14.661: Recitation 8

Chris Walters

November 3, 2010

1 Measurement Error

1.1 Classical RHS measurement error

We covered this case in a previous recitation. The model of interest is

$$y_i^* = \alpha + \beta x_i^* + \epsilon_i$$

with $Cov(x_i^*, \epsilon_i) = 0$. We observe y_i^* and

$$x_i = x_i^* + v_i,$$

with $Cov(x_i^*, v_i) = Cov(\epsilon_i, v_i) = 0$

Note that we can write

$$y_i^* = \alpha + \beta(x_i - v_i) + \epsilon_i$$

$$\implies y_i^* = \alpha + \beta x_i + \epsilon_i - \beta v_i$$

We have correlation between x_i and our new error term $\epsilon_i - \beta v_i$, so we can expect problems with OLS. If we regress y_i^* on x_i, our estimator will plim to

$$plim\hat{\beta}_{OLS} = \frac{Cov\left(y_i^*, x_i\right)}{Var(x_i)}$$

$$= \frac{Cov\left(\alpha + \beta x_i^* + \epsilon_i, x_i^* + v_i\right)}{Var\left(x_i^*\right) + Var(v_i) + 2Cov(x_i^*, v_i)}$$

$$= \beta \cdot \frac{Var\left(x_i^*\right)}{Var\left(x_i^*\right) + Var\left(v_i\right)}$$

This is classical attenuation bias, which pulls our estimate of β towards zero.

1.2 Classical LHS measurement error

Now suppose we observe the true x_i^*, but instead of y_i^* we observe the dependent variable with error:

$$y_i = y_i^* + u_i$$

with $Cov(u_i, x_i^*) = Cov(u_i, \epsilon_i) = 0$. Our OLS estimate of β will plim to

$$plim\hat{\beta}_{OLS} = \frac{Cov\left(x_i^*, y_i\right)}{Var\left(x_i^*\right)}$$

$$= \frac{Cov\left(x_i^*, \alpha + \beta x_i^* + \epsilon_i + v_i\right)}{Var\left(x_i^*\right)}$$

$$= \beta \cdot \frac{Var\left(x_i^*\right)}{Var\left(x_i^*\right)}$$

$$= \beta$$

So there is no problem with classical measurement error on the left-hand side. Intuitively, such measurement error is just adding to the residual variance in the model (so it will increase our standard errors), but it doesn't affect consistency.

1.3 Division bias

Suppose we are not worried about OVB for the moment, and we want to estimate the equation

$$\log h_i^* = \alpha + \delta \log w_i^* + \eta_i$$

If we had data on true hours and wages h_i^* and w_i^* we would be able to consistently estimate δ; in other words, $Cov(\log w_i^*, \eta_i) = 0$.

However, many people do not report an explicit hourly wage, so we have to compute w. If we had perfect measures of hours and earnings, we could successfully compute the true w_i^* :

$$w_i^* = \frac{y_i^*}{h_i^*}$$

Furthermore, hours are measured with multiplicative error v_i, which we can assume is independent of everything else in the model. We observe:

$$h_i = h_i^* \cdot v_i,\, y_i^*$$

So observed wages are

$$w_i = \frac{y_i^*}{h_i} = \frac{y_i^*}{v_i h_i^*}$$

$$\implies \log w_i = \log w_i^* - \log v_i$$

Suppose we regress observed hours on observed wages using OLS. Then we obtain

$$plim\hat{\delta} = \frac{Cov\left(\log h_i, \log w_i\right)}{Var\left(\log w_i\right)}$$

$$= \frac{Cov\left(\log h_i^* + v_i, \log w_i^* - \log v_i\right)}{Var\left(\log w_i^* - \log v_i\right)}$$

$$= \frac{Cov\left(\alpha + \delta \log w_i^* + \eta_i + \log v_i, \log w_i^* - \log v_i\right)}{Var\left(\log w_i^* - \log v_i\right)}$$

$$= \delta \frac{Var\left(\log w_i^*\right)}{Var\left(\log w_i^*\right) + Var\left(\log v_i\right)} - \frac{Var\left(\log v_i\right)}{Var\left(\log w_i^*\right) + Var\left(\log v_i\right)}$$

The first term shows the usual attenuation bias result – δ is multiplied by a positive number less than one, so the measurement error in w will pull our estimate towards zero. However, we also have a second term that is unambiguously negative; this term comes from the correlation between the measurement error on the left hand side and the measurement error in the denominator of our right hand side variable of interest. With a positive δ, this makes attenuation bias worse and can even reverse the sign of the coefficient.

