Chapter Nine

More on Specification and Data Problems

In Chapter 8, we dealt with one failure of the Gauss-Markov assumptions. Heteroskedasticity in the errors can be viewed as a model misspecification, but it is a relatively minor one. The presence of heteroskedasticity does not cause bias or inconsistency in the OLS estimators. Also, it is fairly easy to adjust confidence intervals and t and F statistics to obtain valid inference after OLS estimation, or even to get more efficient estimators by using weighted least squares.

In this chapter, we return to the much more serious problem of correlation between the error, u, and one or more of the explanatory variables. Remember from Chapter 3 that if u is, for whatever reason, correlated with the explanatory variable x_j, then we say that x_j is an **endogenous explanatory variable**. We also provide a more detailed discussion on three reasons why an explanatory variable can be endogenous; in some cases, we discuss possible remedies.

We have already seen in Chapters 3 and 5 that omitting a key variable can cause correlation between the error and some of the explanatory variables, which generally leads to bias and inconsistency in *all* of the OLS estimators. In the special case that the omitted variable is a function of an explanatory variable in the model, the model suffers from **functional form misspecification**.

We begin in the first section by discussing the consequences of functional form misspecification and how to test for it. In Section 9.2, we show how the use of proxy variables can solve, or at least mitigate, omitted variables bias. In Section 9.3, we derive and explain the bias in OLS that can arise under certain forms of **measurement error**. Additional data problems are discussed in Section 9.4.

All of the procedures in this chapter are based on OLS estimation. As we will see, certain problems that cause correlation between the error and some explanatory variables cannot be solved by using OLS on a single cross section. We postpone a treatment of alternative estimation methods until Part 3.

9.1 FUNCTIONAL FORM MISSPECIFICATION

A multiple regression model suffers from functional form misspecification when it does not properly account for the relationship between the dependent and the observed explanatory variables. For example, if hourly wage is determined by $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$, but we omit the squared experience term, $exper^2$, then we are

committing a functional form misspecification. We already know from Chapter 3 that this generally leads to biased estimators of β_0, β_1, and β_2. (We do not estimate β_3 because $exper^2$ is excluded from the model.) Thus, misspecifying how *exper* affects log(*wage*) generally results in a biased estimator of the return to education, β_1. The amount of this bias depends on the size of β_3 and the correlation among *educ*, *exper*, and $exper^2$.

Things are worse for estimating the return to experience: even if we could get an unbiased estimator of β_2, we would not be able to estimate the return to experience because it equals $\beta_2 + 2\beta_3 exper$ (in decimal form). Just using the biased estimator of β_2 can be misleading, especially at extreme values of *exper*.

As another example, suppose the log(*wage*) equation is

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \beta_5 female \cdot educ + u, \tag{9.1}$$

where *female* is a binary variable. If we omit the interaction term, *female*·*educ*, then we are misspecifying the functional form. In general, we will not get unbiased estimators of any of the other parameters, and since the return to education depends on gender, it is not clear what return we would be estimating by omitting the interaction term.

Omitting functions of independent variables is not the only way that a model can suffer from misspecified functional form. For example, if (9.1) is the true model satisfying the first four Gauss-Markov assumptions, but we use *wage* rather than log(*wage*) as the dependent variable, then we will not obtain unbiased or consistent estimators of the partial effects. The tests that follow have some ability to detect this kind of functional form problem, but there are better tests that we will mention in the subsection on testing against nonnested alternatives.

Misspecifying the functional form of a model can certainly have serious consequences. Nevertheless, in one important respect, the problem is minor: by definition, we have data on all the necessary variables for obtaining a functional relationship that fits the data well. This can be contrasted with the problem addressed in the next section, where a key variable is omitted on which we cannot collect data.

We already have a very powerful tool for detecting misspecified functional form: the *F* test for joint exclusion restrictions. It often makes sense to add quadratic terms of any significant variables to a model and to perform a joint test of significance. If the additional quadratics are significant, they can be added to the model (at the cost of complicating the interpretation of the model). However, significant quadratic terms can be symptomatic of other functional form problems, such as using the level of a variable when the logarithm is more appropriate, or vice versa. It can be difficult to pinpoint the precise reason that a functional form is misspecified. Fortunately, in many cases, using logarithms of certain variables and adding quadratics is sufficient for detecting many important nonlinear relationships in economics.

EXAMPLE 9.1
(Economic Model of Crime)

Table 9.1 contains OLS estimates of the economic model of crime (see Example 8.3). We first estimate the model without any quadratic terms; those results are in column (1).

Table 9.1

Dependent Variable: *narr86*

Independent Variables	(1)	(2)
pcnv	−.133 (.040)	.533 (.154)
$pcnv^2$	———	−.730 (.156)
avgsen	−.011 (.012)	−.017 (.012)
tottime	.012 (.009)	.012 (.009)
ptime86	−.041 (.009)	.287 (.004)
$ptime86^2$	———	−.0296 (.0039)
qemp86	−.051 (.014)	−.014 (.017)
inc86	−.0015 (.0003)	−.0034 (.0008)
$inc86^2$	———	.000007 (.000003)
black	.327 (.045)	.292 (.045)
hispan	.194 (.040)	.164 (.039)
intercept	.596 (.036)	.505 (.037)
Observations *R*-Squared	2725 .0723	2725 .1035

QUESTION 9.1

Why do we not include the squares of *black* and *hispan* in column (2) of Table 9.1?

In column (2), the squares of *pcnv*, *ptime86*, and *inc86* are added; we chose to include the squares of these variables because each one is significant in column (1). The variable *qemp86* is a discrete variable taking on only five values, so we do not include its square in column (2).

Each of the squared terms is significant and together they are jointly very significant ($F = 31.37$, with $df = 3$ and 2713; the p-value is essentially zero). Thus, it appears that the initial model overlooked some potentially important nonlinearities.

The presence of the quadratics makes interpreting the model somewhat difficult. For example, *pcnv* no longer has a strict deterrent effect: the relationship between *narr86* and *pcnv* is positive up until *pcnv* = .365, and then the relationship is negative. We might conclude that there is little or no deterrent effect at lower values of *pcnv*; the effect only kicks in at higher prior conviction rates. We would have to use more sophisticated functional forms than the quadratic to verify this conclusion. It may be that *pcnv* is not entirely exogenous. For example, men who have not been convicted in the past (so that *pcnv* = 0) are perhaps casual criminals, and so they are less likely to be arrested in 1986. This could be biasing the estimates.

Similarly, the relationship between *narr86* and *ptime86* is positive up until *ptime86* = 4.85 (almost five months in prison), and then the relationship is negative. The vast majority of men in the sample spent no time in prison in 1986, so again we must be careful in interpreting the results.

Legal income has a negative effect on *narr86* until *inc86* = 242.85; since income is measured in hundreds of dollars, this means an annual income of $24,285. Only 46 of the men in the sample have incomes above this level. Thus, we can conclude that *narr86* and *inc86* are negatively related with a diminishing effect.

Example 9.1 is a tricky functional form problem due to the nature of the dependent variable. There are other models that are theoretically better suited for handling dependent variables that take on a small number of integer values. We will briefly cover these models in Chapter 17.

RESET as a General Test for Functional Form Misspecification

There are some tests that have been proposed to detect general functional form misspecification. Ramsey's (1969) **regression specification error test** (**RESET**) has proven to be useful in this regard.

The idea behind RESET is fairly simple. If the original model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \qquad \textbf{(9.2)}$$

satisfies MLR.3, then no nonlinear functions of the independent variables should be significant when added to equation (9.2). In Example 9.1, we added quadratics in the significant explanatory variables. While this often detects functional form problems, it has the

drawback of using up many degrees of freedom if there are many explanatory variables in the original model (much as the straight form of the White test for heteroskedasticity consumes degrees of freedom). Further, certain kinds of neglected nonlinearities will not be picked up by adding quadratic terms. RESET adds polynomials in the OLS fitted values to equation (9.2) to detect general kinds of functional form misspecification.

In order to implement RESET, we must decide how many functions of the fitted values to include in an expanded regression. There is no right answer to this question, but the squared and cubed terms have proven to be useful in most applications.

Let $\hat{y}$ denote the OLS fitted values from estimating (9.2). Consider the expanded equation

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error. \quad \textbf{(9.3)}$$

This equation seems a little odd, because functions of the fitted values from the initial estimation now appear as explanatory variables. In fact, we will not be interested in the estimated parameters from (9.3); we only use this equation to test whether (9.2) has missed important nonlinearities. The thing to remember is that $\hat{y}^2$ and $\hat{y}^3$ are just nonlinear functions of the x_j.

The null hypothesis is that (9.2) is correctly specified. Thus, RESET is the F statistic for testing H_0: $\delta_1 = 0$, $\delta_2 = 0$ in the expanded model (9.3). A significant F statistic suggests some sort of functional form problem. The distribution of the F statistic is approximately $F_{2,n-k-3}$ in large samples under the null hypothesis (and the Gauss-Markov assumptions). The *df* in the expanded equation (9.3) is $n - k - 1 - 2 = n - k - 3$. An *LM* version is also available (and the chi-square distribution will have two *df*). Further, the test can be made robust to heteroskedasticity using the methods discussed in Section 8.2.

