PROBLEM SET 7

PAUL SCHRIMPF

DUE: TUESDAY MARCH 24TH, 2015 AT THE START OF LECTURE

UNIVERSITY OF BRITISH COLUMBIA

ECONOMICS 326

Problem 1 (Measurement error)**:** Consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i$$

with assumptions $MLR.1-4$. In particular, assume that $\mathrm{E}[x_i^* \epsilon_i] = 0$. However, instead of observing x_i^* we only observed

$$x_i = x_i^* + e_i.$$

We can think of x_i as some measurement of x_i^* that is subject to error. We will assume that $\mathrm{E}[e_i] = 0$, $\mathrm{E}[e_i \epsilon_i] = 0$ and $\mathrm{E}[e_i x_i^*] = 0$.

(i) Suppose we estimate the model using OLS with the observed x_i in place of x_i^*,

$$y_i = \beta_0 + \beta_1 x_i + \underbrace{(\epsilon_i - \beta_1 e_i)}_{u_i}.$$

Note that the error term in the regression is $u_i = \epsilon_i - \beta_1 e_i$. Show that

$$\operatorname{plim} \hat{\beta}_1^{OLS} = \beta_1 \frac{\mathrm{Var}(x^*)}{\mathrm{Var}(x^*) + \mathrm{Var}(e)}.$$

When there is measurement error, the OLS estimate is closer to zero than β_1. This is called attenuation bias.

(ii) Suppose we have a second measurement of x_i^*, z_i such that $\mathrm{Cov}(z, x^*) \neq 0$, $\mathrm{E}[ze] = 0$, and $\mathrm{E}[z\epsilon] = 0$. Argue that

$$\hat{\beta}^{IV} = \frac{\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z}) y_i}{\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z}) x_i}$$

is a consistent estimate of β_1.

Problem 2 (Wald estimator (based on Wooldridge (2013) 15.3))**:** Consider the simple regression model

$$y = \beta_0 + \beta_1 x + \epsilon$$

and let z be a binary instrumental variable for x (i.e. z_i is either 0 or 1). Show that $\hat{\beta}_1^{IV}$ can be written

$$\hat{\beta}_1^{IV} = \hat{\beta}_1^{W} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

where $\bar{y}_0$ and $\bar{x}_0$ are the sample averages of y_i and x_i over the part of the sample with $z_i = 0$ and $\bar{y}_1$ and $\bar{x}_1$ are the sample averages of y_i and x_i over the part of the sample with $z_i = 1$. This estimator was first suggested by Wald (1940) to deal with measurement error in x_i. Wooldridge calls it a *grouping estimator*. Angrist and Pischke (2009) and I (and others) call it the *Wald estimator*.

Problem 3 (Based on Angrist (2009) Problem Set 6)**:** (This problem uses the same class size data you used to do problem 4 in problem set 3. Example code that complete portions of parts (iii) and (iv) will be posted on the course web page.)

In Israel, class size is capped at integer multiples of 40. In other words, if there are 40 kids in your grade, you're in a class of 40 (usually), but if there are 41, the class is (usually) split.

Assuming a new class is added every time enrollment exceeds an integer multiple of 40 generates the following predicted class size variable for class size in a school with enrollment e_s

$$z_s = \frac{e_s}{\text{floor}((e_s - 1)/40) + 1},$$

where floor(x) is the largest integer less than or equal to x. For example floor$(10.354) = 10$ and floor$(3) = 3$. Angrist and Lavy (1999) use z_s as an instrumental variable for class size in regressions of test scores on enrollment and class size. They call z_s "Maimonides Rule," because the medieval Talmudic scholar Moses Maimonides proposed that class size be capped at 40.

(i) Explain the rationale for this IV strategy.
(ii) Why is it a good idea to control for enrollment when using z_s as an instrumental variable?
(iii) Replicate the first stage and reduced form estimates in Angrist and Lavy (1999) Table III Columns 1-6 of Panel A[1]
(iv) Replicate the 2SLS estimates corresponding to these reduced form estimates (columns 1,2,7, and 8 of table IV).
(v) Why do you think the 2SLS estimates show benefits from reducing class size while the OLS estimates do not (you calculated the OLS estimates in problem set 4), even with controls?

REFERENCES

Angrist, J.D. and V. Lavy. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *The Quarterly Journal of Economics* 114 (2):533–575. URL `http://qje.oxfordjournals.org/content/114/2/533.short`.

Angrist, J.D. and J.S. Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, Joshua. 2009. "14.32 Econometrics." Unpublished course material.

Wald, Abraham. 1940. "The Fitting of Straight Lines if Both Variables are Subject to Error." *The Annals of Mathematical Statistics* 11 (3):pp. 284–300. URL `http://www.jstor.org/stable/2235677`.

Wooldridge, J.M. 2013. *Introductory econometrics: A modern approach*. South-Western.

[1] All of table III is labeled as "reduced-form estimates." It is common to call both what I have been referring to as the reduced form (the regression of y on z and w) and the first stage (the regression of x on z and w) reduced forms.