1.4 Additional controls

Suppose the true model is as before, we observe x_i instead of x_i^*, and we include an additional regressor a_i (even though there is no omitted variable bias). Assume our measurement error and ϵ are uncorrelated with a_i. As usual, we can think about running our multivariate regression in 2 steps:

1. Run $x_i = \pi_0 + \pi_1 a_i + \eta_i$, and compute the residuals $\tilde{x}_i = x_i - \hat{\pi}_0 - \hat{\pi}_1 a_i$
2. Regress y_i^* on $\tilde{x}_i$

The plim of our multivariate regression coefficient will then be

$$plim\hat{\beta}_{OLS} = \frac{Cov\left(y_i^*, \tilde{x}_i\right)}{Var\left(\tilde{x}_i\right)}$$

Since we are working in plims we can replace the estimated regression coefficients in $\tilde{x}$ with their population counterparts. This quantity then becomes

$$plim\hat{\beta}_{OLS} = \frac{Cov\left(\alpha + \beta x_i^* + \epsilon_i, x_i - \pi_0 - \pi_1 a_i\right)}{Var\left(x_i - \pi_0 - \pi_1 a_i\right)}$$

$$= \frac{\beta Var\left(x_i^*\right) - \beta \pi_1 Cov\left(x_i^*, a_i\right)}{Var\left(x_i\right) + \pi_1^2 Var\left(a_i\right) - 2\pi_1 Cov\left(x_i, a_i\right)}$$

Since π_1 is a regression coefficient, it is

$$\pi_1 = \frac{Cov\left(x_i, a_i\right)}{Var\left(a_i\right)} = \frac{Cov\left(x_i^*, a_i\right)}{Var\left(a_i\right)}$$

so

$$plim\hat{\beta}_{OLS} = \frac{\beta Var\left(x_i^*\right) - \beta\frac{Cov\left(x_i^*,a_i\right)}{Var(a_i)}Cov\left(x_i^*,a_i\right)}{Var\left(x_i^*\right) + Var(v_i) + \frac{Cov\left(x_i^*,a_i\right)^2}{Var(a_i)^2}Var\left(a_i\right) - 2\frac{Cov\left(x_i^*,a_i\right)}{Var(a_i)}Cov\left(x_i^*,a_i\right)}$$

$$= \frac{\beta Var\left(x_i^*\right) - \beta\frac{Cov(x_i^*,a_i)^2}{Var(a_i)}}{Var\left(x_i^*\right) + Var(v_i) - \frac{Cov\left(x_i^*,a_i\right)^2}{Var(a_i)}}$$

It is a fact that the R^2 from a univariate regression is equal to the square of the correlation coefficient between two variables. So

$$R^2_{xa} = \frac{Cov(x_i^*,a_i)^2}{Var(a_i)Var(x_i^*)}$$

We can therefore re-write the plim of our estimator as

$$plim\hat{\beta}_{OLS} = \frac{\beta Var\left(x_i^*\right) - \beta R^2_{xa} Var\left(x_i^*\right)}{Var\left(x_i^*\right) + Var(v_i) - R^2_{xa} Var(x_i^*)}$$

$$\implies plim\hat{\beta}_{OLS} = \beta\frac{\left(1 - R^2_{xa}\right) Var\left(x_i^*\right)}{\left(1 - R^2_{xa}\right) Var\left(x_i^*\right) + Var\left(v_i\right)}$$

Note that this quantity is increasing in R^2_{xw}– it yields normal attenuation when a and x aren't related, and a zero coefficient as a gets close to explaining all of x. This occurs because in our setup, partialling out a is predicting the signal in x but not the noise. Of course, if we are including a_i because leaving it out would case omitted variable bias, then it's not obvious that including it makes things worse; there is a tradeoff between exacerbating measurement error and eliminating OVB.

1.5 Fixed effects/first differencing

A special case of the "additional controls" scenario is individual fixed effects. Fixed effects can be especially problematic if there is a high degree of serial correlation in the variable of interest. In this case, removing individual means will eliminate a large part of the signal. I will do the derivation for first differences, which is equivalent to fixed effects if there are two time periods.

Suppose the model is

$$y^*_{it} = \alpha + \beta x^*_{it} + \epsilon_{it}$$

We observe y^*_{it} and

$$x_{it} = x^*_{it} + v_{it}$$

and $Cov\left(x^*_{is}, \epsilon_{it}\right) = Cov\left(x^*_{is}, v_{it}\right) = Cov\left(\epsilon_{is}, v_{it}\right) = 0 \ \forall t, s$, as well as $Cov(x^*_{is}, \eta_{it}) = Cov(v_{it}, v_{is}) = 0$ for $s \neq t$. The first-differenced regression is

$$\Delta y^*_{it} = \beta \Delta x_{it} + u_{it}$$

This regression will give us

$$plim\hat{\beta}_{FD} = \frac{Cov\left(\Delta x_{it}, \Delta y^*_{it}\right)}{Var\left(\Delta x_{it}\right)}$$

$$= \frac{Cov\left(\Delta x^*_{it} + \Delta v_{it}, \beta \Delta x^*_{it} + \Delta \epsilon_{it}\right)}{Var\left(\Delta x^*_{it} + \Delta v_{it}\right)}$$

$$= \beta \cdot \frac{Var\left(\Delta x^*_{it}\right)}{Var\left(\Delta x^*_{it}\right) + 2Var\left(v_{it}\right)}$$