EXAMPLE 9.2
(Housing Price Equation)

Using the data in HPRICE1.RAW, we estimate two models for housing prices. The first one has all variables in level form:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u. \quad \textbf{(9.4)}$$

The second one uses the logarithms of all variables except *bdrms*:

$$lprice = \beta_0 + \beta_1 llotsize + \beta_2 lsqrft + \beta_3 bdrms + u. \quad \textbf{(9.5)}$$

Using $n = 88$ houses in HPRICE3.RAW, the RESET statistic for equation (9.4) turns out to be 4.67; this is the value of an $F_{2,82}$ random variable ($n = 88$, $k = 3$), and the associated p-value is .012. This is evidence of functional form misspecification in (9.4).

The RESET statistic in (9.5) is 2.56, with p-value = .084. Thus, we do not reject (9.5) at the 5% significance level (although we would at the 10% level). On the basis of RESET, the log-log model in (9.5) is preferred.

In the previous example, we tried two models for explaining housing prices. One was rejected by RESET, while the other was not (at least at the 5% level). Often, things are not so simple. A drawback with RESET is that it provides no real direction on how to proceed if the model is rejected. Rejecting (9.4) by using RESET does not immediately suggest that (9.5) is the next step. Equation (9.5) was estimated because constant elasticity models are easy to interpret and can have nice statistical properties. In this example, it so happens that it passes the functional form test as well.

Some have argued that RESET is a very general test for model misspecification, including unobserved omitted variables and heteroskedasticity. Unfortunately, such use of RESET is largely misguided. It can be shown that RESET has no power for detecting omitted variables whenever they have expectations that are linear in the included independent variables in the model [see Wooldridge (1995) for a precise statement]. Further, if the functional form is properly specified, RESET has no power for detecting heteroskedasticity. The bottom line is that RESET is a functional form test, and nothing more.

Tests Against Nonnested Alternatives

Obtaining tests for other kinds of functional form misspecification—for example, trying to decide whether an independent variable should appear in level or logarithmic form—takes us outside the realm of classical hypothesis testing. It is possible to test the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad \textbf{(9.6)}$$

against the model

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u, \qquad \textbf{(9.7)}$$

and vice versa. However**,** these are **nonnested models** (see Chapter 6), and so we cannot simply use a standard F test. Two different approaches have been suggested. The first is to construct a comprehensive model that contains each model as a special case and then to test the restrictions that led to each of the models. In the current example, the comprehensive model is

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log(x_1) + \gamma_4 \log(x_2) + u. \qquad \textbf{(9.8)}$$

We can first test H_0: $\gamma_3 = 0$, $\gamma_4 = 0$ as a test of (9.6). We can also test H_0: $\gamma_1 = 0$, $\gamma_2 = 0$ as a test of (9.7). This approach was suggested by Mizon and Richard (1986).

Another approach has been suggested by Davidson and MacKinnon (1981). They point out that, if (9.6) is true, then the fitted values from the *other* model, (9.7), should be insignificant in (9.6). Thus, to test (9.6), we first estimate model (9.7) by OLS to obtain the fitted values. Call these $\hat{\hat{y}}$. Then, the **Davidson-MacKinnon test** is based on the t statistic on $\hat{\hat{y}}$ in the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}} + error.$$

A signficant t statistic (against a two-sided alternative) is a rejection of (9.6).

Similarly, if $\hat{y}$ denotes the fitted values from estimating (9.6), the test of (9.7) is the t statistic on $\hat{y}$ in the model

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_1 \hat{y} + error;$$

a significant t statistic is evidence against (9.7). The same two tests can be used for testing any two nonnested models with the same dependent variable.

There are a few problems with nonnested testing. First, a clear winner need not emerge. Both models could be rejected or neither model could be rejected. In the latter case, we can use the adjusted R-squared to choose between them. If both models are rejected, more work needs to be done. However, it is important to know the practical consequences from using one form or the other: if the effects of key independent variables on y are not very different, then it does not really matter which model is used.

A second problem is that rejecting (9.6) using, say, the Davidson-MacKinnon test, does not mean that (9.7) is the correct model. Model (9.6) can be rejected for a variety of functional form misspecifications.

An even more difficult problem is obtaining nonnested tests when the competing models have different dependent variables. The leading case is y versus $\log(y)$. We saw in Chapter 6 that just obtaining goodness-of-fit measures that can be compared requires some care. Tests have been proposed to solve this problem, but they are beyond the scope of this text. [See Wooldridge (1994a) for a test that has a simple interpretation and is easy to implement.]

9.2 USING PROXY VARIABLES FOR UNOBSERVED EXPLANATORY VARIABLES

A more difficult problem arises when a model excludes a key variable, usually because of data inavailability. Consider a wage equation that explicitly recognizes that ability (*abil*) affects log(*wage*):

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u. \qquad \textbf{(9.9)}$$

This model shows explicitly that we want to hold ability fixed when measuring the return to *educ* and *exper*. If, say, *educ* is correlated with *abil*, then putting *abil* in the error term causes the OLS estimator of β_1 (and β_2) to be biased, a theme that has appeared repeatedly.

Our primary interest in equation (9.9) is in the slope parameters β_1 and β_2. We do not really care whether we get an unbiased or consistent estimator of the intercept β_0; as we will see shortly, this is not usually possible. Also, we can never hope to estimate β_3 because *abil* is not observed; in fact, we would not know how to interpret β_3 anyway, since ability is at best a vague concept.

How can we solve, or at least mitigate, the omitted variables bias in an equation like (9.9)? One possibility is to obtain a **proxy variable** for the omitted variable. Loosely speaking, a proxy variable is something that is related to the unobserved variable that we would like to control for in our analysis. In the wage equation, one possibility is to use the intelligence quotient, or IQ, as a proxy for ability. This *does not* require IQ to

be the same thing as ability; what we need is for IQ to be correlated with ability, something we clarify in the following discussion.

All of the key ideas can be illustrated in a model with three independent variables, two of which are observed:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u. \tag{9.10}$$

We assume that data are available on y, x_1, and x_2—in the wage example, these are log(*wage*), *educ*, and *exper*, respectively. The explanatory variable x_3^* is unobserved, but we have a proxy variable for x_3^*. Call the proxy variable x_3.

What do we require of x_3? At a minimum, it should have some relationship to x_3^*. This is captured by the simple regression equation

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3, \tag{9.11}$$

where v_3 is an error due to the fact that x_3^* and x_3 are not exactly related. The parameter δ_3 measures the relationship between x_3^* and x_3; typically, we think of x_3^* and x_3 as being positively related, so that $\delta_3 > 0$. If $\delta_3 = 0$, then x_3 is not a suitable proxy for x_3^*. The intercept δ_0 in (9.11), which can be positive or negative, simply allows x_3^* and x_3 to be measured on different scales. (For example, unobserved ability is certainly not required to have the same average value as IQ in the U.S. population.)

How can we use x_3 to get unbiased (or at least consistent) estimators of β_1 and β_2? The proposal is to pretend that x_3 and x_3^* are the same, so that we run the regression of

$$y \text{ on } x_1, x_2, x_3. \tag{9.12}$$

We call this the **plug-in solution to the omitted variables problem** because x_3 is just plugged in for x_3^* before we run OLS. If x_3 is truly related to x_3^*, this seems like a sensible thing. However, since x_3 and x_3^* are not the same, we should determine when this procedure does in fact give consistent estimators of β_1 and β_2.

The assumptions needed for the plug-in solution to provide consistent estimators of β_1 and β_2 can be broken down into assumptions about u and v_3:

(1) The error u is uncorrelated with x_1, x_2, and x_3^*, which is just the standard assumption in model (9.10). In addition, u is uncorrelated with x_3. This latter assumption just means that x_3 is irrelevant in the population model, once x_1, x_2, and x_3^* have been included. This is essentially true by definition, since x_3 is a proxy variable for x_3^*: it is x_3^* that directly affects y, not x_3. Thus, the assumption that u is uncorrelated with x_1, x_2, x_3^*, and x_3 is not very controversial. (Another way to state this assumption is that the expected value of u, given all these variables, is zero.)

(2) The error v_3 is uncorrelated with x_1, x_2, and x_3. Assuming that v_3 is uncorrelated with x_1 and x_2 requires x_3 to be a "good" proxy for x_3^*. This is easiest to see by writing the analog of these assumptions in terms of conditional expectations:

$$\mathrm{E}(x_3^*|x_1,x_2,x_3) = \mathrm{E}(x_3^*|x_3) = \delta_0 + \delta_3 x_3. \tag{9.13}$$

The first equality, which is the most important one, says that, once x_3 is controlled for, the expected value of x_3^* does not depend on x_1 or x_2. Alternatively, x_3^* has zero correlation with x_1 and x_2 once x_3 is partialled out.

