = y_i = \alpha + \beta_1 \left(x_{1i} + v_{1i} \right) + \varepsilon_i \,.$$

Distributing by the slope we have the following regression equation:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_1 v_{1i} + \varepsilon_i \,.$$

So the error term now has two components ($\beta_1 v_{1i}$ and ε_i) and x is usually correlated with at least one of them (although, in some cases, it may not be). Thus, measurement error is a form of misspecification bias and violates the independence assumption. But the estimated slope is biased, usually towards zero (*attenuation bias*). The degree of bias is typically unknown, though; it depends on how much error there is in the measurement of x. The standard error of the slope is also incorrect since the sum of squares of x is not accurate.

Unfortunately, variables in the social and behavioral sciences are often measured with error, especially when using survey data. Stata offers several models designed to address measurement error, including two- and three-stage least squares (instrumental variables approaches; see `ivregress`), a command called *`eivreg`* (errors-in-variables regression) where the analyst may set the reliabilities of the variables measured with error, and several other approaches. Type `search measurement error, all` in the command window (or use the help menu) to examine some of these options. Chapter 10 discusses some commonly used techniques for estimating latent variables, which allow one type of adjustment for measurement error.

9. Influential Observations

Strictly speaking, this is not an assumption of the OLS regression model. However, you may recall that influential observations include outliers and high leverage values that can affect the model in untoward ways. Some leverage points are bad and others are not so bad. Leverage points that fall along the regression line or plane are usually acceptable, even though they are relatively far away from the other values. Bad leverage points are extreme on the joint distribution of the explanatory variables. Their main influence is on the standard errors of the coefficients (consider the standard error equation shown earlier in the chapter; leverage points can affect the denominator). Outliers—also known as *vertical outliers*—are extreme on the outcome variable and do not fall near the regression line or surface. Outliers are especially problematic because they can influence the coefficients to a substantial degree (recall that OLS minimizes the *squared distances* from the observations to the regression line or surface).

Consider figure 1.9, which provides a scatterplot using a small set of data. It shows the effect of an outlier on the slope, with the solid line representing the slope with the outlier and the dashed line representing the slope without the outlier. Of course, the first question we should ask is, why does this outlier exist?

Stata allows us to save studentized residuals, leverage values, and Cook's *D* values (as well as other quantities such as DFFITS and Welch's distances) to check for these types of observations. It also has some automated graphs, such as the `lvr2plot`. This graph examines the leverage values against the squared values of the standardized residuals (which are positive), thus providing a useful way to check for high leverage values and outliers at the same time. Figure 1.10 provides an example from our *sei* regression model.

```
lvr2plot                                    * after Example 1.2
```

The two reference lines are the means for the leverage values (horizontal) and for the normalized residual squared (vertical). It appears there is a cluster of observations that are relatively high leverage values and one observation in particular that is a substantial outlier (see the right-hand portion of figure 1.10).

