1 Measurement error

Until now, we have always assumed that all relevant variables \((y \text{ and } x)\) were measured without error. In reality, however, there are many reasons to believe that a substantial amount of error is introduced in the measurement of standard economic variables. For example, people may not recall exactly when asked about recent consumption expenditures, hours of work, earnings, etc. A number of validation surveys have indeed indicated that these errors can be substantial by comparing detailed employer administrative data about earnings to self reports by individuals. In light of this, it is important to assess how standard econometric estimates are affected by measurement error in either (or both) the explanatory or dependent variables.

1.1 Classical measurement error

The most standard form of measurement error that has been considered in the literature is what is called “classical measurement error”. Under classical measurement error, it is assumed that the observed value of \(y \text{ (or } x\) is equal to the true value of \(y \text{ (or } x\) plus a purely random component. More specifically, we can write the measured value of \(y\) as the sum of the true value \(y^*\) plus a measurement error \(u_y\):

\[
y = y^* + u_y.
\]

A first important result is that measurement error in the dependent variable does not affect the consistency of OLS estimates of \(\beta\). To see this, let’s first write the true model
y^* = x^* \beta + \varepsilon. \quad (2)

From equation (1) it follows that \( y^* = y - u_y \). Substituting back into equation (2) yields

\[ y = x^* \beta + \varepsilon + u_y. \quad (3) \]

So measurement error in \( y \) only changes the interpretation of the error term where we both have the standard error term, \( \varepsilon \), plus a component due to measurement error, \( u_y \). This does not affect the consistency of the OLS estimate \( \hat{\beta} \) since

\[
\text{plim}(\hat{\beta}) = \frac{\text{cov}(y, x^*)}{\text{var}(x^*)} = \frac{\text{cov}(x^* \beta + \varepsilon + u_y, x^*)}{\text{var}(x^*)} = \beta + \frac{\text{cov}(\varepsilon + u_y, x^*)}{\text{var}(x^*)} = \beta \quad (4)
\]

because \( \text{cov}(\varepsilon, x^*) = \text{cov}(u_y, x^*) = 0 \). Note, however, that adding a measurement error term does increase that variance of the error term from \( \text{var}(\varepsilon) \) to \( \text{var}(\varepsilon) + \text{var}(u_y) \). As a result, it also increases the variance (or standard error) of \( \hat{\beta} \) from \( (X'X)^{-1}\text{var}(\varepsilon) \) to \( (X'X)^{-1}[\text{var}(\varepsilon) + \text{var}(u_y)] \).

A second important result is that measurement error in the explanatory variable, \( x \), does result in a bias in the OLS estimate of \( \beta \). To see this, consider the measured value of \( x \) as the sum of the true value \( x^* \) plus a measurement error \( u_x \):

\[ x = x^* + u_x \quad (5) \]

Substituting \( x^* = x - u_x \) in equation (2) we now get that

\[ y^* = x \beta + (\varepsilon - \beta u_x) \quad (6) \]

This is obviously a problem since the presence of \( u_x \) in the error term generates a mechanical correlation between the error term, \( \varepsilon - \beta u_x \), and the explanatory variable, \( x = x^* + u_x \). In fact, it now follows that

\[
\text{plim}(\hat{\beta}) = \frac{\text{cov}(y^*, x)}{\text{var}(x)} = \frac{\text{cov}(x^* \beta + \varepsilon + u_y, x^* + u_x)}{\text{var}(x^* + u_x)}
\]