In the wage equation (9.9), where IQ is the proxy for ability, condition (9.13) becomes

$$\mathrm{E}(abil|educ,exper,IQ) = \mathrm{E}(abil|IQ) = \delta_0 + \delta_3 IQ.$$

Thus, the average level of ability only changes with *IQ*, not with *educ* and *exper*. Is this reasonable? Maybe it is not exactly true, but it may be close to being true. It is certainly worth including *IQ* in the wage equation to see what happens to the estimated return to education.

We can easily see why the previous assumptions are enough for the plug-in solution to work. If we plug equation (9.11) into equation (9.10) and do simple algebra, we get

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3.$$

Call the composite error in this equation $e = u + \beta_3 v_3$; it depends on the error in the model of interest, (9.10), and the error in the proxy variable equation, v_3. Since u and v_3 both have zero mean and each is uncorrelated with x_1, x_2, and x_3, e also has zero mean and is uncorrelated with x_1, x_2, and x_3. Write this equation as

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e,$$

where $\alpha_0 = (\beta_0 + \beta_3\delta_0)$ is the new intercept and $\alpha_3 = \beta_3\delta_3$ is the slope parameter on the proxy variable x_3. As we alluded to earlier, when we run the regression in (9.12), we will not get unbiased estimators of β_0 and β_3; instead, we will get unbiased (or at least consistent) estimators of α_0, β_1, β_2, and α_3. The important thing is that we get good estimates of the parameters β_1 and β_2.

In many cases, the estimate of α_3 is actually more interesting than an estimate of β_3, anyway. For example, in the wage equation, α_3 measures the return to wage, given one more point on IQ score. Since the distribution of IQ in most populations is readily available, it is possible to see how large a ceteris paribus effect IQ has on wage.

EXAMPLE 9.3
(IQ as a Proxy for Ability)

The file WAGE2.RAW, from Blackburn and Neumark (1992), contains information on monthly earnings, education, several demographic variables, and IQ scores for 935 men in 1980. As a method to account for omitted ability bias, we add *IQ* to a standard log wage equation. The results are shown in Table 9.2.

Our primary interest is in what happens to the estimated return to education. Column (1) contains the estimates without using *IQ* as a proxy variable. The estimated return to education is 6.5%. If we think omitted ability is positively correlated with *educ*, then we assume that this estimate is too high. (More precisely, the average estimate across all random samples would be too high.) When IQ is added to the equation, the return to education falls to 5.4%, which corresponds with our prior beliefs about omitted ability bias.

Table 9.2

Dependent Variable: log(*wage*)

Independent Variables	(1)	(2)	(3)
educ	.065 (.006)	.054 (.007)	.018 (.041)
exper	.014 (.003)	.014 (.003)	.014 (.003)
tenure	.012 (.002)	.011 (.002)	.011 (.002)
married	.199 (.039)	.200 (.039)	.201 (.039)
south	−.091 (.026)	−.080 (.026)	−.080 (.026)
urban	.184 (.027)	.182 (.027)	.184 (.027)
black	−.188 (.038)	−.143 (.039)	−.147 (.040)
IQ	———	.0036 (.0010)	−.0009 (.0052)
educ·IQ	———	———	.00034 (.00038)
intercept	5.395 (.113)	5.176 (.128)	5.648 (.546)
Observations *R*-Squared	935 .253	935 .263	935 .263

The effect of IQ on socioeconomic outcomes has been recently documented in the controversial book, *The Bell Curve*, by Herrnstein and Murray (1994). Column (2) shows that IQ does have a statistically significant, positive effect on earnings, after controlling for several other factors. Everything else being equal, an increase of 10 IQ points is predicted to raise monthly earnings by 3.6%. The standard deviation of IQ in the U.S. population is 15, so a one standard deviation increase in IQ is associated with an elevation in earnings of 5.4%. This is identical to the predicted increase in wage due to another year of education. It is

clear from column (2) that education still has an important role in increasing earnings, even though the effect is not as large as originally estimated.

Some other interesting observations emerge from columns (1) and (2). Adding IQ to the equation only increases the *R*-squared from .253 to .263. Most of the variation in log(*wage*) is not explained by the factors in column (2). Also, adding IQ to the equation does not eliminate the estimated earnings difference between black and white men: a black man with the same IQ, education, experience, and so on as a white man is predicted to earn about 14.3% less, and the difference is very statistically significant.

Column (3) in Table 9.2 includes the interaction term *educ·IQ*. This allows for the possibility that *educ* and *abil* interact in determining log(*wage*). We might think that the return to education is higher for people with more ability, but this turns out not to be the case: the interaction term is not significant, and its addition makes *educ* and *IQ* individually insignificant while complicating the model. Therefore, the estimates in column (2) are preferred.

QUESTION 9.2

What do you conclude about the small and statistically insignificant coefficient on *educ* in column (3) of Table 9.2? (*Hint*: When *educ·IQ* is in the equation, what is the interpretation of the coefficient on *educ*?)

There is no reason to stop at a single proxy variable for ability in this example. The data set WAGE2.RAW also contains a score for each man on the *Knowledge of the World of Work* (KWW) test. This provides a different measure of ability, which can be used in place of IQ or along with IQ, to estimate the return to education (see Exercise 9.7).

It is easy to see how using a proxy variable can still lead to bias, if the proxy variable does not satisfy the preceding assumptions. Suppose that, instead of (9.11), the unobserved variable, x_3^*, is related to all of the observed variables by

$$x_3^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + v_3, \tag{9.14}$$

where v_3 has a zero mean and is uncorrelated with x_1, x_2, and x_3. Equation (9.11) assumes that δ_1 and δ_2 are both zero. By plugging equation (9.14) into (9.10), we get

$$y = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)x_1 + (\beta_2 + \beta_3\delta_2)x_2 + \beta_3\delta_3 x_3 + u + \beta_3 v_3, \tag{9.15}$$

from which it follows that $\text{plim}(\hat{\beta}_1) = \beta_1 + \beta_3\delta_1$ and $\text{plim}(\hat{\beta}_2) = \beta_2 + \beta_3\delta_2$. [This follows because the error in (9.15), $u + \beta_3 v_3$, has zero mean and is uncorrelated with x_1, x_2, and x_3.] In the previous example where $x_1 = educ$ and $x_3^* = abil$, $\beta_3 > 0$, so there is a positive bias (inconsistency), if *abil* has a positive partial correlation with *educ* ($\delta_1 > 0$). Thus, we could still be getting an upward bias in the return to education, using *IQ* as a proxy for *abil*, if *IQ* is not a good proxy. But we can reasonably hope that this bias is smaller than if we ignored the problem of omitted ability entirely.

Proxy variables can come in the form of binary information as well. In Example 7.9 [see equation (7.15)], we discussed Krueger's (1993) estimates of the return to using a

computer on the job. Krueger also included a binary variable indicating whether the worker uses a computer at home (as well as an interaction term between computer usage at work and at home). His primary reason for including computer usage at home in the equation was to proxy for unobserved "technical ability" that could affect wage directly and be related to computer usage at work.

Using Lagged Dependent Variables as Proxy Variables

In some applications, like the earlier wage example, we have at least a vague idea about which unobserved factor we would like to control for. This facilitates choosing proxy variables. In other applications, we suspect that one or more of the independent variables is correlated with an omitted variable, but we have no idea how to obtain a proxy for that omitted variable. In such cases, we can include, as a control, the value of the dependent variable from an earlier time period. This is especially useful for policy analysis.

Using a **lagged dependent variable** in a cross-sectional equation increases the data requirements, but it also provides a simple way to account for historical factors that cause *current* differences in the dependent variable that are difficult to account for in other ways. For example, some cities have had high crime rates in the past. Many of the same unobserved factors contribute to both high current and past crime rates. Likewise, some universities are traditionally better in academics than other universities. Inertial effects are also captured by putting in lags of y.

Consider a simple equation to explain city crime rates:

$$crime = \beta_0 + \beta_1 unem + \beta_2 expend + \beta_3 crime_{-1} + u, \qquad \textbf{(9.16)}$$

where *crime* is a measure of per capita crime, *unem* is the city unemployment rate, *expend* is per capita spending on law enforcement, and $crime_{-1}$ indicates the crime rate measured in some earlier year (this could be the past year or several years ago). We are interested in the effects of *unem* on *crime*, as well as of law enforcement expenditures on crime.

What is the purpose of including $crime_{-1}$ in the equation? Certainly we expect that $\beta_3 > 0$, since crime has inertia. But the main reason for putting this in the equation is that cities with high historical crime rates may spend more on crime prevention. Thus, factors unobserved to us (the econometricians) that affect *crime* are likely to be correlated with *expend* (and *unem*). If we use a pure cross-sectional analysis, we are unlikely to get an unbiased estimator of the causal effect of law enforcement expenditures on crime. But, by including $crime_{-1}$ in the equation, we can at least do the following experiment: if two cities have the same previous crime rate and current unemployment rate, then β_2 measures the effect of another dollar of law enforcement on crime.