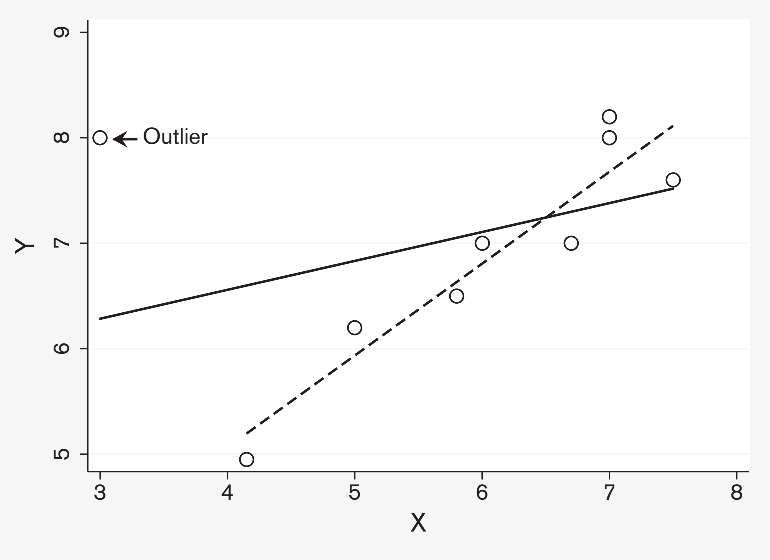

FIGURE 1.9

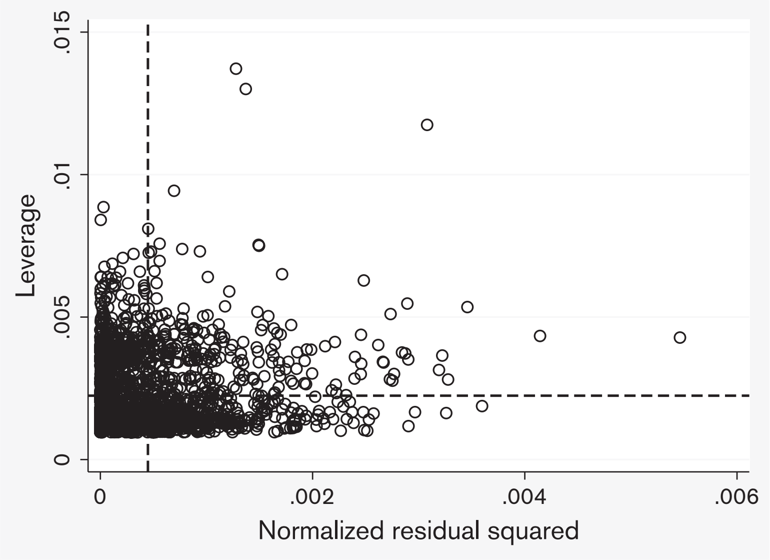

FIGURE 1.10

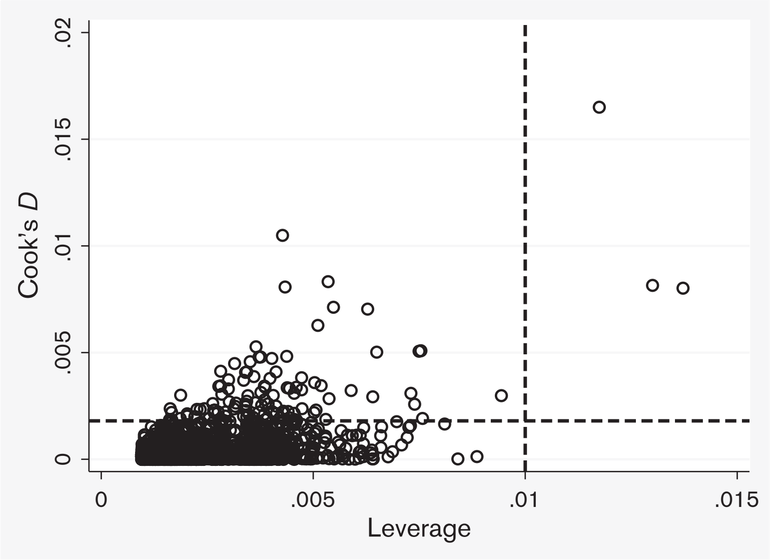

FIGURE 1.11

The Cook's *D* and leverage values may be saved using the commands given below. It is then a good idea to examine these values with some exploratory methods, such as stem-and-leaf plots and box-and-whisker plots. Another useful visualization is a scatterplot of Cook's *D* values and leverage values with reference lines at the thresholds of each. A common Cook's *D* threshold is $4/(n - k - 1)$ or, for our model, $4/(2{,}231 - 4 - 1) = 0.0018$. There are also two frequently used thresholds for determining high leverage values: $2(k + 1)/n$ and $3\hat{h}$ or three times the average leverage value. Given the large sample size, we use the latter threshold, which rounds to 0.01 for our model. Figure 1.11 provides the graph.

```
predict cook, c                                         * after Example 1.2
predict leverage, leverage
twoway scatter cook leverage, yline(0.0018) xline(0.01)
```

The `twoway` subcommand places reference lines in the graph that designate the thresholds. The graph shows a few outliers that are not high leverage values (in the upper left quadrant) and at least three points that are outliers and high leverage points (upper right quadrant). These deserve further exploration (jittering the points using *jitter(5)* as a subcommand may reveal the number of points better).

At this juncture, we may wish to use a robust regression model or a median regression model to minimize the effects of the outliers. As an example, Example 1.8 provides a median regression model in Stata that is invoked with the *quantile regression* (`qreg`) command.

Example 1.8

```
qreg sei female educate nonwhite pasei
```

```
Median regression                                  Number of obs =      2,231
  Raw sum of deviations 18357.65 (about 39.700001)
  Min sum of deviations 14061.14                   Pseudo R2     =     0.2340

------------------------------------------------------------------------------
         sei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3982264   .9046558    -0.44   0.660    -2.172284    1.375831
     educate |   4.482618   .1692831    26.48   0.000     4.150648    4.814587
    nonwhite |  -2.093388   1.276964    -1.64   0.101    -4.597552    .4107768
       pasei |   .1090661   .0263158     4.14   0.000     .0574601    .1606721
       _cons |  -18.73586   2.304638    -8.13   0.000    -23.25532   -14.21639
------------------------------------------------------------------------------
```

Median regression differs from OLS regression by minimizing the distances between the predicted values and the median values for the outcome variable. So it is not as affected by outliers (i.e., it is a *robust* technique). Note that the general results of this model are not much different than what we have seen before. The *nonwhite* coefficient has a *p*-value somewhat above the common threshold level, but education and parents' socioeconomic status remain predictive of *sei*. Alternatives to `qreg` include `rreg` (robust regression) and `mmregress`, which has some desirable properties when the model is affected by influential observations. Kneib (2013) provides a good overview of some alternatives to linear regression based on means.

MODIFYING THE OLS REGRESSION MODEL

We have now reviewed how to test the assumptions of the model and discussed a few ideas about what to do when they are violated. However, we focused on simply correcting the model for some of these violations, such as heteroscedasticity and influential observations, with an eye toward getting unbiased or more precise coefficients. But we should also think carefully about model specification in light of what the tests have shown. For example, recall that education and *sei* may be involved in a nonlinear association, that the residuals from the model are not quite normally distributed, and that education is involved in the heteroscedastic errors. These may or may not be related, but are probably worth exploring in more detail. Some detective work also allows us to illustrate alternative regression models.

