In this case, we happen to know the population mean of IQ, but in most cases we will not know the population mean of a proxy variable. Then, we should use the sample average to demean the proxy before interacting it with x_K; see Problem 4.8. Technically, using the sample average to estimate the population average should be reflected in the OLS standard errors. But, as you are asked to show in Problem 6.10 in Chapter 6, the adjustments generally have very small impacts on the standard errors and can safely be ignored.

In his study on the effects of computer usage on the wage structure in the United States, Krueger (1993) uses computer usage at home as a proxy for unobservables that might be correlated with computer usage at work; he also includes an interaction between the two computer usage dummies. Krueger does not demean the "uses computer at home" dummy before constructing the interaction, so his estimate on "uses a computer at work" does not have an average treatment effect interpretation. However, just as in Example 4.5, Krueger found that the interaction term is insignificant.

4.4 Properties of OLS under Measurement Error

As we saw in Section 4.1, another way that endogenous explanatory variables can arise in economic applications occurs when one or more of the variables in our model contains **measurement error**. In this section, we derive the consequences of measurement error for ordinary least squares estimation.

The measurement error problem has a statistical structure similar to the omitted variable–proxy variable problem discussed in the previous section. However, they are conceptually very 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 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 explanatory variable is the one whose effect is of primary interest. In the proxy variable case, we cannot estimate the effect of the omitted variable.

Before we turn to the analysis, it is important to remember that measurement error is an issue only when the variables on which we can collect data differ from the variables that influence decisions by individuals, families, firms, and so on. For example,

suppose we are estimating the effect of peer group behavior on teenage drug usage, where the behavior of one's peer group is self-reported. Self-reporting may be a mismeasure of actual peer group behavior, but so what? We are probably more interested in the effects of how a teenager perceives his or her peer group.

4.4.1 Measurement Error in the Dependent Variable

We begin with the case where the dependent variable is the only variable 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 saving. The regression model has the usual linear form

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + v \tag{4.37}$$

and we assume that it satisfies at least Assumptions OLS.1 and OLS.2. Typically, we are interested in $\mathrm{E}(y^* \mid x_1, \ldots, x_K)$. We let y represent the observable measure of y^* where $y \neq y^*$.

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

$$e_0 = y - y^* \tag{4.38}$$

For a random draw i from the population, we can write $e_{i0} = y_i - y_i^*$, but what is important 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 (4.37), and rearrange:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + v + e_0 \tag{4.39}$$

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 (4.37) satisfies Assumption OLS.1, v 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, this fact only affects estimation of the intercept, β_0. Much more important is what we assume 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, which implies that e_0 is uncorrelated with $\mathbf{x}$. Then, the OLS estimators from equation (4.39) are consistent (and possibly unbiased as well). Further, the usual OLS inference procedures (t statistics, F statistics, LM statistics) are asymptotically valid under appropriate homoskedasticity assumptions.

If e_0 and v are uncorrelated, as is usually assumed, then $\text{Var}(v + e_0) = \sigma_v^2 + \sigma_0^2 > \sigma_v^2$. Therefore, measurement error in the dependent variable results in a larger error variance than when the dependent variable is not measured with error. This result is hardly surprising and translates into larger asymptotic variances for the OLS estimators than if we could observe y^*. But the larger error variance violates none of the assumptions needed for OLS estimation to have its desirable large-sample properties.

Example 4.6 (Saving Function with Measurement Error): Consider a saving function

$$\text{E}(sav^* \mid inc, size, educ, age) = \beta_0 + \beta_1 inc + \beta_2 size + \beta_3 educ + \beta_4 age$$

but where actual saving (sav^*) may deviate from reported saving (sav). The question is whether the size of the measurement error in sav is systematically related to the other variables. It may be reasonable to assume that the measurement error is not correlated with *inc*, *size*, *educ*, and *age*, but we might expect that families with higher incomes, or more education, report their saving more accurately. Unfortunately, without more information, we cannot know whether the measurement error is correlated with *inc* or *educ*.