EXAMPLE 9.4
(City Crime Rates)

We estimate a constant elasticity version of the crime model in equation (9.16) (*unem*, since it is a percent, is left in level form). The data in CRIME2.RAW are from 46 cities for the year

Table 9.3

Dependent Variable: log($crmrte_{87}$)

Independent Variables	(1)	(2)
$unem_{87}$	−.029 (.032)	.009 (.020)
log($lawexpc_{87}$)	.203 (.173)	−.140 (.109)
log($crmrte_{82}$)	—	1.194 (.132)
intercept	3.34 (1.25)	.076 (.821)
Observations *R*-Squared	46 .057	46 .680

1987. The crime rate is also available for 1982, and we use that as an additional independent variable in trying to control for city unobservables that affect crime and may be correlated with current law enforcement expenditures. Table 9.3 contains the results.

Without the lagged crime rate in the equation, the effects of the unemployment rate and expenditures on law enforcement are counterintuitive; neither is statistically significant, although the t statistic on log($lawexpc_{87}$) is 1.17. One possibility is that increased law enforcement expenditures improve reporting conventions, and so more crimes are *reported*. But it is also likely that cities with high recent crime rates spend more on law enforcement.

Adding the log of the crime rate from five years earlier has a large effect on the expenditures coefficient. The elasticity of the crime rate with respect to expenditures becomes −.14, with $t = -1.28$. This is not strongly significant, but it suggests that a more sophisticated model with more cities in the sample could produce significant results.

Not surprisingly, the current crime rate is strongly related to the past crime rate. The estimate indicates that if the crime rate in 1982 was 1% higher, then the crime rate in 1987 is predicted to be about 1.19% higher. We cannot reject the hypothesis that the elasticity of current crime with respect to past crime is unity [$t = (1.194 - 1)/.132 \approx 1.47$]. Adding the past crime rate increases the explanatory power of the regression markedly, but this is no surprise. The primary reason for including the lagged crime rate is to obtain a better estimate of the ceteris paribus effect of log($lawexpc_{87}$) on log($crmrte_{87}$).

The practice of putting in a lagged y as a general way of controlling for unobserved variables is hardly perfect. But it can aid in getting a better estimate of the effects of policy variables on various outcomes.

Adding a lagged value of y is not the only way to use two years of data to control for omitted factors. When we discuss panel data methods in Chapters 13 and 14, we will cover other ways to use repeated data on the same cross-sectional units at different points in time.

9.3 PROPERTIES OF OLS UNDER MEASUREMENT ERROR

Sometimes, in economic applications, we cannot collect data on the variable that truly affects economic behavior. A good example is the marginal income tax rate facing a family that is trying to choose how much to contribute to charity in a given year. The marginal rate may be hard to obtain or summarize as a single number for all income levels. Instead, we might compute the average tax rate based on total income and tax payments.

When we use an imprecise measure of an economic variable in a regression model, then our model contains measurement error. In this section, we derive the consequences of measurement error for ordinary least squares estimation. OLS will be consistent under certain assumptions, but there are others under which it is inconsistent. In some of these cases, we can derive the size of the asymptotic bias.

As we will see, the measurement error problem has a similar statistical structure to the omitted variable-proxy variable problem discussed in the previous section, but they are conceptually different. In the proxy variable case, we are looking for a variable that is somehow associated with the unobserved variable. In the measurement error case, the variable that we do not observe has a well-defined, quantitative meaning (such as a marginal tax rate or annual income), but our recorded measures of it may contain error. For example, reported annual income is a measure of actual annual income, whereas IQ score is a proxy for ability.

Another important difference between the proxy variable and measurement error problems is that, in the latter case, often the mismeasured independent variable is the one of primary interest. In the proxy variable case, the partial effect of the omitted variable is rarely of central interest: we are usually concerned with the effects of the other independent variables.

Before we consider details, we should remember that measurement error is an issue only when the variables for which the econometrician can collect data differ from the variables that influence decisions by individuals, families, firms, and so on.

Measurement Error in the Dependent Variable

We begin with the case where only the dependent variable is measured with error. Let y^* denote the variable (in the population, as always) that we would like to explain. For example, y^* could be annual family savings. The regression model has the usual form

$$y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad \textbf{(9.17)}$$

and we assume it satisfies the Gauss-Markov assumptions. We let y represent the observable measure of y^*. In the savings case, y is reported annual savings. Unfor-

tunately, families are not perfect in their reporting of annual family savings; it is easy to leave out categories or to overestimate the amount contributed to a fund. Generally, we can expect y and y^* to differ, at least for some subset of families in the population.

The measurement error (in the population) is defined as the difference between the observed value and the actual value:

$$e_0 = y - y^*. \tag{9.18}$$

For a random draw i from the population, we can write $e_{i0} = y_i - y_i^*$, but the important thing is how the measurement error in the population is related to other factors. To obtain an estimable model, we write $y^* = y - e_0$, plug this into equation (9.17), and rearrange:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0. \tag{9.19}$$

The error term in equation (9.19) is $u + e_0$. Since $y, x_1, x_2, \ldots, x_k$ are observed, we can estimate this model by OLS. In effect, we just ignore the fact that y is an imperfect measure of y^* and proceed as usual.

When does OLS with y in place of y^* produce consistent estimators of the β_j? Since the original model (9.17) satisfies the Gauss-Markov assumptions, u has zero mean and is uncorrelated with each x_j. It is only natural to assume that the measurement error has zero mean; if it does not, then we simply get a biased estimator of the intercept, β_0, which is rarely a cause for concern. Of much more importance is our assumption about the relationship between the measurement error, e_0, and the explanatory variables, x_j. The usual assumption is that the measurement error in y is statistically independent of each explanatory variable. If this is true, then the OLS estimators from (9.19) are unbiased and consistent. Further, the usual OLS inference procedures (t, F, and LM statistics) are valid.

If e_0 and u are uncorrelated, as is usually assumed, then $\text{Var}(u + e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$. This means that measurement error in the dependent variable results in a larger error variance than when no error occurs; this, of course, results in larger variances of the OLS estimators. This is to be expected, and there is nothing we can do about it (except collect better data). The bottom line is that, if the measurement error is uncorrelated with the independent variables, then OLS estimation has good properties.

EXAMPLE 9.5
(Savings Function with Measurement Error)

Consider a savings function

$$sav^* = \beta_0 + \beta_1 inc + \beta_2 size + \beta_3 educ + \beta_4 age + u,$$

but where actual savings (*sav**) may deviate from reported savings (*sav*). The question is whether the size of the measurement error in *sav* is systematically related to the other variables. It might be reasonable to assume that the measurement error is not correlated with *inc*, *size*, *educ*, and *age*. On the other hand, we might think that families with higher incomes, or more education, report their savings more accurately. We can never know

whether the measurement error is correlated with *inc* or *educ*, unless we can collect data on *sav**; then the measurement error can be computed for each observation as $e_{i0} = sav_i - sav_i^*$.

When the dependent variable is in logarithmic form, so that $\log(y^*)$ is the dependent variable, it is natural for the measurement error equation to be of the form

$$\log(y) = \log(y^*) + e_0. \tag{9.20}$$

This follows from a **multiplicative measurement error** for y: $y = y^*a_0$, where $a_0 > 0$ and $e_0 = \log(a_0)$.

EXAMPLE 9.6
(Measurement Error in Scrap Rates)

In Section 7.6, we discussed an example where we wanted to determine whether job training grants reduce the scrap rate in manufacturing firms. We certainly might think the scrap rate reported by firms is measured with error. (In fact, most firms in the sample do not even report a scrap rate.) In a simple regression framework, this is captured by

$$\log(scrap^*) = \beta_0 + \beta_1 grant + u,$$

where *scrap** is the true scrap rate and *grant* is the dummy variable indicating whether a firm received a grant. The measurement error equation is

$$\log(scrap) = \log(scrap^*) + e_0.$$

Is the measurement error, e_0, independent of whether the firm receives a grant? A cynical person might think that a firm receiving a grant is more likely to underreport its scrap rate in order to make the grant look effective. If this happens, then, in the estimable equation,

$$\log(scrap) = \beta_0 + \beta_1 grant + u + e_0,$$

the error $u + e_0$ is negatively correlated with *grant*. This would produce a downward bias in β_1, which would tend to make the training program look more effective than it actually was. (Remember, a more negative β_1 means the program was more effective, since increased worker productivity is associated with a lower scrap rate.)

The bottom line of this subsection is that measurement error in the dependent variable *can* cause biases in OLS if it is systematically related to one or more of the explanatory variables. If the measurement error is just a random reporting error that is independent of the explanatory variables, as is often assumed, then OLS is perfectly appropriate.

Measurement Error in an Explanatory Variable

Traditionally, measurement error in an explanatory variable has been considered a much more important problem than measurement error in the dependent variable. In this subsection, we will see why this is the case.

We begin with the simple regression model

$$y = \beta_0 + \beta_1 x_1^* + u, \tag{9.21}$$

and we assume that this satisfies at least the first four Gauss-Markov assumptions. This means that estimation of (9.21) by OLS would produce unbiased and consistent estimators of β_0 and β_1. The problem is that x_1^* is not observed. Instead, we have a measure of x_1^*, call it x_1. For example, x_1^* could be actual income, and x_1 could be reported income.