To begin, remember that the residuals from the model, as well as the original *sei* variable, are not normally distributed. So we should consider a transformation to normality, if we can find one. Recall also that taking the natural logarithm ($\log_e$) of a variable with a long tail often normalizes its distribution. In Stata, we may do this with the following command:

```
generate logsei = log(sei)
```

According to a kernel density graph, it appears we have made things worse; there is now a clear bimodal distribution that was induced by the transformation (determine this for yourself; why did it occur?). Nonetheless, for illustrative purposes, reestimate the model using *logsei* as the outcome variable (see Example 1.9).[9]

The directions of the effects are the same, and the standardized effects based on the beta weights are similar. But, of course, the scale has changed so the interpretations are different. For example, we could now interpret the education coefficient in this way:

- Adjusting for the effects of sex, nonwhite, and parents' socioeconomic status, each 1-year increase in education is associated with a 0.074 log-unit increase in socioeconomic status.

Example 1.9

```
regress logsei female educate nonwhite pasei, beta
```

```
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(4, 2226)      =    291.03
       Model |  120.196619         4  30.0491548   Prob > F        =    0.0000
    Residual |  229.833462     2,226  .103249534   R-squared       =    0.3434
-------------+----------------------------------   Adj R-squared   =    0.3422
       Total |  350.030081     2,230  .156964162   Root MSE        =    .32132

------------------------------------------------------------------------------
      logsei |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      female |  -.0100858   .0136505    -0.74   0.460                -.0127139
     educate |   .0743256   .0025543    29.10   0.000                 .5403255
    nonwhite |  -.0800009   .0192683    -4.15   0.000                -.0718826
       pasei |   .0016843   .0003971     4.24   0.000                 .0792274
       _cons |   2.738645   .0347751    78.75   0.000                        .
------------------------------------------------------------------------------
```

Another convenient way to interpret regression coefficients when we have a *log-linear model* (since the logarithm of the outcome variable is used, the model is assumed linear on a logarithmic scale) is to use a *percentage change formula* that, in general form, is

$$\% \text{ change} = 100 \times \left\{ \exp\left(\beta_i\right) - 1 \right\}.$$

In our example, we can transform the education coefficient using this formula. Taking advantage of Stata's `display` command (which acts as a high-end calculator), we find

[9] See http://blog.stata.com/2011/08/22/use-poisson-rather-than-regress-tell-a-friend/, for an alternative approach to this model.

```
display 100 × (exp(0.0743) − 1)
7.713
```

How can we use this number in our interpretation? Here is one approach:

- Adjusting for the effects of sex, nonwhite, and parents' socioeconomic status, each one-year increase in education is associated with a 7.71% increase in the socioeconomic status score.

The nonwhite coefficient may be treated in a similar way, keeping in mind that it is a binary variable.

- Adjusting for the effects of sex, education, and parents' socioeconomic status, socioeconomic status scores among nonwhites are expected to be 7.69% lower than among whites.

It is important to reexamine the assumptions of the model to see what implications logging *sei* has for them. The diagnostics reveal some issues that may need to be addressed. The partial residual plot with education still looks odd. Moreover, there remains a problem with heteroscedasticity according to `imtest`, but not `hettest`. Glejser's test shows no signs of this problem, so perhaps we have solved it. There are still some outliers and high leverage points. These could be addressed using a robust regression technique, such as `qreg` or `mmregress`, or by assessing the variables more carefully.

EXAMINING EFFECT MODIFICATION WITH INTERACTION TERMS

In this section, we briefly review *interaction terms*. These are used when the analyst suspects that there is effect modification in the model. Another way of saying this is that some third variable moderates (or modifies) the association between an explanatory and the outcome variable. For example, if you've worked previously with the *GSS* data set used here, you may recall that education moderates the association between gender and personal income.

Do you have any ideas for moderating effects in our socioeconomic status model? Consider education, parents' socioeconomic status, and socioeconomic status. One hypothesis is that parents' status matters a great deal for one's own status, but this can be overcome by completing more years of formal education. Hence, education may moderate the association between parents' status and one's own status. To test this model, we introduce an interaction term in the model using Stata's factor variables options (`help factor variables`). Since including interaction terms often induces collinearity issues, the regression model is followed by a request for the VIFs (see Example 1.10).

Example 1.10

```
regress sei female nonwhite c.educate##c.pasei
```

```
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(5, 2225)      =    245.12
       Model |  294190.629         5  58838.1259   Prob > F        =    0.0000
    Residual |  534074.451     2,225  240.033461   R-squared       =    0.3552
-------------+----------------------------------   Adj R-squared   =    0.3537
       Total |  828265.081     2,230  371.419319   Root MSE        =    15.493

------------------------------------------------------------------------------------
               sei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
            female |  -.2225939   .6582242    -0.34   0.735    -1.513392    1.068204
          nonwhite |  -2.928681   .9299653    -3.15   0.002    -4.752371    -1.10499
           educate |   3.235946   .3204869    10.10   0.000     2.607461     3.86443
             pasei |   -.077266   .0954341    -0.81   0.418    -.2644152    .1098833
                   |
c.educate#c.pasei  |   .0106026   .0064437     1.65   0.100    -.0020337     .023239
                   |
             _cons |   2.064548   4.498666     0.46   0.646    -6.757475    10.88657
------------------------------------------------------------------------------------
```

```
vif
```

```
    Variable |       VIF       1/VIF
-------------+----------------------
      female |      1.00    0.996049
    nonwhite |      1.02    0.982143
     educate |      7.92    0.126332
       pasei |     29.39    0.034030
  c.educate#
     c.pasei |     46.35    0.021574
-------------+----------------------
    Mean VIF |     17.14
```

Although we might wish to make some interpretations from this model, we should first pause and consider what the VIFs indicate. These are typically evaluated in models with interaction terms. When we multiply the values of two variables together, the resulting variable is usually highly correlated with its constituent variables. For example, the correlation between *pasei* and the interaction term is 0.923. The VIFs show the implications of the high correlations. Consider the square root of the largest VIF: $\sqrt{46.35} = 6.8$. This is substantial (Fox, 2016). What implication does it have for the model? It is unclear, but we can be confident that the standard errors for these two variables are inflated in the regression model. What is a solution? One approach is to take the *z*-scores of education and parents' status and reestimate the model using the *z*-scores of the variables along with the updated interaction term. Example 1.11 provides the steps for carrying this out.