When the dependent variable is in logarithmic form, so that $\log(y^*)$ is the dependent variable, a natural measurement error equation is

$$\log(y) = \log(y^*) + e_0 \tag{4.40}$$

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

Example 4.7 (Measurement Error in Firm Scrap Rates): In Example 4.4, we might think that the firm scrap rate is mismeasured, leading us to postulate the model $\log(scrap^*) = \beta_0 + \beta_1 grant + v$, where $scrap^*$ is the true scrap rate. 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? Not if a firm receiving a grant is more likely to underreport its scrap rate in order to make it look as if the grant had the intended effect. If underreporting occurs, then, in the estimable equation $\log(scrap) = \beta_0 + \beta_1 grant + v + e_0$, the error $u = v + e_0$ is negatively correlated with *grant*. This result would produce a downward bias in β_1, tending to make the training program look more effective than it actually was.

These examples show that measurement error in the dependent variable *can* cause biases in OLS if the measurement error is systematically related to one or more of the explanatory variables. If the measurement error is uncorrelated with the explanatory variables, OLS is perfectly appropriate.

4.4.2 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 response variable. This point was suggested by Example 4.2, and in this subsection we develop the general case.

We consider the model with a single explanatory measured with error:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K^* + v \tag{4.41}$$

where $y, x_1, \ldots, x_{K-1}$ are observable but x_K^* is not. We assume at a minimum that v has zero mean and is uncorrelated with $x_1, x_2, \ldots, x_{K-1}, x_K^*$; in fact, we usually have in mind the structural model $\mathrm{E}(y \mid x_1, \ldots, x_{K-1}, x_K^*) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K^*$. If x_K^* were observed, OLS estimation would produce consistent estimators. Instead, we have a measure of x_K^*; call it x_K. A maintained assumption is that v is also uncorrelated with x_K. This follows under the redundancy assumption $\mathrm{E}(y \mid x_1, \ldots, x_{K-1}, x_K^*, x_K) = \mathrm{E}(y \mid x_1, \ldots, x_{K-1}, x_K^*)$, an assumption we used in the proxy variable solution to the omitted variable problem. This means that x_K has no effect on y once the other explanatory variables, including x_K^*, have been controlled for. Since x_K^* is assumed to be the variable that affects y, this assumption is uncontroversial.

The measurement error in the population is simply

$$e_K = x_K - x_K^* \tag{4.42}$$

and this can be positive, negative, or zero. We assume that the average measurement error in the population is zero: $\mathrm{E}(e_K) = 0$, which has no practical consequences because we include an intercept in equation (4.41). Since v is assumed to be uncorrelated with x_K^* and x_K, v is also uncorrelated with e_K.

We want to know the properties of OLS if we simply replace x_K^* with x_K and run the regression of y on $1, x_1, x_2, \ldots, x_K$. These depend crucially on the assumptions we make about the measurement error. An assumption that is almost always maintained is that e_K is uncorrelated with the explanatory variables not measured with error: $\mathrm{E}(x_j e_K) = 0$, $j = 1, \ldots, K-1$.

The key assumptions involve the relationship between the measurement error and x_K^* and x_K. Two assumptions have been the focus in the econometrics literature, and these represent polar extremes. The first assumption is that e_K is uncorrelated with the *observed* measure, x_K:

$$\mathrm{Cov}(x_K, e_K) = 0 \tag{4.43}$$

From equation (4.42), if assumption (4.43) is true, then e_K must be correlated with the unobserved variable x_K^*. To determine the properties of OLS in this case, we write $x_K^* = x_K - e_K$ and plug this into equation (4.41):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + (v - \beta_K e_K) \tag{4.44}$$

Now, we have assumed that v and e_K both have zero mean and are uncorrelated with each x_j, including x_K; therefore, $v - \beta_K e_K$ has zero mean and is uncorrelated with the x_j. It follows that OLS estimation with x_K in place of x_K^* produces consistent estimators of all of the β_j (assuming the standard rank condition Assumption OLS.2). Since v is uncorrelated with e_K, the variance of the error in equation (4.44) is $\text{Var}(v - \beta_K e_K) = \sigma_v^2 + \beta_K^2 \sigma_{e_K}^2$. Therefore, except when $\beta_K = 0$, measurement error increases the error variance, which is not a surprising finding and violates none of the OLS assumptions.

