Measurement Error Correction

Notes for Summer School
Moscow State University, Faculty of Economics

Andrey Simonov*

June 2013

1 Measurement error

1.1 Problem

Consider two examples. We would like to run a regression

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_i$$

but we do not observe x_2. Instead, in the first case we observe a z_2 which is x_2 and some measurement error:

$$z_{2i} = x_{2i} + \nu_i$$

where ν_i is a random variable with zero mean, independent of x_{2i} (think about hiring a bad RA who will consistently make mistakes). Using z_{2i} instead of x_{2i} implies running a regression:

$$y_i = x_{1i}\beta_1 + z_{2i}\beta_2 + \epsilon_i - \nu_i\beta_2 = x_{1i}\beta_1 + z_{2i}\beta_2 + \eta_i$$

Running OLS regression yields

$$plim(\hat{\beta}_2) = \frac{cov(z^*_{2i}, y^*_i)}{var(z^*_{2i})} = \beta_2 + \frac{cov(z^*_{2i}, \eta_i)}{var(z^*_{2i})} = \beta_2 - \beta_2\frac{\sigma_\nu}{\sigma_{x_2} + \sigma_\nu} = \beta_2\frac{\sigma_{x_2}}{\sigma_{x_2} + \sigma_\nu} \tag{1}$$

where $z^*_{2i} = M_1 z_{2i}$, $y_i = M_1 y^*_i$, and $M_1 = (I - x_1(x_1'x_1)^{-1}x_1')$. This implies that if variance of ν_i is greater than zero (that is, if there is a measurement error), estimate would be *attenuated*.

Moreover, it appears that estimator of β_1 would generally also be inconsistent:

$$plim(\hat{\beta}_1) = \tilde{\beta}_1 - plim(\hat{\beta}_2)\gamma \tag{2}$$

*University of Chicago, Booth School of Business. All errors and typos are of my own. Please report these or any other questions to asimonov@chicagobooth.edu

where $\tilde{\beta}_1$ is a coefficient in population regression of y on x_1 alone, and γ is the coefficient in population regression of x_2 on x_1. In general from equation (2) we cannot conclude about the direction of bias. Moreover, if x_1 is correlated with ν bias is even harder to predict.

On the other hand, if x_1 is uncorrelated with x_2 γ would be zero, which would imply that 1) $\tilde{\beta}_1 = \beta_1$ 2) $plim(\hat{\beta}_1) = \tilde{\beta}_1$. Hence, OLS estimator of β_1 would be consistent.

Now lets think about a different case. Again, we would like to run a regression

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_i$$

and again we do not observe x_2. Instead, we observe a proxy z_2 such that

$$x_{2i} = z_{2i} + u_i$$

where u_i is a prediction error such that $E(u|z_2) = 0$. Using z_2 instead of x_2 we get

$$y_i = x_{1i}\beta_1 + z_{2i}\beta_2 + \epsilon_i + u_i\beta_2 = x_{1i}\beta_1 + z_{2i}\beta_2 + \eta_i$$

Running OLS:

$$plim(\hat{\beta}_2) = \frac{cov(z^*_{2i}, y^*_i)}{var(z^*_{2i})} = \beta_2 + \frac{cov(z^*_{2i}, \eta_i)}{var(z^*_{2i})} = \beta_2 \tag{3}$$

as $E(\eta_i|x_1, z_2) = 0$, so OLS estimator is consistent.

The bottom line is: it is very important what we assume as our *data generating process (DGP)*. Some people (i.e. Derek Neal) would claim that assumptions on DGP is the part from which every paper should start. How to support this assumptions? That is why economics is also an art, not only a sciences.

1.2 Solutions

Reconsider results in (1) and (2). If we have several proxies which are correlated with the unobserved component (example 1), which would we choose? Equation (1) tells us that we need to choose the one with the minimum variance of the unobserved component ν_i. Equation (2) tells that we need to choose the one which is not correlated with x_1 (however, it is simple to show that in this case we can just omit x_2).

1.2.1 Instrumental variables

Another familiar solution to this problem is *instrumental variables (IV)* approach. Assume the same regression as before

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_i$$

with x_{2i} being unobserved, but instead we observe a proxy

$$\tilde{x}_{2i} = x_{2i} + \nu_i$$

Suppose we can find an instrument z_i, such that

- Instrument is correlated with x_{2i};
- Instrument is not correlated with either ν_i or ϵ_i.