The measurement error in the population is simply

$$e_1 = x_1 - x_1^*, \tag{9.22}$$

and this can be positive, negative, or zero. We assume that the *average* measurement error in the population is zero: $E(e_1) = 0$. This is natural, and, in any case, it does not affect the important conclusions that follow. A maintained assumption in what follows is that u is uncorrelated with x_1^* *and* x_1. In conditional expectation terms, we can write this as $E(y|x_1^*,x_1) = E(y|x_1^*)$, which just says that x_1 does not affect y after x_1^* has been controlled for. We used the same assumption in the proxy variable case, and it is not controversial; it holds almost by definition.

We want to know the properties of OLS if we simply replace x_1^* with x_1 and run the regression of y on x_1. They depend crucially on the assumptions we make about the measurement error. Two assumptions have been the focus in econometrics literature, and they both represent polar extremes. The first assumption is that e_1 is uncorrelated with the *observed* measure, x_1:

$$\text{Cov}(x_1,e_1) = 0. \tag{9.23}$$

From the relationship in (9.22), if assumption (9.23) is true, then e_1 must be correlated with the unobserved variable x_1^*. To determine the properties of OLS in this case, we write $x_1^* = x_1 - e_1$ and plug this into equation (9.21):

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1). \tag{9.24}$$

Since we have assumed that u and e_1 both have zero mean and are uncorrelated with x_1, $u - \beta_1 e_1$ has zero mean and is uncorrelated with x_1. It follows that OLS estimation with x_1 in place of x_1^* produces a consistent estimator of β_1 (and also β_0). Since u is uncorrelated with e_1, the variance of the error in (9.23) is $\text{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$. Thus, except when $\beta_1 = 0$, measurement error increases the error variance. But this does not affect any of the OLS properties (except that the variances of the $\hat{\beta}_j$ will be larger than if we observe x_1^* directly).

The assumption that e_1 is uncorrelated with x_1 is analogous to the proxy variable assumption we made in Section 9.2. Since this assumption implies that OLS has all of its nice properties, this is not usually what econometricians have in mind when they refer to measurement error in an explanatory variable. The **classical errors-in-variables** (**CEV**) assumption is that the measurement error is uncorrelated with the *unobserved* explanatory variable:

$$\text{Cov}(x_1^*,e_1) = 0. \tag{9.25}$$

This assumption comes from writing the observed measure as the sum of the true explanatory variable and the measurement error,

$$x_1 = x_1^* + e_1,$$

and then assuming the two components of x_1 are uncorrelated. (This has nothing to do with assumptions about u; we always maintain that u is uncorrelated with x_1^* and x_1, and therefore with e_1).

If assumption (9.25) holds, then x_1 and e_1 *must* be correlated:

$$\text{Cov}(x_1,e_1) = \text{E}(x_1e_1) = \text{E}(x_1^*e_1) + \text{E}(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2. \tag{9.26}$$

Thus, the covariance between x_1 and e_1 is equal to the variance of the measurement error under the CEV assumption.

Referring to equation (9.24), we can see that correlation between x_1 and e_1 is going to cause problems. Because u and x_1 are uncorrelated, the covariance between x_1 and the composite error $u - \beta_1 e_1$ is

$$\text{Cov}(x_1,u - \beta_1 e_1) = -\beta_1 \text{Cov}(x_1,e_1) = -\beta_1 \sigma_{e_1}^2.$$

Thus, in the CEV case, the OLS regression of y on x_1 gives a biased and inconsistent estimator.

Using the asymptotic results in Chapter 5, we can determine the amount of inconsistency in OLS. The probability limit of $\hat{\beta}_1$ is β_1 plus the ratio of the covariance between x_1 and $u - \beta_1 e_1$ and the variance of x_1:

$$\begin{aligned} \text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\text{Cov}(x_1,u - \beta_1 e_1)}{\text{Var}(x_1)} \\ &= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} = \beta_1\left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right) \\ &= \beta_1\left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}\right), \end{aligned} \tag{9.27}$$

where we have used the fact that $\text{Var}(x_1) = \text{Var}(x_1^*) + \text{Var}(e_1)$.

Equation (9.27) is very interesting. The term multiplying β_1, which is the ratio $\text{Var}(x_1^*)/\text{Var}(x_1)$, is always less than one [an implication of the CEV assumption (9.25)]. Thus, $\text{plim}(\hat{\beta}_1)$ is always closer to zero than is β_1. This is called the **attenuation bias**

in OLS due to classical errors-in-variables: on average (or in large samples), the estimated OLS effect will be *attenuated*. In particular, if β_1 is positive, $\hat{\beta}_1$ will tend to underestimate β_1. This is an important conclusion, but it relies on the CEV setup.

If the variance of x_1^* is large, relative to the variance in the measurement error, then the inconsistency in OLS will be small. This is because $\text{Var}(x_1^*)/\text{Var}(x_1)$ will be close to unity, when $\sigma_{x_1^*}^2/\sigma_{e_1}^2$ is large. Therefore, depending on how much variation there is in x_1^*, relative to e_1, measurement error need not cause large biases.

Things are more complicated when we add more explanatory variables. For illustration, consider the model

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + \beta_3 x_3 + u, \tag{9.28}$$

where the first of the three explanatory variables is measured with error. We make the natural assumption that u is uncorrelated with x_1^*, x_2, x_3, and x_1. Again, the crucial assumption concerns the measurement error e_1. In almost all cases, e_1 is assumed to be uncorrelated with x_2 and x_3—the explanatory variables not measured with error. The key issue is whether e_1 is uncorrelated with x_1. If it is, then the OLS regression of y on x_1, x_2, and x_3 produces consistent estimators. This is easily seen by writing

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u - \beta_1 e_1, \tag{9.29}$$

where u and e_1 are both uncorrelated with all the explanatory variables.

Under the CEV assumption (9.25), OLS will be biased and inconsistent, because e_1 is correlated with x_1 in equation (9.29). Remember, this means that, in general, *all* OLS estimators will be biased, not just $\hat{\beta}_1$. What about the attenuation bias derived in equation (9.27)? It turns out that there is still an attentuation bias for estimating β_1: It can be shown that

$$\text{plim}(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma_{r_1^*}^2}{\sigma_{r_1^*}^2 + \sigma_{e_1}^2} \right), \tag{9.30}$$

where r_1^* is the population error in the equation $x_1^* = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + r_1^*$. Formula (9.30) also works in the general k variable case when x_1 is the only mismeasured variable.

Things are less clear-cut for estimating the β_j on the variables not measured with error. In the special case that x_1^* is uncorrelated with x_2 and x_3, $\hat{\beta}_2$ and $\hat{\beta}_3$ are consistent. But this is rare in practice. Generally, measurement error in a single variable causes inconsistency in all estimators. Unfortunately, the sizes, and even the directions of the biases, are not easily derived.

EXAMPLE 9.7
(GPA Equation with Measurement Error)

Consider the problem of estimating the effect of family income on college grade point average, after controlling for *hsGPA* and *SAT*. It could be that, while family income is important

for performance before college, it has no direct effect on college performance. To test this, we might postulate the model

$$colGPA = \beta_0 + \beta_1 faminc^* + \beta_2 hsGPA + \beta_3 SAT + u,$$

where *faminc** is actual annual family income. (This might appear in logarithmic form, but for the sake of illustration we leave it in level form.) Precise data on *colGPA*, *hsGPA*, and *SAT* are relatively easy to obtain. But family income, especially as reported by students, could be easily mismeasured. If $faminc = faminc^* + e_1$ and the CEV assumptions hold, then using reported family income in place of actual family income will bias the OLS estimator of β_1 towards zero. One consequence of this is that a test of H_0: $\beta_1 = 0$ will have less chance of detecting $\beta_1 > 0$.

Of course, measurement error can be present in more than one explanatory variable, or in some explanatory variables and the dependent variable. As we discussed earlier, any measurement error in the dependent variable is usually assumed to be uncorrelated with all the explanatory variables, whether it is observed or not. Deriving the bias in the OLS estimators under extensions of the CEV assumptions is complicated and does not lead to clear results.

In some cases, it is clear that the CEV assumption in (9.25) cannot be true. Consider a variant on Example 9.7:

$$colGPA = \beta_0 + \beta_1 smoked^* + \beta_2 hsGPA + \beta_3 SAT + u,$$

where *smoked** is the number of times a student smoked marijuana in the last 30 days. The variable *smoked* is the answer to the question: On how many separate occasions did you smoke marijuana in the last 30 days? Suppose we postulate the standard measurement error model

$$smoked = smoked^* + e_1.$$

Even if we assume that students try to report the truth, the CEV assumption is unlikely to hold. People who do not smoke marijuana at all—so that $smoked^* = 0$—are likely to report $smoked = 0$, so the measurement error is probably zero for students who never smoke marijuana. When $smoked^* > 0$, it is much more likely that the student miscounts how many times he or she smoked marijuana in the last 30 days. This means that the measurement error e_1 and the *actual* number of times smoked, *smoked**, are correlated, which violates the CEV assumption in (9.25). Unfortunately, deriving the implications of measurement error that do not satisfy (9.23) or (9.25) is difficult and beyond the scope of this text.

