Econometrics II: Assignment 3

Due: Thursday, March 6th

1 Binary Variables IV Estimator A researcher wants to study the effect of Malaria nets on child mortality in a developing country. She observes y_i, an indicator of infant death, and x_i, and indicator of whether the household in question had purchased a Malaria net. Consider estimation of the model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

by OLS.

1.1 Show that in this particular setting,

$$\begin{aligned} \hat{\beta} &= \overline{y}_1 - \overline{y}_0 \\ \hat{\alpha} &= \overline{y}_0 \\ \hat{y}_1 &= \overline{y}_1 \\ \hat{y}_0 &= \overline{y}_0, \end{aligned}$$

where $\overline{y}_x = \frac{\sum_{i=1}^n y_i \cdot 1\{x_i = x\}}{\sum_{i=1}^n 1\{x_i = x\}}$ is the sample average of y_i for the subsample where $x_i = x$, and where $\hat{y}_x$ is the fitted value for $x_i = x$.

1.2 Is $\hat{\beta}$ a credible estimate for the causal effect of β? The consideration I am after is *omitted variables*: x_i is likely correlated with other, unobserved traits of households.

1.3 Suppose an experiment had been conducted and Malaria nets had been randomly assigned to households. Suppose perfect compliance with the experiment, thus a household used a Malaria net if, and only if, it was randomized into *treatment*; the other households are *control*. Can you now estimate the causal effect of Malaria nets?

1.4 Suppose now more realistically that compliance is imperfect: Some households discard their malaria nets. And maybe there is a secondary market. (We continue to assume, however, that the true causal effect of a malaria net, if it were used, would be the same across households. This *homogeneous treatment effect* assumption is of course questionable.) However, we assume that having received a malaria net from the experimenter increases the chance that a household uses one. Letting the r.v. z_i denote **receipt** of a Malaria net and x_i **use** of a Malaria net, argue that you can estimate the causal effect of Malaria nets and express the estimator similarly to the simplification in 2.1.

2 Measurement Error Consider the model

$$Y^* = \beta_0 + \beta_1 X^* + \varepsilon,$$

where OLS assumptions hold with important exceptions that I am about to explain. (Assume throughout that moments exist as needed.) The substantive motivation for this exercise are different forms of measurement error. That is, you are invited to think of (Y^*, X^*) as "true" (and causally relevant) quantities but you observe some of them subject to additive measurement error.

2.1 Measurement Error in Outcome You do not observe Y^* but $Y = Y^* + \eta$, where η is i.i.d. and independent from all other random variables with mean 0 and variance σ_η^2. Argue that you can still estimate β_1 by OLS of Y on X^*. What is the estimator's asymptotic distribution?

2.2 Errors-in-Variables You observe Y^*, however you do not observe X^* but $X = X^* + \eta$, where η is just as before. Argue that β_1 can not be estimated by OLS of Y^* on X. Can you characterize the OLS estimator's bias?

2.3 Dual Measurements As in the previous question but in addition, you observe $\tilde{X} = X^* + \nu$ of X^*. Here, ν is i.i.d. and independent from all other random variables with mean 0 and variance σ_ν^2.

Argue that you *can* now estimate β_1. How? Provide an estimator and characterize the estimator's asymptotic distribution.

3. Empirical Exercise The data for this exercise are the "Card" data at `https://www.ssc.wisc.edu/ bhansen/econometrics/`. They belong to the paper Card (1995) which is uploaded.

3.1 Please replicate the column 2SLS(a) in Table 12.1 and the final column of Table 12.2 in Hansen's textbook. Note that the variable `experience` has to be created as `age-educ-6`.

3.2 Add `nearc2` ("grew up near a 2 year college") to the first stage/reduced form equation. Do results change appreciably?

3.3 Estimate the structural equation by TSLS but add instruments `nearc4a,nearc4b,near4ca*age76,near4ca*age76squared/100` (the last two are generated interaction variables whose names should be self-explanatory). Do results change appreciably?

Answer Key

1.1 Let n_x be the number of observations with $x_i = x$, then

$$b = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{Y})}{\sum_{i=1}^{n} X_i(X_i - \overline{X})} = \frac{n_1(\overline{X}_1 - \overline{Y})}{n_1(1 - \overline{X})} = \frac{\overline{Y}_1 - \overline{Y}}{1 - \overline{X}}$$
$$= \frac{\overline{Y}_1 - \left(\frac{n_0}{n}\overline{Y}_0 + \frac{n_1}{n}\overline{Y}_1\right)}{1 - n_1/n} = \frac{(n - n_1)\,\overline{Y}_1 - n_0\overline{Y}_0}{n_0} = \overline{Y}_i - \overline{Y}_0.$$

The other result follows from $\overline{Y} = a + b\overline{X}$, and the fitted values are then easily verified. (This is an example of a saturated regression.)

1.2 No, because x_i is likely correlated with overall conscientousness of a household.

1.3 Now we can regress outcomes on assignment to malaria nets.

1.4 Use assignment as instrumental variable. Note this requires that both assignment and actual use are observed. If only the former is observed, we could try to estimate a so-called intention-to-treat effect but not (without further assumptions) a treatment effect.

2.1 In this case, we can rewrite the model as

$$Y = \beta_0 + \beta_1 X^* + \varepsilon - \eta,$$

and everything is as before after defining $\epsilon = \varepsilon + \eta$. The estimator's asymptotic variance is higher than before because $\text{var}(\epsilon) = \text{var}(\varepsilon) + \text{var}(\eta)$.

2.2 This case is very different. While we can write

$$Y^* = \beta_0 + \beta_1 X + \beta_1(X^* - X) + \varepsilon = \beta_0 + \beta_1 X + \underbrace{\beta_1\eta + \varepsilon}_{\equiv\epsilon},$$

but now the OLS estimator will be biased and inconsistent because $\text{cov}(X, \eta) = \text{cov}(X * +\eta, \eta) = \sigma_\eta$.

Using b_1 for the population projection coefficient, which is consistently estimated by OLS, we can now write

$$b_1 = \frac{\text{cov}(Y^*, X)}{\text{var}(X)} = \frac{\text{cov}(\beta_0 + \beta_1 X^* + \varepsilon, X^* + \eta)}{\text{var}(X^* + \eta)} = \frac{\beta_1 \,\text{var}(X^*)}{\text{var}(X^*) + \text{var}(\eta)},$$

and so the estimator will be biased toward zero ("attenuation bias").

(Note: I used the term "bias" in a loose sense in the question, analogous to the common usage of "omitted variable bias." My main intention was to use assumptions under which the above correctly characterizes the plim. Note also that this famous result depends on specifics of the model; measurement error does not "always" induce attenuation bias.)

2.3 Now we can use one measurement as instrument for the other one.