\[
= \frac{\text{var}(x^*)}{\text{var}(x^* + u_x)} \beta + \frac{\text{cov}(x^* \beta, u_x) + \text{cov}(\varepsilon, x^* + u_x)}{\text{var}(x^* + u_x)} = \frac{\text{var}(x^*)}{\text{var}(x^*) + \text{var}(u_x)} \beta. \quad (7)
\]
So \( \hat{\beta} \) now only converges in probability to a fraction \( \frac{\text{var}(x^*)}{\text{var}(x^*) + \text{var}(u_x)} < 1 \) of the true \( \beta \). This bias is called an attenuation bias since \( \hat{\beta} \) is biased towards 0 regardless of whether \( \beta \) is positive or negative. Note that the attenuation bias term is closely linked to what is called the signal-to-noise ratio in \( x \) since \( \text{var}(x^*) \) is the variance of the “signal” while \( \text{var}(u_x) \) is the variance of the “noise”. The larger the variance of the noise relative to the signal, the larger is the magnitude of the attenuation bias.

Note also that this simple and clear result about the form of the bias only works in the case of a univariate regression. With multiple regressors where only one of the element of (the vector) \( x \) is measured with error, the bias will also depend on the correlation between the \( x \) variable measured with error, \( x_1 \), and the other explanatory variables, \( x_2 \). The general intuition here is that if \( x_1 \) and \( x_2 \) are positively correlated, the coefficient on \( x_2 \) will be biased up (when \( \beta_1 \) is positive) as \( x_2 \) captures some of the effect of \( x_1 \) on \( y \) now that the regression coefficient on \( x_1 \) is “too low” because of the attenuation bias. In this setting, measurement error in \( x_1 \) could explain why a variable \( x_2 \) that should have no effect on \( y \) after conditioning on \( x_1 \) may be found to have significant effect in practice.

1.2 Measurement error bias in first-differences

When \( x \) is measured with error, first-differencing that is used to eliminate a fixed effect will generally increase the magnitude of the measurement error bias. To see this, consider a modified version of the above model where the variables are replaced with their first differences:

\[
\text{plim} \left( \hat{\beta}^{FD} \right) = \frac{\text{var}(\Delta x^*)}{\text{var}(\Delta x^*) + \text{var}(\Delta u_x)} \beta
\]  

The problem is that the variance in the noise term doubles from \( \text{var}(u_x) \) to \( \text{var}(\Delta u_x) = 2\text{var}(u_x) \). By contrast, the variance of the signal declines if there is substantial serial correlation in \( x^* \), which is typically the case with panel data (the serial correlation is equal to one in the extreme but common case of time-invariant regressors). If \( \rho_x \) is the autocorrelation in \( x^* \), it follows that

\[
\text{var}(\Delta x^*) = 2(1 - \rho_x)\text{var}(x^*)
\]

which is smaller than \( \text{var}(x^*) \) when \( \rho_x > .5 \). So unless \( \rho_x \leq 0 \), which is unlikely, first differencing will tend to magnify the attenuation bias due to measurement error. Using a similar approach, in can be shown that using the within (or difference from means) approach instead of first differences does not generally increase the measurement error
bias as much. Taking longer (say $t$ vs. $t - 3$) differences is another way of removing the fixed effect without increasing as much the measurement error bias as in the case of first differences. The intuition is that the autocorrelation term $\rho_x$ tends to decline for longer differences (for example in an AR(1) model), which results in a larger variance in the signal and a smaller attenuation bias. Griliches and Hausman have a classic piece on this that was published in the Journal of Econometrics in 1986.

In practical terms, if you find that first-difference estimates are closer to zero than within estimates or estimates based on longer differences, measurement error may well be the culprit. In such a case, you should trust more the latter set of estimates than standard first differences. It can also be shown that measurement error tends to magnify the attenuation bias in the random effect estimator relative to OLS. This is an additional reason for simply correcting the OLS standard errors by clustering instead of estimating the more efficient (but also more sensitive to measurement error) random effects model.