Let's define

$$\rho \equiv Corr\left(x^*_{it}, x^*_{it-1}\right)$$

so that

$$Cov\left(x^*_{it}, x^*_{it-1}\right) = \rho Var\left(x^*_{it}\right)$$

Then

$$Var\left(\Delta x^*_{it}\right) = 2Var\left(x^*_{it}\right) - 2Cov\left(x^*_{it}, x^*_{it-1}\right)$$

$$= 2\left(1-\rho\right) Var\left(x^*_{it}\right)$$

So

$$plim\hat{\beta}_{FD} = \beta \cdot \frac{\left(1-\rho\right) Var\left(x^*_{it}\right)}{\left(1-\rho\right) Var\left(x^*_{it}\right) + Var\left(v_{it}\right)}$$

The attenuation bias in this model is increasing in ρ, the serial correlation in x^*; as $\rho \to 1$, it approaches zero. Again, note that it is only strictly "worse" to difference this equation if we don't need to do it because of OVB; if OVB is present, OLS will be inconsistent too.

1.6 Mismeasured Controls

Now, suppose our model of interest is

$$y^*_i = \alpha + \beta x^*_i + \gamma w^*_i + \epsilon_i$$

with

$$Cov(x^*_i, \epsilon_i) = Cov(w^*_i, \epsilon_i) = 0$$

β is the parameter of interest.

We observe y^*_i and x^*_i, but instead of observing the covariate w^*_i we observe

$$w_i = w^*_i + \mu_i$$

where

$$Cov(\mu_i, w_i^*) = Cov(\mu_i, x_i^*) = Cov(\mu_i, \epsilon_i) = 0.$$

We are worried about omitted variables bias due to w_i^*, so we include w_i to “proxy” for this variable. Using our standard partialling out argument, we have

$$plim\hat{\beta}_{OLS} = \frac{Cov\left(y_i, \tilde{x}_i^*\right)}{Var\left(\tilde{x_i}^*\right)}$$

where

$$\tilde{x}_i^* = x_i^* - \pi_0 - \pi_1 w_i$$

Here π_0 and π_1 are population coefficients from regressing x_i^* on w_i. Then

$$plim\hat{\beta}_{OLS} = \frac{Cov\left(\alpha + \beta x_i^* + \gamma w_i^* + \epsilon_i, x_i^* - \pi_0 - \pi_1 w_i\right)}{Var\left(x_i^* - \pi_0 - \pi_1 w_i\right)}$$

$$= \frac{Cov\left(\beta x_i^* + \gamma w_i^*, x_i^* - \pi_1 w_i\right)}{Var\left(x_i^*\right) + \pi_1^2 Var(w_i)}$$

$$= \frac{\beta Var\left(x_i^*\right) - \pi_1 \beta Cov\left(x_i^*, w_i\right) + \gamma Cov\left(x_i^*, w_i^*\right) - \pi_1 \gamma Cov\left(w_i^*, w_i\right)}{Var\left(x_i^*\right) + \pi_1^2 Var(w_i)}$$

$$= \frac{\beta \left(Var\left(x_i^*\right) - \frac{Cov(x_i^*, w_i)^2}{Var(w_i)}\right) + \gamma Cov\left(x_i^*, w_i^*\right) - \pi_1 \gamma Cov\left(w_i^*, w_i\right)}{Var\left(x_i^*\right) + \frac{Cov(x_i^*, w_i)^2}{Var(w_i)}}$$

$$= \beta + \gamma \cdot \frac{Cov\left(x_i^*, w_i^*\right) - \frac{Cov(x_i^*, w_i)}{Var(w_i)} Var(w_i^*)}{Var\left(x_i^*\right) + \frac{Cov(x_i^*, w_i)^2}{Var(w_i)}}$$

$$= \beta + \gamma \frac{Cov\left(x_i^*, w_i^*\right)}{Var(x_i^*)} \cdot \left[\frac{1 - \frac{Var(w_i^*)}{Var(w_i)}}{1 + \frac{Cov(x_i^*, w_i)^2}{Var(w_i) Var(x_i^*)}}\right]$$

$$= \beta + \gamma \frac{Cov\left(x_i^*, w_i^*\right)}{Var(x_i^*)} \cdot \left[\frac{1 - \frac{Var(w_i^*)}{Var(w_i)}}{1 + R_{xw}^2}\right]$$

So there is still some portion of the omitted variables bias left; how much depends on the reliability ratio for the covariate w_i^*.