The assumption that e_K is uncorrelated with x_K is analogous to the proxy variable assumption we made in the Section 4.3.2. Since this assumption implies that OLS has all its nice properties, this is not usually what econometricians have in mind when referring to measurement error in an explanatory variable. The **classical errors-in-variables** (**CEV**) assumption replaces assumption (4.43) with the assumption that the measurement error is uncorrelated with the *unobserved* explanatory variable:

$$\text{Cov}(x_K^*, e_K) = 0 \tag{4.45}$$

This assumption comes from writing the observed measure as the sum of the true explanatory variable and the measurement error, $x_K = x_K^* + e_K$, and then assuming the two components of x_K are uncorrelated. (This has nothing to do with assumptions about v; we are always maintaining that v is uncorrelated with x_K^* and x_K, and therefore with e_K.)

If assumption (4.45) holds, then x_K and e_K *must* be correlated:

$$\text{Cov}(x_K, e_K) = \text{E}(x_K e_K) = \text{E}(x_K^* e_K) + \text{E}(e_K^2) = \sigma_{e_K}^2 \tag{4.46}$$

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

Looking at equation (4.44), we see that correlation between x_K and e_K causes problems for OLS. Because v and x_K are uncorrelated, the covariance between x_K and the composite error $v - \beta_K e_K$ is $\text{Cov}(x_K, v - \beta_K e_K) = -\beta_K \text{Cov}(x_K, e_K) = -\beta_K \sigma_{e_K}^2$. It follows that, in the CEV case, the OLS regression of y on $x_1, x_2, \ldots, x_K$ generally gives inconsistent estimators of *all* of the β_j.

The plims of the $\hat{\beta}_j$ for $j \neq K$ are difficult to characterize except under special assumptions. If x_K^* is uncorrelated with x_j, all $j \neq K$, then so is x_K, and it follows that plim $\hat{\beta}_j = \beta_j$, all $j \neq K$. The plim of $\hat{\beta}_K$ can be characterized in any case. Problem 4.10 asks you to show that

$$\text{plim}(\hat{\beta}_K) = \beta_K \left(\frac{\sigma^2_{r_K^*}}{\sigma^2_{r_K^*} + \sigma^2_{e_K}} \right) \tag{4.47}$$

where r_K^* is the linear projection error in

$$x_K^* = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_{K-1} x_{K-1} + r_K^*$$

An important implication of equation (4.47) is that, because the term multiplying β_K is always between zero and one, $|\text{plim}(\hat{\beta}_K)| < |\beta_K|$. 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* as a result of the presence of classical errors-in-variables. If β_K is positive, $\hat{\beta}_K$ will tend to underestimate β_K; if β_K is negative, $\hat{\beta}_K$ will tend to overestimate β_K.

In the case of a single explanatory variable ($K = 1$) measured with error, equation (4.47) becomes

$$\text{plim}\ \hat{\beta}_1 = \beta_1 \left(\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} \right) \tag{4.48}$$

The term multiplying β_1 in equation (4.48) is $\text{Var}(x_1^*)/\text{Var}(x_1)$, which is always less than unity under the CEV assumption (4.45). As $\text{Var}(e_1)$ shrinks relative to $\text{Var}(x_1^*)$, the attentuation bias disappears.

In the case with multiple explanatory variables, equation (4.47) shows that it is not $\sigma^2_{x_K^*}$ that affects plim$(\hat{\beta}_K)$ but the variance in x_K^* after netting out the other explanatory variables. Thus, the more collinear x_K^* is with the other explanatory variables, the worse is the attenuation bias.

Example 4.8 (Measurement Error in Family Income): Consider the problem of estimating the causal effect of family income on college grade point average, after controlling for high school grade point average and SAT score:

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

where $faminc^*$ is actual annual family income. Precise data on *colGPA*, *hsGPA*, and *SAT* are relatively easy to obtain from school records. But family income, especially