### 1.3 IV as a possible solution

What can be done about measurement error in the explanatory variables? Since the problem in equation (6) is that the error term is correlated with $x$, the standard solution is to find an instrumental variable for $x$ that is correlated with $x$, but not with the error term $\varepsilon - \beta u_x$. As it turns out, finding such a variable is not quite as difficult as in the standard case where the correlation with the error term is due to a deeper economic mechanism. In particular, when the problem is that $x$ is only an imperfect proxy for the true $x^*$, there may also be other proxies for $x^*$ that are available. For instance consider a second proxy $z$, where

$$z = x^* + u_z$$

Provided that the measurement error in $z$, $u_z$, is uncorrelated with the measurement error in $x$, $u_x$, we get that $z$ is a valid instrumental variable since:

$$\text{cov}(z, x) = \text{var}(x^*) > 0$$

and

$$\text{cov}(z, \varepsilon - \beta u_x) = \text{cov}(x^* + u_z, \varepsilon - \beta u_x) = 0$$

So the simple solution is to use one proxy as an instrument for the other proxy, and consistently estimate $\beta$ by TSLS. Note that $z$ will remain a valid instrument for $x$ (but not the other way around) even if $z$ has a different “scale”, i.e. $z = \gamma x^* + u_z$, with $\gamma \neq 1$.  

4
This can be quite useful with panel data, since \( z \) can then consist of lagged values of \( x \) where \( z \) is correlated with \( x \) (provided that \( \rho_x \neq 0 \)) but where the measurement errors are not correlated under the assumption that measurement error is iid (classical measurement error).

Another approach is to use group means as an instrumental variable. For example, if we have data on firms, the average value of \( x \) in the industry can be used as an instrument for firm-level \( x \). The two variables tend to be strongly correlated if \( x \) varies substantially across industries, while measurement error in the industry-level variable tends to be small since firm-level measurement errors get averaged out to zero when we compute the industry mean. Even better, the mean \( x \) among all other firms in the industry (the so-called “leave-one-out measure”) can be used as an instrument to make sure the firm-specific measurement error does not become part of the industry average. Note, however, that the exclusion restriction (assumption that the mean of \( x \) for other firms does not belong in the regression model) is not completely innocuous. Indeed, we can easily think of scenarios where characteristics of competing firms (e.g. how much R&D they do, etc.) may well have a direct impact on the firm by, for instance, reducing its share of the market.

### 1.4 Non-classical measurement error

Many validation studies have found that the assumption underlying classical measurement error was incorrect, and that measurement error was often correlated with the \( x \) variable. For example, several validation studies indicate that high-income people tend to under-report their income, while low-income tend to over-report. This means that \( u_x \) is negatively correlated with \( x^* \), which complicates the formula for the attenuation bias. The IV approach will still work provided that \( u_z \) remains uncorrelated with \( u_x \), but this won’t happen if the measurement error \( u_z \) is also non-classical and correlated with \( x^* \).

Another case where non-classical measurement error can prevail is when both \( y \) and \( x \) are defined partly in terms of the same underlying variables. Take for instance the case where both \( y \) and \( x \) are defined on a per capita basis. For example, say we have a sample of regions with regional level aggregate variables \( Y \) and \( X \) that are divided by population \( P \) to get a per capita number. When working with logs, we have

\[
    y = \log(Y) - \log(P) \quad \text{and} \quad x = \log(X) - \log(P)
\]

If measured population is based on a noisy estimate of the true population \( P^* \), where
log(P) = log(P*) + u_p, while Y and X are measured without error, we then get that
\[ y = \log(Y^*) - \log(P) = \log(Y^*) - \log(P^*) - u_p = y^* - u_p \text{ and } \]
\[ x = \log(X^*) - \log(P) = \log(X^*) - \log(P^*) - u_p = x^* - u_p. \]

So we now have the same measurement error in x and y, which results in inconsistent OLS estimates since
\[
\text{plim}(\hat{\beta}) = \frac{\text{cov}(y, x)}{\text{var}(x)} = \frac{\text{cov}(y^* - u_p, x^* - u_p)}{\text{var}(x^* - u_p)}
\]
\[
= \frac{\text{cov}(x^* \beta + \varepsilon - u_p, x^* - u_p)}{\text{var}(x^* - u_p)} = \frac{\beta \cdot \text{var}(x^*) + \text{var}(u_p)}{\text{var}(x^*) + \text{var}(u_p)} \neq \beta
\]
(13)