Example 1.11

```
egen zeducate = std(educate)                 * calculate z-scores of educate
egen zpasei = std(pasei)                        * calculate z-scores of pasei
regress sei female nonwhite c.zeducate##c.zpasei
```

```
      Source |       SS           df       MS      Number of obs   =     2,231
-------------+----------------------------------   F(5, 2225)      =    245.12
       Model |  294190.629         5  58838.1259   Prob > F        =    0.0000
    Residual |  534074.451     2,225  240.033461   R-squared       =    0.3552
-------------+----------------------------------   Adj R-squared   =    0.3537
       Total |  828265.081     2,230  371.419319   Root MSE        =    15.493

------------------------------------------------------------------------------------
                   sei |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------+------------------------------------------------------------
                female |  -.2225939   .6582242    -0.34   0.735    -1.513392    1.068204
              nonwhite |  -2.928681   .9299653    -3.15   0.002    -4.752371    -1.10499
              zeducate |   10.95155   .3619516    30.26   0.000     10.24175    11.66135
                zpasei |   1.199235   .3820587     3.14   0.002     .4500064    1.948464
                       |
c.zeducate#c.zpasei    |   .5779742   .3512613     1.65   0.100    -.1108601    1.266808
                       |
                 _cons |   48.36958   .5122487    94.43   0.000     47.36505    49.37412
------------------------------------------------------------------------------------
```

```
vif
```

```
    Variable |       VIF       1/VIF
-------------+----------------------
      female |      1.00    0.996049
    nonwhite |      1.02    0.982143
    zeducate |      1.18    0.849157
      zpasei |      1.36    0.735944
  c.zeducate#|
    c.zpasei |      1.15    0.866698
-------------+----------------------
    Mean VIF |      1.14
```

The main implication is that the coefficient for parents' status is now statistically significant. This might be because, before taking the *z*-scores, its standard error was unduly inflated by collinearity (although there could be other reasons). Moreover, there are positive coefficients for education, parents' status, and their interaction term. This suggests that the slope of the association between parents' status and one's own status is slightly steeper at higher levels of education, although note that the slope for the interaction term is not statistically significant ($p = 0.1$). This fails to support the hypothesis.

Nonetheless, assume that we did find an interesting association. Instead of relying only on the regression coefficients and their directional effects, it is helpful to graph predicted values for the different groups represented by the moderator. Here is one approach to this that uses the predicted values from the OLS regression model and examines three categories

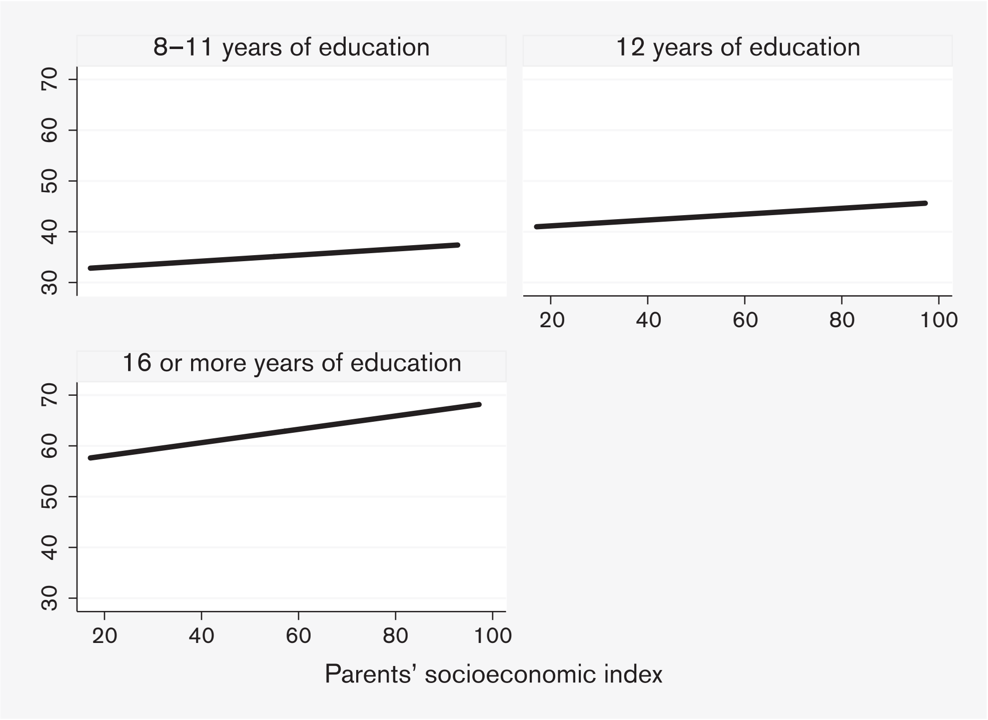

FIGURE 1.12

of education. (The `margins` and `marginsplot` commands could also be used to construct a similar graph.)

```
predict seihat, xb                                  * after Example 1.10
recode educate (0/7=.)(8/11=1)(12=2)
  (13/15=.)(16/max=3), generate(cateducate)
```