QUESTION 9.3

Let *educ** be actual amount of schooling, measured in years (which can be a noninteger) and let *educ* be reported highest grade completed. Do you think *educ* and *educ** are related by the classical errors-in-variables model?

Before leaving this section, we emphasize that, a priori, the CEV assumption (9.25) is no better or worse than assumption (9.23), which implies that OLS is consistent. The truth is probably somewhere in between, and if e_1 is correlated with both x_1^* and x_1, OLS is inconsistent. This raises

an important question: Must we live with inconsistent estimators under classical errors-in-variables, or other kinds of measurement error that are correlated with x_1? Fortunately, the answer is no. Chapter 15 shows how, under certain assumptions, the parameters can be consistently estimated in the presence of general measurement error. We postpone this discussion until later, because it requires us to leave the realm of OLS estimation.

9.4 MISSING DATA, NONRANDOM SAMPLES, AND OUTLYING OBSERVATIONS

The measurement error problem discussed in the previous section can be viewed as a data problem: we cannot obtain data on the variables of interest. Further, under the classical errors-in-variables model, the composite error term is correlated with the mismeasured independent variable, violating the Gauss-Markov assumptions.

Another data problem we discussed frequently in earlier chapters is multicollinearity among the explanatory variables. Remember that correlation among the explanatory variables does not violate any assumptions. When two independent variables are highly correlated, it can be difficult to estimate the partial effect of each. But this is properly reflected in the usual OLS statistics.

In this section, we provide an introduction to data problems that can violate the random sampling assumption, MLR.2. We can isolate cases where nonrandom sampling has no practical effect on OLS. In other cases, nonrandom sampling causes the OLS estimators to be biased and inconsistent. A more complete treatment that establishes several of the claims made here is given in Chapter 17.

Missing Data

The **missing data** problem can arise in a variety of forms. Often, we collect a random sample of people, schools, cities, and so on, and then discover later that information is missing on some key variables for several units in the sample. For example, in the data set BWGHT.RAW, 197 of the 1,388 observations have no information on either mother's education, father's education, or both. In the data set on median starting law school salaries, LAWSCH85.RAW, six of the 156 schools have no reported information on median LSAT scores for the entering class; other variables are also missing for some of the law schools.

If data are missing for an observation on either the dependent variable or one of the independent variables, then the observation cannot be used in a standard multiple regression analysis. In fact, provided missing data have been properly indicated, all modern regression packages keep track of missing data and simply ignore observations when computing a regression. We saw this explicitly in the birth weight Example 4.9, when 197 observations were dropped due to missing information on parents' education.

Other than reducing the sample size available for a regression, are there any *statistical* consequences of missing data? It depends on why the data are missing. If the data are missing at random, then the size of the random sample available from the population is simply reduced. While this makes the estimators less precise, it does not introduce any bias: the random sampling assumption, MLR.2, still holds. There are ways to

use the information on observations where only some variables are missing, but this is not often done in practice. The improvement in the estimators is usually slight, while the methods are somewhat complicated. In most cases, we just ignore the observations that have missing information.

Nonrandom Samples

Missing data is more problematic when it results in a **nonrandom sample** from the population. For example, in the birth weight data set, what if the probability that education is missing is higher for those people with lower than average levels of education? Or, in Section 9.2, we used a wage data set that included IQ scores. This data set was constructed by omitting several people from the sample for whom IQ scores were not available. If obtaining an IQ score is easier for those with higher IQs, the sample is not representative of the population. The random sampling assumption MLR.2 is violated, and we must worry about these consequences for OLS estimation.

Certain types of nonrandom sampling do *not* cause bias or inconsistency in OLS. Under the Gauss-Markov assumptions (but without MLR.2), it turns out that the sample can be chosen on the basis of the *independent* variables without causing any statistical problems. This is called *sample selection based on the independent variables*, and it is an example of **exogenous sample selection**. To illustrate, suppose that we are estimating a saving function, where annual saving depends on income, age, family size, and perhaps some other factors. A simple model is

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u. \qquad \textbf{(9.31)}$$

Suppose that our data set was based on a survey of people over 35 years of age, thereby leaving us with a nonrandom sample of all adults. While this is not ideal, we can still get unbiased and consistent estimators of the parameters in the population model (9.31), using the nonrandom sample. We will not show this formally here, but the reason OLS on the nonrandom sample is unbiased is that the regression function $E(saving|income,age,size)$ is the same for any subset of the population described by *income*, *age*, or *size*. Provided there is enough variation in the independent variables in the sub-population, selection on the basis of the independent variables is not a serious problem, other than that it results in inefficient estimators.

In the IQ example just mentioned, things are not so clear-cut, because no fixed rule based on IQ is used to include someone in the sample. Rather, the *probability* of being in the sample increases with IQ. If the other factors determining selection into the sample are independent of the error term in the wage equation, then we have another case of exogenous sample selection, and OLS using the selected sample will have all of its desirable properties under the other Gauss-Markov assumptions.

Things are much different when selection is based on the dependent variable, *y*, which is called *sample selection based on the dependent variable* and is an example of **endogenous sample selection**. If the sample is based on whether the dependent variable is above or below a given value, bias always occurs in OLS in estimating the population model. For example, suppose we wish to estimate the relationship between individual wealth and several other factors in the population of all adults:

$$wealth = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u. \quad \textbf{(9.32)}$$

Suppose that only people with wealth below $75,000 dollars are included in the sample. This is a nonrandom sample from the population of interest, and it is based on the value of the dependent variable. Using a sample on people with wealth below $75,000 will result in biased and inconsistent estimators of the parameters in (9.32). Briefly, the reason is that the population regression E(*wealth*|*educ*, *exper*, *age*) is not the same as the expected value conditional on *wealth* being less than $75,000.

Other sample selection issues are more subtle. For instance, in several previous examples, we have estimated the effects of various variables, particularly education and experience, on hourly wage. The data set WAGE1.RAW that we have used throughout is essentially a random sample of *working* individuals. Labor economists are often interested in estimating the effect of, say, education on the wage *offer*. The idea is this: Every person of working age faces an hourly wage offer, and he or she can either work at that wage or not work. For someone who does work, the wage offer is just the wage earned. For people who do not work, we usually cannot observe the wage offer. Now, since the wage offer equation

$$\log(wage^o) = \beta_0 + \beta_1 educ + \beta_2 exper + u, \quad \textbf{(9.33)}$$

represents the population of all working age people, we cannot estimate it using a random sample from this population; instead, we have data on the wage offer only for working people (although we can get data on *educ* and *exper* for nonworking people). If we use a random sample on working people to estimate (9.33), will we get unbiased estimators? This case is not clear-cut. Since the sample is selected based on someone's decision to work (as opposed to the size of the wage offer), this is not like the previous case. However, since the decision to work might be related to unobserved factors that affect the wage offer, selection might be endogenous, and this can result in a sample selection bias in the OLS estimators. We will cover methods that can be used to test and correct for sample selection bias in Chapter 17.

QUESTION 9.4

Suppose we are interested in the effects of campaign expenditures by incumbents on voter support. Some incumbents choose not to run for reelection. If we can only collect voting and spending outcomes on incumbents that actually do run, is there likely to be endogenous sample selection?

Outlying Observations

In some applications, especially, but not only, with small data sets, the OLS estimates are influenced by one or several observations. Such observations are called **outliers** or **influential observations**. Loosely speaking, an observation is an outlier if dropping it from a regression analysis makes the OLS estimates change by a practically "large" amount.

OLS is susceptible to outlying observations because it minimizes the sum of squared residuals: large residuals (positive or negative) receive a lot of weight in the least squares minimization problem. If the estimates change by a practically large amount when we slightly modify our sample, we should be concerned.

When statisticians and econometricians study the problem of outliers theoretically, sometimes the data are viewed as being from a random sample from a given population—albeit with an unusual distribution that can result in extreme values—and sometimes the outliers are assumed to come from a different population. From a practical perspective, outlying observations can occur for two reasons. The easiest case to deal with is when a mistake has been made in entering the data. Adding extra zeros to a number or misplacing a decimal point can throw off the OLS estimates, especially in small sample sizes. It is always a good idea to compute summary statistics, especially minimums and maximums, in order to catch mistakes in data entry. Unfortunately, incorrect entries are not always obvious.

Outliers can also arise when sampling from a small population if one or several members of the population are very different in some relevant aspect from the rest of the population. The decision to keep or drop such observations in a regression analysis can be a difficult one, and the statistical properties of the resulting estimators are complicated. Outlying observations can provide important information by increasing the variation in the explanatory variables (which reduces standard errors). But OLS results should probably be reported with and without outlying observations in cases where one or several data points substantially change the results.