As before, the bias goes away when the variance of the signal, \(\text{var}(x^*)\), becomes very large relative to the variance in the noise, \(\text{var}(u_p)\). But when the signal-to-noise ratio is not that large, the mechanical correlation between y and x due to \(u_p\) tends to bias \(\hat{\beta}\) towards 1. For example, even if the true \(\beta\) is zero, we have
\[
\text{plim}(\hat{\beta}) = \frac{\text{var}(u_p)}{\text{var}(x^*) + \text{var}(u_p)}
\]
(14)
which lies between 0 and 1, depending on the signal-to-noise ratio.

2 Appendix: measurement error with 2 explanatory variables

As noted earlier, with multiple regressors where only one of the element of (the vector) \(x\) is measured with error, the bias will also depend on the correlation between the \(x\) variable measured with error, \(x_1\), and the other explanatory variables, \(x_2\). Consider the case where the true \(\beta_1\) is positive. In the model
\[ y = x_1^* \beta_1 + x_2^* \beta_2 + \varepsilon \]
(15)
we have:
\[
\text{plim}(\hat{\beta}) = Var(x^*)^{-1} Cov(x^*, y)
\]
(16)
and
\[
\text{plim}(\hat{\beta}_1) = D^{-1}[\sigma_2^2 \sigma_{y1} - \sigma_{12} \sigma_{y2}] > 0 \tag{17}
\]

where \(D\) is the determinant of the matrix \(Var(x)\). When \(x_1^*\) is measured with error we have instead the model

\[
y = x_1 \beta_1 + x_2^* \beta_2 + (\varepsilon - \beta_1 u_x) \tag{18}
\]

It follows that

\[
\text{plim}(\hat{\beta}_2) = D^{-1}[\sigma_1^2 + \sigma_u^2]\sigma_{y2} - \sigma_{12} \sigma_{y1}] \tag{19}
\]

where the matrix determinant is:

\[
D = [(\sigma_1^2 + \sigma_u^2)\sigma_2^2 - \sigma_{12}^2] \tag{20}
\]

and:

\[
\text{plim}(\hat{\beta}_2) = \frac{(\sigma_1^2 + \sigma_u^2)\sigma_{y2} - \sigma_{12} \sigma_{y1}}{(\sigma_1^2 + \sigma_u^2)\sigma_2^2 - \sigma_{12}^2} \tag{21}
\]

The effect of increase the variance of the measurement error \(\sigma_u^2\) is:

\[
\frac{d\text{plim}(\hat{\beta}_2)}{d\sigma_u^2} = \frac{\sigma_{y2}}{(\sigma_1^2 + \sigma_u^2)\sigma_2^2 - \sigma_{12}^2} - \frac{\sigma_u^2[(\sigma_1^2 + \sigma_u^2)\sigma_{y2} - \sigma_{12} \sigma_{y1}]}{[(\sigma_1^2 + \sigma_u^2)\sigma_2^2 - \sigma_{12}^2]^2} \tag{22}
\]

which can be simplified as

\[
\frac{d\text{plim}(\hat{\beta}_2)}{d\sigma_u^2} = \frac{\sigma_{12}[\sigma_2^2 \sigma_{y1} - \sigma_{12} \sigma_{y2}]}{[(\sigma_1^2 + \sigma_u^2)\sigma_2^2 - \sigma_{12}^2]^2} \tag{23}
\]

where the numerator is positive when \(\text{plim}(\hat{\beta}_1) = D^{-1}[\sigma_2^2 \sigma_{y1} - \sigma_{12} \sigma_{y2}]\) is also positive.