International School of Economics at TSU

Econometrics 2

Lasha Chochua

Problem Set 9

Instructions: You are encouraged to solve the problems before the recitation. Additionally, you are encouraged to work in groups. It is **not mandatory** to submit solutions unless stated otherwise. However, if you would like to share your solution, I would be happy to review it.

Problem 1: Consider the binary variable version of the fixed effects model discussed in the class except with an additional regressor, $D1_i$; that is, let

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_1 D1_i + \gamma_2 D2_i + \cdots + \gamma_n Dn_i + u_{it}.$$

a. Suppose that $n = 3$. Show that the binary regressors and the "constant" regressor are perfectly multicollinear; that is, express one of the variables $D1_i, D2_i, D3_i$, and $X_{0,it}$ as a perfect linear function of the others, where $X_{0,it} = 1$ for all i, t.

b. Show the result in (a) for general n.

c. What will happen if you try to estimate the coefficients of the regression by OLS?

Solution:

a. Multicollinearity for $n = 3$

- Let $X_{0,it} = 1$, and define $D1_i, D2_i, D3_i$ as:
  - $D1_i = 1$ if $i = 1$, 0 otherwise
  - $D2_i = 1$ if $i = 2$, 0 otherwise
  - $D3_i = 1$ if $i = 3$, 0 otherwise
- Since each individual falls into exactly one category:

$$D1_i + D2_i + D3_i = 1 = X_{0,it}$$

- So:
$$X_{0,it} = D1_i + D2_i + D3_i$$
- Thus, any one of these four regressors can be written as a linear combination of the other three ⇒ **perfect multicollinearity**.

b. Generalization to Arbitrary n

- For general n, define $D1_i, D2_i, \ldots, Dn_i$ as binary dummies:
$$D1_i + D2_i + \cdots + Dn_i = 1 = X_{0,it}$$
- Therefore, the regressors $D1_i, \ldots, Dn_i, X_{0,it}$ are linearly dependent.
- Any one of these variables can be written as a function of the others ⇒ **perfect multicollinearity persists**.

c. Consequence for OLS Estimation

- OLS cannot estimate the model with perfectly collinear regressors.
- The matrix $X'X$ will be **singular** (non-invertible).
- Statistical software will either:
  - Drop one dummy (e.g., omit $D1_i \rightarrow$ baseline group), or

To resolve this: omit the intercept β_0, or drop one dummy regressor.

Problem 2: A researcher believes that traffic fatalities increase when roads are icy and thinks that therefore states with more snow will have more fatalities than other states. Comment on the following methods designed to estimate the effect of snow on fatalities:

a. The researcher collects data on the average snowfall for each state and adds this regressor ($AverageSnow_i$) to the regressions given in Table 10.1.

b. The researcher collects data on the snowfall in each state for each year in the sample and adds this regressor to the regressions.

Solution:

a. Average snow fall does not vary over time, and thus will be perfectly collinear with (and will be controlled for by) the state fixed effect.

b. $Snow_{it}$ does vary with time, and so this method can be used along with state fixed effects.

Problem 3:

a. In the fixed effects regression model, are the fixed entity effects, α_i, consistently estimated as $n \longrightarrow \infty$ with T fixed?

b. If n is large (say, $n = 2000$) but T is small (say, $T = 4$), do you think that the estimated values of α_i are approximately normally distributed? Why or why not?

Solution:

a. $\widehat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ which has variance $\frac{\sigma_u^2}{T}$. Because T is not growing, the variance is not getting small. $\widehat{\alpha}_i$ is not consistent.

b. The average in (a) is computed over T observations. In this case T is small ($T = 4$), so the normal approximation from the CLT is not likely to be very good.

Problem 4: Let $\widehat{\beta}_1^{DM}$ denote the entity-demeaned estimator given in Equation (10.22), and let $\widehat{\beta}_1^{BA}$ denote the "before and after" estimator without an intercept, so that

$$\widehat{\beta}_1^{BA} = \left[\sum_{i=1}^{n}(X_{i2} - X_{i1})(Y_{i2} - Y_{i1})\right] \Big/ \left[\sum_{i=1}^{n}(X_{i2} - X_{i1})^2\right].$$

Show that, if $T = 2$, $\widehat{\beta}_1^{DM} = \widehat{\beta}_1^{BA}$.

Solution:

Start from the **demeaned variables** for $T = 2$:

$$\tilde{X}_{i1} = -\frac{1}{2}(X_{i2} - X_{i1}), \quad \tilde{X}_{i2} = \frac{1}{2}(X_{i2} - X_{i1})$$

$$\tilde{Y}_{i1} = -\frac{1}{2}(Y_{i2} - Y_{i1}), \quad \tilde{Y}_{i2} = \frac{1}{2}(Y_{i2} - Y_{i1})$$

The demeaned estimator is:

$$\widehat{\beta}_1^{DM} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{2}\tilde{X}_{it}\tilde{Y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{2}\tilde{X}_{it}^2}$$