EXAMPLE 9.8
(R&D Intensity and Firm Size)

Suppose that R&D expenditures as a percentage of sales (*rdintens*) are related to *sales* (in millions) and profits as a percentage of sales (*profmarg*):

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 profmarg + u. \tag{9.34}$$

The OLS equation using data on 32 chemical companies in RDCHEM.RAW is

$$\widehat{rdintens} = 2.625 + .000053\ sales + .0446\ profmarg$$
$$(0.586)\quad (.000044)\qquad (.0462)$$
$$n = 32, R^2 = .0761, \bar{R}^2 = .0124.$$

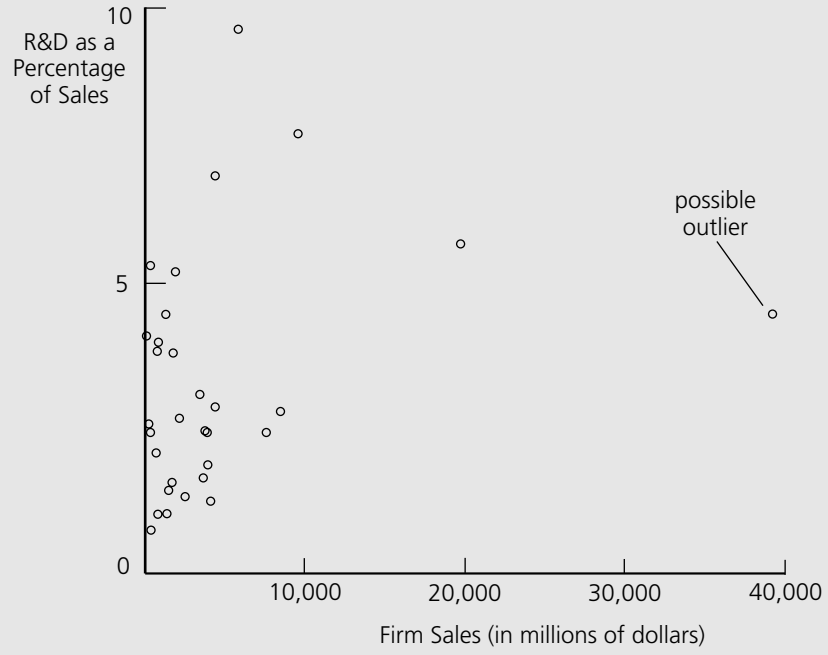
Neither *sales* nor *profmarg* is statistically significant at even the 10% level in this regression.

Of the 32 firms, 31 have annual sales less than $20 billion. One firm has annual sales of almost $40 billion. Figure 9.1 shows how far this firm is from the rest of the sample. In terms of sales, this firm is over twice as large as every other firm, so it might be a good idea to estimate the model without it. When we do this, we obtain

$$\widehat{rdintens} = 2.297 + .000186\ sales + .0478\ profmarg$$
$$(0.592)\quad (.000084)\qquad (.0445)$$
$$n = 31, R^2 = .1728, \bar{R}^2 = .1137.$$

If the largest firm is dropped from the regression, the coefficient on *sales* more than triples, and it now has a *t* statistic over two. Using the sample of smaller firms, we would conclude that there is a statistically significant positive effect between R&D intensity and firm size. The profit margin is still not significant, and its coefficient has not changed by much.

Figure 9.1
Scatterplot of R&D intensity against firm sales.



Sometimes outliers are defined by the size of the residual in an OLS regression where all of the observations are used. This is *not* a good idea. In the previous example, using all firms in the regression, a firm with sales of just under \$4.6 billion had the largest residual by far (about 6.37). The residual for the largest firm was -1.62, which is less than one estimated standard deviation from zero ($\hat{\sigma} = 1.82$). Dropping the observation with the largest residual does not change the results much at all.

Certain functional forms are less sensitive to outlying observations. In Section 6.2, we mentioned that, for most economic variables, the logarithmic transformation significantly narrows the range of the data and also yields functional forms—such as constant elasticity models—that can explain a broader range of data.

EXAMPLE 9.9
(R&D Intensity)

We can test whether R&D intensity increases with firm size by starting with the model

$$rd = sales^{\beta_1}\exp(\beta_0 + \beta_2 profmarg + u). \qquad \textbf{(9.35)}$$

Then, holding other factors fixed, R&D intensity increases with *sales* if and only if $\beta_1 > 1$. Taking the log of (9.35) gives

$$\log(rd) = \beta_0 + \beta_1 \log(sales) + \beta_2 profmarg + u. \tag{9.36}$$

When we use all 32 firms, the regression equation is

$$\widehat{\log(rd)} = \underset{(0.468)}{-4.378} + \underset{(0.062)}{1.084}\log(sales) + \underset{(.0128)}{.0217}\, profmarg,$$
$$n = 32, R^2 = .9180, \bar{R}^2 = .9123,$$

while dropping the largest firm gives

$$\widehat{\log(rd)} = \underset{(0.511)}{-4.404} + \underset{(0.067)}{1.088}\log(sales) + \underset{(.0130)}{.0218}\, profmarg,$$
$$n = 31, R^2 = .9037, \bar{R}^2 = .8968.$$

Practically, these results are the same. In neither case do we reject the null H_0: $\beta_1 = 1$ against H_1: $\beta_1 > 1$ (Why?).

In some cases, certain observations are suspected at the outset of being fundamentally different from the rest of the sample. This often happens when we use data at very aggregated levels, such as the city, county, or state level. The following is an example.

EXAMPLE 9.10
(State Infant Mortality Rates)

Data on infant mortality, per capita income, and measures of health care can be obtained at the state level from the *Statistical Abstract of the United States*. We will provide a fairly simple analysis here just to illustrate the effect of outliers. The data are for the year 1990, and we have all 50 states in the United States, plus the District of Columbia (D.C.). The variable *infmort* is number of deaths within the first year per 1,000 live births, *pcinc* is per capita income, *physic* is physicians per 100,000 members of the civilian population, and *popul* is the population (in thousands). We include all independent variables in logarithmic form:

$$\widehat{infmort} = \underset{(20.43)}{33.86} - \underset{(2.60)}{4.68}\log(pcinc) + \underset{(1.51)}{4.15}\log(physic) - \underset{(.287)}{.088}\log(popul)$$
$$n = 51, R^2 = .139, \bar{R}^2 = .084. \tag{9.37}$$

Higher per capita income is estimated to lower infant mortality, an expected result. But more physicians per capita is associated with *higher* infant mortality rates, something that is counterintuitive. Infant mortality rates do not appear to be related to population size.

The District of Columbia is unusual in that it has pockets of extreme poverty and great wealth in a small area. In fact, the infant mortality rate for D.C. in 1990 was 20.7, compared with 12.4 for the next highest state. It also has 615 physicians per 100,000 of the civilian population, compared with 337 for the the next highest state. The high number of physicians coupled with the high infant mortality rate in D.C. could certainly influence the results. If we drop D.C. from the regression, we obtain

$$\widehat{infmort} = \underset{(12.42)}{23.95} - \underset{(1.64)}{.57}\log(pcinc) - \underset{(1.19)}{2.74}\log(physic) + \underset{(.191)}{.629}\log(popul) \quad \textbf{(9.38)}$$
$$n = 50,\ R^2 = .273,\ \bar{R}^2 = .226.$$

We now find that more physicians per capita lowers infant mortality, and the estimate is statistically different from zero at the 5% level. The effect of per capita income has fallen sharply and is no longer statistically significant. In equation (9.38), infant mortality rates are higher in more populous states, and the relationship is very statistically significant. Also, much more variation in *infmort* is explained when D.C. is dropped from the regression. Clearly, D.C. had substantial influence on the initial estimates, and we would probably leave it out of any further analysis.

Rather than having to personally determine the influence of certain observations, it is sometimes useful to have statistics that can detect such influential observations. These statistics do exist, but they are beyond the scope of this text. [See, for example, Belsley, Kuh, and Welsch (1980).]

Before ending this section, we mention another approach to dealing with influential observations. Rather than trying to find outlying observations in the data before applying least squares, we can use an estimation method that is less sensitive to outliers than OLS. This obviates the need to explicitly search for outliers before estimation. One such method is called *least absolute deviations*, or LAD. The LAD estimator minimizes the sum of the absolute deviation of the residuals, rather than the sum of squared residuals. Compared with OLS, LAD gives less weight to large residuals. Thus, it is less influenced by changes in a small number of observations.

While LAD helps to guard against outliers, it does have some drawbacks. First, there are no formulas for the estimators; they can only be found by using iterative methods on a computer. This is not very difficult with the powerful personal computers of today, but large data sets can involve time-consuming computations. Second, LAD consistently estimates the parameters in the population regression function (the conditional mean), only when the distribution of the error term u is symmetric. And third, if the error u is normally distributed, LAD is less efficient (asymptotically) than OLS. Of course, if the error is truly normally distributed, the probability of getting a large outlier is small, and we would probably be satisfied with OLS.

Least absolute deviations is a special case of what is often called *robust regression.* In statistical terms, a robust regression estimator is relatively insensitive to extreme

observations: effectively, larger residuals are given less weight than in the least squares approach. While this characterization is accurate, usage of the term "robust" in this context can cause confusion. As mentioned earlier, the LAD estimator requires the error distribution to be symmetric about zero in order to consistently estimate the parameters in the conditional mean. This is not required of OLS. (Recall that the Gauss-Markov assumptions do not include symmetry of the error distribution.)