We have divided the education variable into three parts: those who did not graduate from high school, those who graduated from high school but did not attend college, and those who graduated from college, some of whom may have attended graduate school. We then request graphs that include linear fit lines for each parent socioeconomic status–*sei* association (see figure 1.12). This provides a simple way to compare the slopes for particular education groups to see if they appear distinct.

```
twoway lfit seihat pasei, by(cateducate)
```

There are only slight differences in these three slopes. Recall that, even after taking care of the collinearity problem, the coefficient for the interaction term was not statistically significant, so the graphs should not be surprising. The combination of these graphs and the unremarkable interaction term in Example 1.11 should persuade us that there is not much, if any, effect modification in this situation.

ASSESSING ANOTHER OLS REGRESSION MODEL

As a final exercise in this review of OLS regression, use the *USData* data set (*usdata.dta*) and assess a model that examines violent crimes per 100,000 population (*violrate*) as the outcome variable and the following explanatory variables: unemployment rate (*unemprat*), gross state product (*gsprod*), and state migration rate (*mig_rate*) (see Example 1.12). Then, examine some of the assumptions of the model, judge whether there are violations, and consider what we might do about them. The following examples provide some guidance for executing these steps.

Example 1.12

```
regress violrate unemprat gsprod mig_rate
```

```
      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(3, 46)        =     10.07
       Model |  1408073.67         3  469357.889   Prob > F        =    0.0000
    Residual |  2143815.79        46  46604.6911   R-squared       =    0.3964
-------------+----------------------------------   Adj R-squared   =    0.3571
       Total |  3551889.46        49  72487.5399   Root MSE        =    215.88

------------------------------------------------------------------------------
    violrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    unemprat |   71.64414   25.02161     2.86   0.006     21.27825      122.01
      gsprod |   80.38644   19.76907     4.07   0.000     40.59336    120.1795
    mig_rate |   .0176516   .0084813     2.08   0.043     .0005795    .0347236
       _cons |   29.22665   132.9562     0.22   0.827    -238.4003    296.8536
------------------------------------------------------------------------------
```

```
cprplot mig_rate, lowess
```

What does the graph in figure 1.13 suggest? What might we do about it? Check the other explanatory variables also.

```
rvfplot
```

What does the residuals-by-fitted plot show (figure 1.14)? What other tests that are related to the test that this graph provides would you like to see? Is there a problem and, if yes, what should we do about it?

Check for multicollinearity. Are there any problems? How can you tell?

What does the following test indicate about an assumption of the model? Which assumption? Should we have any additional concerns regarding this assumption after using this test?