Numerator:

$$\sum_{i=1}^{n}\left(\tilde{X}_{i1}\tilde{Y}_{i1}+\tilde{X}_{i2}\tilde{Y}_{i2}\right)=\sum_{i=1}^{n}\left(\frac{1}{4}(X_{i2}-X_{i1})(Y_{i2}-Y_{i1})+\frac{1}{4}(X_{i2}-X_{i1})(Y_{i2}-Y_{i1})\right)$$

$$=\sum_{i=1}^{n}\frac{1}{2}(X_{i2}-X_{i1})(Y_{i2}-Y_{i1})$$

Denominator:

$$\sum_{i=1}^{n}\left(\tilde{X}_{i1}^{2}+\tilde{X}_{i2}^{2}\right)=\sum_{i=1}^{n}\left(\frac{1}{4}(X_{i2}-X_{i1})^{2}+\frac{1}{4}(X_{i2}-X_{i1})^{2}\right)=\sum_{i=1}^{n}\frac{1}{2}(X_{i2}-X_{i1})^{2}$$

$$\hat{\beta}_{1}^{DM}=\frac{\sum_{i=1}^{n}\frac{1}{2}(X_{i2}-X_{i1})(Y_{i2}-Y_{i1})}{\sum_{i=1}^{n}\frac{1}{2}(X_{i2}-X_{i1})^{2}}=\frac{\sum_{i=1}^{n}(X_{i2}-X_{i1})(Y_{i2}-Y_{i1})}{\sum_{i=1}^{n}(X_{i2}-X_{i1})^{2}}=\hat{\beta}_{1}^{BA}$$

Thus, when $T=2$, the **demeaned estimator equals the before-after estimator**.

Problem 5: You wish to study the effects of unionisation on wages using a panel of N individuals and T time periods. You wish to allow for the following phenomena:

- (a) Unionised firms select the higher ability workers
- (b) Workers with bad productivity shocks join the union sector

a. Set up a suitable model and explain how these phenomena are reflected in your specification.

b. Explain how you would estimate this model and present the estimator. Carefully state any assumptions you make.

Solution:

a. Model Specification with Selection Concerns

The model is:

$$w_{it}=\beta_{1}\text{union}_{it}+x_{it}\gamma+f_{i}+v_{it}$$

where x_{it} includes controls like education and age.

This specification captures: - Unionised firms selecting higher ability workers:

$$\mathbb{E}(f_i \mid \text{union}_{it}) > 0$$

- Workers with bad productivity shocks joining unions:

$$\mathbb{E}(v_{it} \mid \text{union}_{it+j}) < 0 \quad \text{for } j \geq 1$$

b. Estimation Strategy

Use first-differences to eliminate f_i:

$$\Delta w_{it} = \beta_1 \Delta \text{union}_{it} + \Delta x_{it}\gamma + \Delta v_{it}$$

However, Δunion_{it} is endogenous because:

$$\mathbb{E}(\Delta v_{it} \mid \Delta \text{union}_{it}) \neq 0$$

since

$$\mathbb{E}(v_{it-1} \mid \text{union}_{it}) \neq 0$$

Instrumental Variables (IV) Approach

We need an instrument for Δunion_{it}. Under the assumption:

$$\mathbb{E}(\Delta v_{it} \cdot \Delta \text{union}_{it-1}) = 0$$

we can use $\Delta \text{union}_{it-1}$ as a valid instrument.

The IV estimator is:

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = (Z'\tilde{X})^{-1} Z' \Delta W$$

where:

- $\tilde{X}_{it} = [\Delta \text{union}_{it} \quad \Delta x_{it}]$
- $Z_{it} = [\Delta \text{union}_{it-1} \quad \Delta x_{it}]$

Note of caution: If union_{it} is a dummy, then $\mathbb{E}(\Delta\text{union}_{it-1} \cdot \Delta\text{union}_{it}) = 0$ may not hold. Use union_{it-1} as instrument instead.

Problem 6: A common setup for program evaluation with two periods of panel data is the following. Let y_{it} denote the outcome of interest for unit i in period t.

At $t = 1$, no one is in the program. At $t = 2$, some units are in the control group and others are in the experimental group. Let $prog_{it}$ be a binary indicator equal to one if unit i is in the program in period t; by design, $prog_{i1} = 0$ for all i.

An unobserved effects model without additional covariates is:

$$y_{it} = \theta_1 + \theta_2 d2_t + \delta_1 prog_{it} + c_i + u_{it}, \qquad \mathbb{E}(u_{it} \mid prog_{i2}, c_i) = 0$$

where $d2_t = 1$ if $t = 2$, and 0 otherwise; c_i is the unobserved individual effect.

a. Explain why including $d2_t$ is important in these contexts. In particular, what problems might be caused by leaving it out?

b. Why is it important to include c_i in the equation?

c. Using the first differencing method, show that:

$$\hat{\theta}_2 = \overline{\Delta y}_{control}, \quad \hat{\delta}_1 = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}$$

where $\overline{\Delta y}_{control}$ is the average change in y for units with $prog_{i2} = 0$, and $\overline{\Delta y}_{treat}$ is the average change for units with $prog_{i2} = 1$. This shows that $\hat{\delta}_1$, the **difference-in-differences estimator**, arises from an unobserved effects panel model.

d. Write down the extension of the model for T time periods.