LAD does consistently estimate the parameters in the conditional median, whether or not the error distribution is symmetric. In some cases, this is of interest, but we will not pursue this idea now. Berk (1990) contains an introductory treatment of robust regression methods.

SUMMARY

We have further investigated some important specification and data issues that often arise in empirical cross-sectional analysis. Misspecified functional form makes the estimated equation difficult to interpret. Nevertheless, incorrect functional form can be detected by adding quadratics, computing RESET, or testing against a nonnested alternative model using the Davidson-MacKinnon test. No additional data collection is needed.

Solving the omitted variables problem is more difficult. In Section 9.2, we discussed a possible solution based on using a proxy variable for the omitted variable. Under reasonable assumptions, including the proxy variable in an OLS regression eliminates, or at least reduces, bias. The hurdle in applying this method is that proxy variables can be difficult to find. A general possibility is to use data on a dependent variable from a prior year.

Applied economists are often concerned with measurement error. Under the classical errors-in-variables (CEV) assumptions, measurement error in the dependent variable has no effect on the statistical properties of OLS. In contrast, under the CEV assumptions for an independent variable, the OLS estimator for the coefficient on the mismeasured variable is biased towards zero. The bias in coefficients on the other variables can go either way and is difficult to determine.

Nonrandom samples from an underlying population can lead to biases in OLS. When sample selection is correlated with the error term u, OLS is generally biased and inconsistent. On the other hand, exogenous sample selection—which is either based on the explanatory variables or is otherwise independent of u—does not cause problems for OLS. Outliers in data sets can have large impacts on the OLS estimates, especially in small samples. It is important to at least informally identify outliers and to reestimate models with the suspected outliers excluded.

KEY TERMS

Attenuation Bias
Classical Errors-in-Variables (CEV)
Davidson-MacKinnon Test
Endogenous Explanatory Variable
Endogenous Sample Selection
Exogenous Sample Selection
Functional Form Misspecification
Influential Observations

Lagged Dependent Variable
Measurement Error
Missing Data
Multiplicative Measurement Error
Nonnested Models
Nonrandom Sample
Outliers
Plug-In Solution to the Omitted Variables Problem
Proxy Variable
Regression Specification Error Test (RESET)

PROBLEMS

9.1 In Exercise 4.11, the *R*-squared from estimating the model

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + \beta_3 profmarg + \beta_4 ceoten + \beta_5 comten + u,$$

using the data in CEOSAL2.RAW, is $R^2 = .353$ ($n = 177$). When $ceoten^2$ and $comten^2$ are added, $R^2 = .375$. Is there evidence of functional form misspecification in this model?

9.2 Let us modify Exercise 8.9 by using voting outcomes in 1990 for incumbents who were elected in 1988. Candidate A was elected in 1988 and was seeking reelection in 1990; *voteA90* is Candidate A's share of the two-party vote in 1990. The 1988 voting share of Candidate A is used as a proxy variable for quality of the candidate. All other variables are for the 1990 election. The following equations were estimated, using the data in VOTE2.RAW:

$$\widehat{voteA90} = \underset{(9.25)}{75.71} + \underset{(.046)}{.312}\, prtystrA + \underset{(1.01)}{4.93}\, democA - \underset{(.684)}{.929} \log(expendA) - \underset{(0.281)}{1.950} \log(expendB)$$
$$n = 186,\ R^2 = .495,\ \bar{R}^2 = .483,$$

and

$$\widehat{voteA90} = \underset{(10.01)}{70.81} + \underset{(.052)}{.282}\, prtystrA + \underset{(1.06)}{4.52}\, democA - \underset{(.687)}{.839} \log(expendA) - \underset{(0.292)}{1.846} \log(expendB) + \underset{(.053)}{.067}\, voteA88$$
$$n = 186,\ R^2 = .499,\ \bar{R}^2 = .485.$$

(i) Interpret the coefficient on *voteA88* and discuss its statistical significance.
(ii) Does adding *voteA88* have much effect on the other coefficients?

9.3 Let *math10* denote the percentage of students at a Michigan high school receiving a passing score on a standardized math test (see also Example 4.2). We are interested in estimating the effect of per student spending on math performance. A simple model is

$$math10 = \beta_0 + \beta_1 \log(expend) + \beta_2 \log(enroll) + \beta_3 poverty + u,$$

where *poverty* is the percentage of students living in poverty.

(i) The variable *lnchprg* is the percentage of students eligible for the federally funded school lunch program. Why is this a sensible proxy variable for *poverty*?

(ii) The table that follows contains OLS estimates, with and without *lnchprg* as an explanatory variable.

Dependent Variable: *math10*

Independent Variables	(1)	(2)
log(*expend*)	11.13 (3.30)	7.75 (3.04)
log(*enroll*)	.022 (.615)	−1.26 (.58)
lnchprg	———	−.324 (.036)
intercept	−69.24 (26.72)	−23.14 (24.99)
Observations *R*-Squared	428 .0297	428 .1893

Explain why the effect of expenditures on *math10* is lower in column (2) than in column (1). Is the effect in column (2) still statistically greater than zero?

(iii) Does it appear that pass rates are lower at larger schools, other factors being equal? Explain.

(iv) Interpret the coefficient on *lnchprg* in column (2).

(v) What do you make of the substantial increase in R^2 from column (1) to column (2)?

9.4 The following equation explains weekly hours of television viewing by a child in terms of the child's age, mother's education, father's education, and number of siblings:

$$tvhours^* = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 motheduc + \beta_4 fatheduc + \beta_5 sibs + u.$$

We are worried that *tvhours** is measured with error in our survey. Let *tvhours* denote the reported hours of television viewing per week.

(i) What do the classical errors-in-variables (CEV) assumptions require in this application?

(ii) Do you think the CEV assumptions are likely to hold? Explain.

9.5 In Example 4.4, we estimated a model relating number of campus crimes to student enrollment for a sample of colleges. The sample we used was not a random sam-

ple of colleges in the United States, because many schools in 1992 did not report campus crimes. Do you think that college failure to report crimes can be viewed as exogenous sample selection? Explain.

COMPUTER EXERCISES

9.6 (i) Apply RESET from equation (9.3) to the model estimated in Problem 7.13. Is there evidence of functional form misspecification in the equation?

(ii) Compute a heteroskedasticity-robust form of RESET. Does your conclusion from part (i) change?

9.7 Use the data set WAGE2.RAW for this exercise.

(i) Use the variable *KWW* (the "knowledge of the world of work" test score) as a proxy for ability in place of *IQ* in Example 9.3. What is the estimated return to education in this case?

(ii) Now use *IQ* and *KWW* together as proxy variables. What happens to the estimated return to education?

(iii) In part (ii), are *IQ* and *KWW* individually significant? Are they jointly significant?

9.8 Use the data from JTRAIN.RAW for this exercise.

(i) Consider the simple regression model

$$\log(scrap) = \beta_0 + \beta_1 grant + u,$$

where *scrap* is the firm scrap rate and *grant* is a dummy variable indicating whether a firm received a job training grant. Can you think of some reasons why the unobserved factors in *u* might be correlated with *grant*?

(ii) Estimate the simple regression model using the data for 1988. (You should have 54 observations.) Does receiving a job training grant significantly lower a firm's scrap rate?

(iii) Now add as an explanatory variable $\log(scrap_{87})$. How does this change the estimated effect of *grant*? Interpret the coefficient on *grant*. Is it statistically significant at the 5% level against the one-sided alternative H_1: $\beta_{grant} < 0$?

(iv) Test the null hypothesis that the parameter on $\log(scrap_{87})$ is one against the two-sided alternative. Report the *p*-value for the test.

(v) Repeat parts (iii) and (iv), using heteroskedasticity-robust standard errors, and briefly discuss any notable differences.

9.9 Use the data for the year 1990 in INFMRT.RAW for this exercise.

(i) Restimate equation (9.37), but now include a dummy variable for the observation on the District of Columbia (called *DC*). Interpret the coefficient on *DC* and comment on its size and significance.

(ii) Compare the estimates and standard errors from part (i) with those from equation (9.38). What do you conclude about including a dummy variable for a single observation?

9.10 Use the data in RDCHEM.RAW to further examine the effects of outliers on OLS estimates. In particular, estimate the model

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \beta_3 profmarg + u$$

with and without the firm having annual sales of almost \$40 billion and discuss whether the results differ in important respects. The equations will be easier to read if you redefine *sales* to be measured in billions of dollars before proceeding (see Problem 6.3).

9.11 Redo Example 4.10 by dropping schools where teacher benefits are less than 1% of salary.

(i) How many observations are lost?

(ii) Does dropping these observations have any important effects on the estimated tradeoff?

9.12 Use the data in LOANAPP.RAW for this exercise.

(i) How many observations have $obrat > 40$, that is, other debt obligations more than 40% of total income?

(ii) Reestimate the model in part (iii) of Exercise 7.16, excluding observations with $obrat > 40$. What happens to the estimate and t statistic on *white*?

(iii) Does it appear that the estimate of β_{white} is overly sensitive to the sample used?