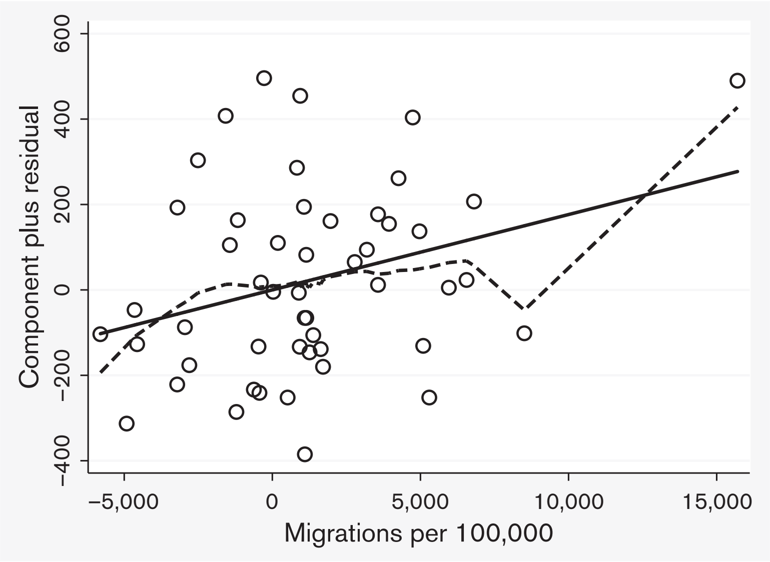

FIGURE 1.13

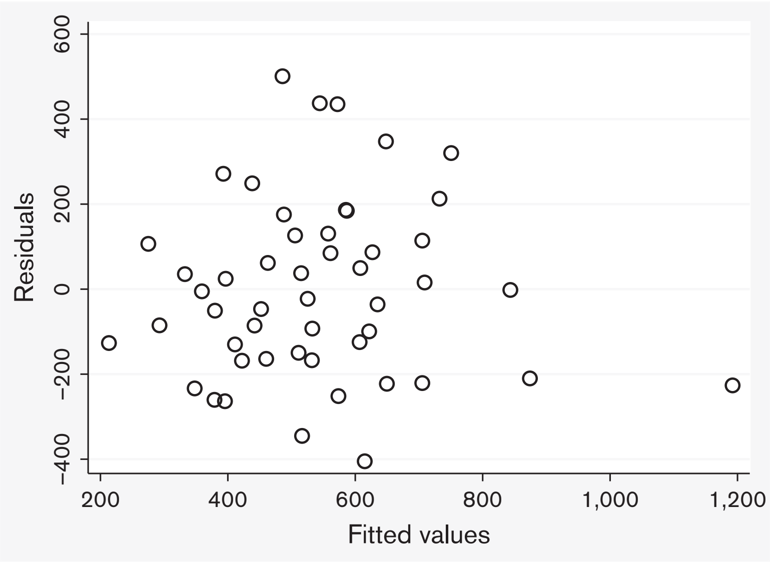

FIGURE 1.14

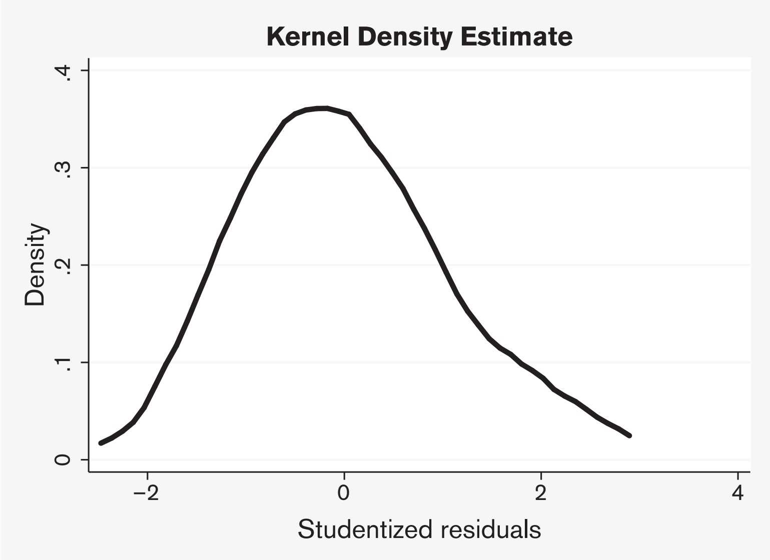

FIGURE 1.15

```
linktest
```

```
      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(2, 47)        =     17.59
       Model |  1520753.83         2  760376.915   Prob > F        =    0.0000
    Residual |  2031135.63        47  43215.6516   R-squared       =    0.4282
-------------+----------------------------------   Adj R-squared   =    0.4038
       Total |  3551889.46        49  72487.5399   Root MSE        =    207.88

------------------------------------------------------------------------------
    violrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   2.110521   .7097014     2.97   0.005     .6827859    3.538256
      _hatsq |  -.0008759   .0005424    -1.61   0.113    -.0019671    .0002153
       _cons |  -319.5151   221.2434    -1.44   0.155    -764.5994    125.5693
------------------------------------------------------------------------------
```

```
kdensity rstudent                      * rstudent = studentized residuals
```

Consider figure 1.15. It shows the distribution of the studentized residuals from our model. Is there any cause for concern? Why or why not?

Finally, after saving two types of values using Stata's `predict` post-command, generate figure 1.16. What does it represent? What is the source of the two reference lines? What do you conclude based on this graph? What should we do in this situation?

```
twoway scatter cook leverage, yline(0.087) xline(0.16)
  mlabel(state)
```

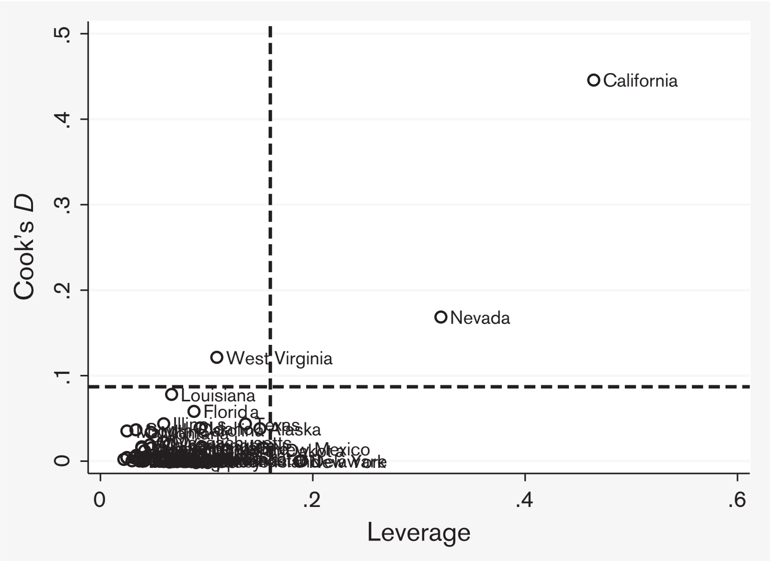

FIGURE 1.16

Do you have any concluding thoughts about this OLS regression model?

Consider the following as a final step in this exercise. Why is California an influential observation in the model? The answer is that it has a very large gross state product value. Examine the distribution of gross state product. Then, as shown in Example 1.13, take its natural logarithm (plus one to avoid negative values) and reestimate the model, check the partial residual plot with the log of gross state product, and reestimate the influential observations graph.