Using z_i as an instrument for $\tilde{x}_{2i}$, we get a consistent estimator of β_2:

$$plim(\hat{\beta}_{2IV}) = \frac{cov(y_i^*, z_i)}{cov(\tilde{x}_{2i}^*, z_i)} = \beta_2$$

An interesting example of such instrument can be (fairly rare) case when we observe two proxies with uncorrelated measurement errors (think about the case when you hire two RA for coding, and they make mean zero mistakes independently). In this case we have

$$x'_{2i} = x_{2i} + \nu'_i$$

$$x''_{2i} = x_{2i} + \nu''_i$$

ν'_i and ν''_i are uncorrelated with each other and with ϵ_i. Then

$$plim(\hat{\beta}_{2IV}) = \frac{cov(y_i^*, x''_{2i})}{cov(x'^*_{2i}, x''_{2i})} = \beta_2$$

1.2.2 High Order Moment Estimator

What to do if we do not have a good instrument? If we can make stronger assumptions on ν_i, ϵ_i and x_{2i} there is a way to proceed. We make three basic assumptions (ν_i, ϵ_i and x_{2i} are iid, ν_i, ϵ_i and x_{2i} have moments of every order, $E(\nu) = E(\epsilon) = 0$) and add two restrictive assumptions:

- ν_i and ϵ_i are distributed *independently* of each other and of x_{2i};
- $\beta_2 \neq 0$ and x_{2i} is not normally distributed (in particular, *skewed*).

Under this conditions one can show that

$$E(y_i^2 \tilde{x}_{2i}) = \beta_2^2 E(x_{2i}^3)$$

$$E(y_i \tilde{x}_{2i}) = \beta_2 E(x_{2i}^3)$$

So if $\beta_2 \neq 0$ and $E(x_{2i}^3) \neq 0$

$$\hat{\beta}_2 = E(y_i^2 \tilde{x}_{2i}) / E(y_i \tilde{x}_{2i}^2) = \beta_2$$

This estimator is called the third-order moment estimator. Interestingly, it can be derived via IV procedure. Under assumptions above $y_i \tilde{x}_{2i}$ appears to be a valid and relevant instrument. IV estimation produces:

$$plim(\hat{\beta}_{2IV}) = \frac{cov(y_i^*, y_i \tilde{x}_{2i})}{cov(\tilde{x}_{2i}^*, y_i \tilde{x}_{2i})} = \frac{cov(x_{2i}, x_{2i}^2)\beta_2^2 + cov(\nu_i, \nu_i^2)\beta_2^2}{cov(x_{2i}, x_{2i}^2)\beta_2 + cov(\nu_i, \nu_i^2)\beta_2} = \beta_2$$

Is it restrictive to assume that distribution of x_{2i} is skewed? Well, at least one needs to argue that. Roberts and Whited (2012) [1] give an example of marginal q in the investment problem. Like many other valuation ratios in finance, marginal q is considered to be highly skewed. That is, a number of papers used Tobin's q as a proxy for marginal q (i.e. see Erickson and Whited 2000 [2], Chen and Chen 2012 [3])

1.2.3 Reverse Regression Bounds

What if assumptions in previous section seem to be not plausible? There is another solution that can be proposed: setting bounds on β_2. Again, consider

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_i$$

with x_{2i} being unobserved, but instead we observe a proxy

$$\tilde{x}_{2i} = x_{2i} + \nu_i$$

Assume both β_1 and β_2 are (strictly) greater than zero[1]. Now run two regressions:

$$y_i = x_{1i}b_1 + \tilde{x}_{2i}b_2 + \epsilon_i$$

$$x_{1i} = y_i\frac{1}{\bar{b}_1} + \tilde{x}_{2i}\frac{-\bar{b}_2}{\bar{b}_1} + \epsilon_i\frac{-1}{\bar{b}_1}$$

Gini[2] (1921) [4] showed that true coefficients β_1 and β_2 must lie between $(b_1, \bar{b}_1)$ and $(b_2, \bar{b}_2)$, respectively. Standard errors for $\bar{b}$ estimators can be computed via delta method.

If measurement error is severe the bounds would be too wide, and this procedure wound not be informative. However, this still appears to be a good diagnostic test.

1.3 Conclusions

The bottom line of this is:

- State assumptions on DGP and on relation between variables/error terms very clearly;
- Defend this assumptions carefully;
- If testing a hypothesis like $H0 : \beta_2 = 0$ - try to use proxies that will make type 1 error less likely (more difficult to reject the null).

[1] This is assumed without loss of generality

[2] Roberts and Whited cite this paper which is in Italian. Derivation of this result is fairly straightforward.

References

[1] Roberts, Michael R., Toni M. Whited (2012), Endogeneity in Empirical Corporate Finance, *Working Paper*

[2] Erickson, Timothy and Toni M. Whited, (2000), Measurement error and the relationship between investment and q, *Journal of Political Economy* 108, 1027-1057.

[3] Chen, Huafeng (Jason) and Shaojun Chen, (2012), Investment-cash flow sensitivity cannot be a good measure of financial constraints: Evidence from the time series, *Journal of Financial Economics*, 103, 2012.

[4] Gini, Corrado, 1921, Sullinterpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali, *Metroeconomica* 1, 63-82.