Example 1.13

```
gen loggsprod = log(gsprod + 1)
regress violrate unemprat loggsprod mig_rate
```

```
      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(3, 46)        =     14.13
       Model |   1703254.4         3  567751.468   Prob > F        =    0.0000
    Residual |  1848635.05        46  40187.7185   R-squared       =    0.4795
-------------+----------------------------------   Adj R-squared   =    0.4456
       Total |  3551889.46        49  72487.5399   Root MSE        =    200.47

------------------------------------------------------------------------------
    violrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    unemprat |   79.36568   22.95864     3.46   0.001     33.15234     125.579
   loggsprod |   300.4235   58.33766     5.15   0.000     182.9959    417.8511
    mig_rate |   .0169867   .0077277     2.20   0.033     .0014316    .0325418
       _cons |  -117.0315   128.1596    -0.91   0.366    -375.0033    140.9403
------------------------------------------------------------------------------
```

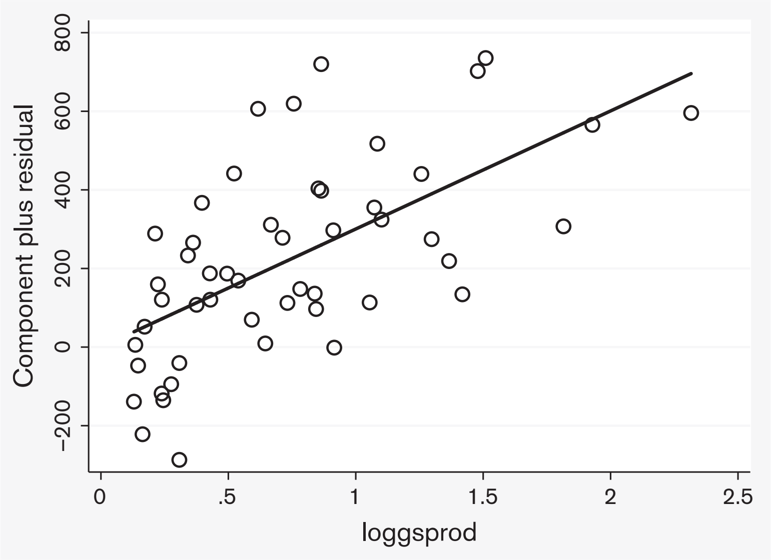

FIGURE 1.17

```
cprplot loggsprod                                        * see figure 1.17

predict cook, c
predict leverage, leverage
twoway scatter cook leverage, yline(0.087) xline(0.16)
  mlabel(state)
```

Nevada remains a high leverage point and is now the most extreme outlier (see figure 1.18). It is not clear why, but an in-depth exploration shows that it has a relatively high migration rate compared to other states. This may lead to the high leverage value. Dropping Nevada from the model leads to a better fit and fewer influential observations, but migration rate is no longer a statistically significant predictor (see Example 1.14 and figure 1.19).

However, before we close the case on migration and violent crimes, we need to think carefully about what we have done. Do we have sufficient justification for dropping Nevada from the analysis? Are there other steps we should take first before doing this? Do not forget that dropping legitimate observations is a dubious practice. It is much better to consider the source; for example, why does Nevada have a migration rate that appears to be substantially higher than in other states? Does the association between violent crimes and the migration rate have any particular implications for a state such as Nevada? There are numerous paths one might take to understand the variables and the model better.

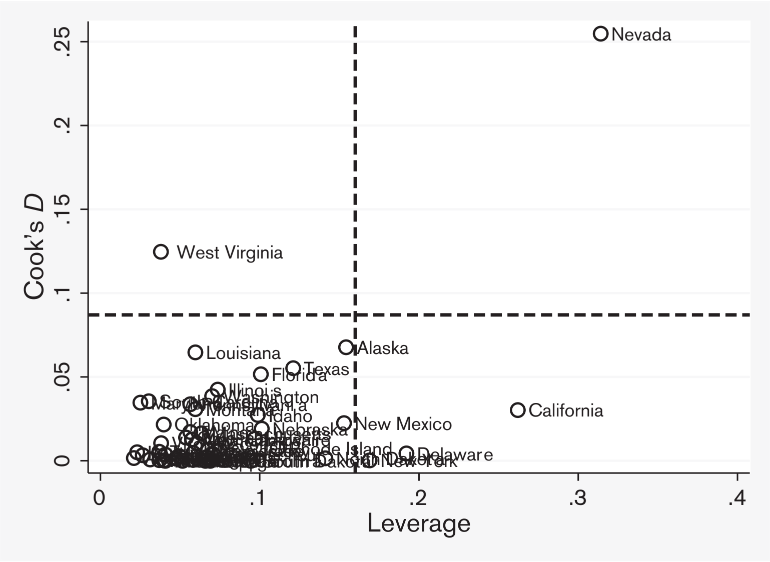

FIGURE 1.18

Example 1.14

```
regress violrate unemprat loggsprod mig_rate if state
  ~= “Nevada”
```

```
      Source |       SS           df       MS      Number of obs   =        49
-------------+----------------------------------   F(3, 45)        =     13.85
       Model |  1624737.77         3  541579.256   Prob > F        =    0.0000
    Residual |  1759165.58        45  39092.5684   R-squared       =    0.4801
-------------+----------------------------------   Adj R-squared   =    0.4455
       Total |  3383903.35        48  70497.9864   Root MSE        =    197.72

------------------------------------------------------------------------------
    violrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    unemprat |    77.8264   22.66651     3.43   0.001     32.17372    123.4791
   loggsprod |   295.7535   57.62004     5.13   0.000     179.7007    411.8062
    mig_rate |   .0095692   .0090626     1.06   0.297    -.0086839    .0278222
       _cons |  -103.7971   126.7036    -0.82   0.417    -358.9913    151.3971
------------------------------------------------------------------------------
```

```
predict cook, c
predict leverage, leverage
twoway scatter cook leverage if state ~= “Nevada”,
  yline(0.091) xline(0.204) mlabel(state)
```

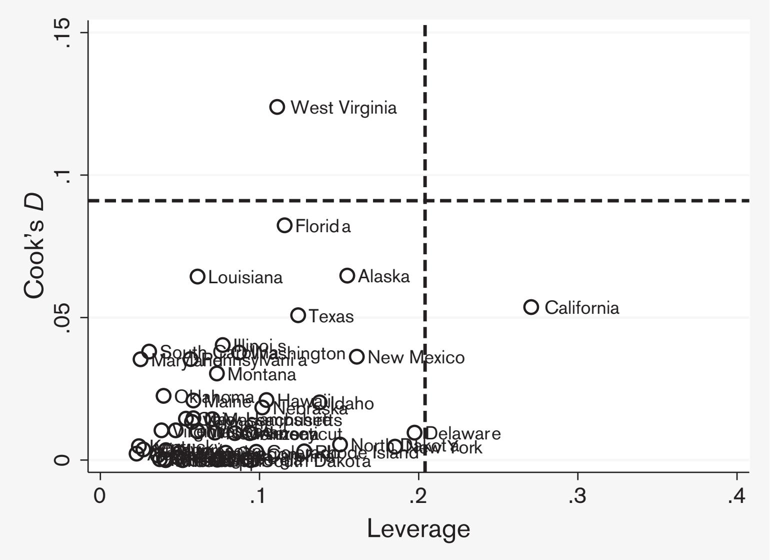

FIGURE 1.19

Finally, why West Virginia (see figure 1.19)? Sort by the Cook's *D* values and list the variables, predicted values, Cook's *D* values, and leverage values. This shows that West Virginia had fewer violent crimes than expected, but it is also relatively high on unemployment. Perhaps something about this is causing it to be an outlier. Given all this effort, should we rely on an OLS regression model to analyze state-level violent crimes or is an alternative model preferable? We do not investigate further here, but you may wish to.

FINAL WORDS

The OLS regression model is a powerful and frequently used statistical tool for assessing associations among variables. If the assumptions discussed in this chapter are satisfied, it can provide good estimates of the coefficients with which to judge these associations. However, meeting these assumptions can be challenging, even though some analysts note that only the first four are critical for achieving the BLUE property. Moreover, even when the errors are not distributed normally, the OLS regression model provides pretty accurate and efficient results. However, there are alternative models that can also provide good results but do not make some of these assumptions. Some of these are also designed for the common situation in the social and behavioral sciences when the outcome variable is not continuous and the relationship between variables is not, strictly speaking, linear. In this situation, linear models may provide questionable results, so, as discussed in subsequent chapters, these alternative models can be valuable.

EXERCISES FOR CHAPTER 1

1. The GPA data set (*gpa.dta*, available on the book's website) includes 20 observations from a sample of college students. Examine the variables to make sure you understand what they are measuring. Then, complete the following exercises.
   a. Construct a scatterplot with *gpa* on the *y*-axis and *sat_quan* on the *x*-axis. What does the scatterplot suggest about their association? Are there any unusual patterns shown in the scatterplot?
   b. Estimate a simple OLS regression model that uses *gpa* as the outcome variable and *sat_quan* as the explanatory variable. Interpret the intercept and the unstandardized coefficient for *sat_quan*.

2. Estimate the following three OLS regression models that each uses *gpa* as the outcome variable:
   a. Use only *hs_engl* as the explanatory variable.
   b. Use *hs_engl* and *sat_verb* as explanatory variables.
   c. Use *hs_engl*, *sat_verb*, and *sat_quan* as explanatory variables.
   d. Interpret the unstandardized coefficient for *hs_engl* from all three models.
   e Interpret the R^2 from the model in 2c.

3. Please describe what happened to the association between *hs_engl* and *gpa* as we moved from the first to the second to the third model. Speculate, in a conceptual way, why this change occurred.

4. Using the model in 2c, check the following OLS regression assumptions:
   a. Normality of the residuals.
   b. Homoscedasticity.

5. Check the model for influential observations.

CHAPTER RESOURCES: SOME USEFUL STATA COMMANDS

Rather than describing each, here is brief list of several of the most useful commands for data management, exploratory analysis, and regression modeling. Type `help` *`command name`* in Stata's command window for more information on each. The Stata User Manuals provide even more information.

```
findit Stata command or statistical procedure
help Stata command or statistical procedure
set memory xx
use "filename.dta"
insheet using "filename.txt"
save "filename.dta"
clear (be careful, this removes data from memory!)
describe
edit
browse
qui command (omits the output)
list varnames
codebook varnames
```

```
reshape data type, variables
recode varname
replace varname
generate newvar = f(oldvarname)
egen newvar = f(oldvarname)
drop varname
summarize varnames, detail
table varname or varname1 varname2
tabulate varname or varname1 varname2
histogram varname, normal
graph bar varname, over(group variable)
graph dot varname, over(group variable)
graph box varname, over(group variable)
kdensity varname
scatter varname1 varname2
lowess varname1 varname2
correlate varlist
spearman varlist
mean varname, over(group variable)
ci varname
ttest varname1 = varname2 or varname, by(group variable)
prtest varname1 = varname2
ranksum varname, by(group variable)
median varname, by(group variable)
regress y-varname x-varnames
predict newvar, type (e.g., residual, predicted value)
test varnames
glm y-varname x-varnames, link(link function) family(distribution)
```