Solution:

a. Importance of $d2_t$

Including the aggregate time effect, $d2_t$, can be very important. Without it, we must assume that any change in average y over the two time periods is due to the program, and not to external factors. For example, if y_{it} is the unemployment rate for city i at time t, and $prog_{it}$ denotes a job creation program, we want to be sure that we account for the fact that the general economy may have worsened or improved over the period. If $d2_t$ is omitted, and $\theta_2 < 0$ (an improving economy, since unemployment has fallen), we might attribute a decrease in unemployment to the job creation program, when in fact it had nothing to do with it. For general T, each time period should have its own intercept (otherwise the analysis is not entirely convincing).

b. Importance of c_i

The presence of c_i allows program participation to be correlated with unobserved individual heterogeneity, something crucial in contexts where the experimental group is not randomly assigned. Two examples are when individuals "self-select" into the program and when program administrators target specific groups that may benefit more or less from the program.

c. First Difference and Estimators

If we first difference the equation, use the fact that $prog_{i1} = 0$ for all i, $d2_1 = 0$, and $d2_2 = 1$, we get

$$y_{i2} - y_{i1} = \theta_2 + \delta_1 prog_{i2} + u_{i2} - u_{i1},$$

or

$$\Delta y_i = \theta_2 + \delta_1 prog_{i2} + \Delta u_i.$$

Now, the FE (and FD) estimates of θ_2 and δ_1 are just the OLS estimators from this equation (on cross section data). From basic two-variable regression with a dummy independent variable, $\hat{\theta}_2$ is the average value of Δy over the group with $prog_{i2} = 0$ – that is, the control group. Also, $\hat{\theta}_2$ and $\hat{\delta}_1$ is the average value of Δy over the group with $prog_{i1} = 1$ – that is, the treatment group. Thus, as asserted, we have

$$\hat{\theta}_2 = \overline{\Delta y}_{control}, \quad \hat{\delta}_1 = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}.$$

If we did not include the $d2_t$, $\hat{\delta}_1 = \overline{\Delta y}_{treat}$, the average change of the treated group. The demonstrates the claim in part b that without the aggregate time effect any change in the average value of y for the treated group is attributed to the program. Differencing and averaging over the treated group allows program participation to depend on time-constant unobservables affecting the level of y, but that does not account for external factors that affect y for everyone.

d. Extension for T Periods

In general, for T time periods we have

$$y_{it} = \theta_1 + \theta_2 d2_t + \theta_3 d3_t + \cdots + \theta_r dT_t + \delta_1 prog_{it} + c_i + u_{it};$$

that is, we have separate year intercepts, an unobserved effect c_i, and the program indicator.

Problem 7: Consider the model:

$$y = z_1\beta + w\alpha + \varepsilon$$

where:

- $E(z\varepsilon) = 0$
- $z = (z_1, z_2)$ is a vector of exogenous variables
- w is endogenous: $E(w\varepsilon) \neq 0$

Suppose we estimate (β, α) using the following two-step procedure:

Step 1: Regress w on z_2 and obtain the fitted values $\widehat{w}$.
Step 2: Regress y on $(z_1, \widehat{w})$ and obtain $(\hat{\beta}, \hat{\alpha})$.

(a) Will $(\hat{\beta}, \hat{\alpha})$ be generally consistent? Show.

(b) When will $(\hat{\beta}, \hat{\alpha})$ be consistent?

Solution

(a)

Use the reduced form for w:

$$w = z_2\gamma + \nu$$

Estimate γ by OLS. Then $E(z_2\nu(\hat{\gamma})) = 0$ by construction,
but it is possible that $E(z_1\nu(\hat{\gamma})) \neq 0$.

Then:

$$\begin{aligned} y &= z_1\beta + w\alpha + \varepsilon \\ &= z_1\beta + (z_2\hat{\gamma} + \hat{\nu})\alpha + \varepsilon \\ &= z_1\beta + \hat{w}\alpha + v, \end{aligned}$$

where $\hat{w} = z_2\hat{\gamma}$ and $v = \hat{\nu}\alpha + \varepsilon$.

Since $E(z_1'\widehat{\nu}) \neq 0$, it follows that $E(z_1'v) \neq 0$

$\widehat{\beta}$ is **inconsistent** due to correlation between z_1 and the error term.

(b)

If:

$$E(w \mid z_1, z_2) = E(w \mid z_2),$$

then $E(z_1'\nu) = 0$, and the estimates will be **consistent**.

Problem 8: Suppose you wish to estimate β in:

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

(a) Derive the consequences for the OLS estimator if y is measured with error that is independent of x.

(b) Instead of measuring x you measure x^* where $x_i^* = x_i + \epsilon_i$ and ϵ_i is a measurement error which is independent across individuals and independent of x. Show that the OLS estimator converges asymptotically to $\delta\beta$ where $0 \leq \delta \leq 1$. Explain the implication of this result for estimating the elasticity of hours worked with respect to wages when wages are measured with iid errors.

If x is measured with error, so that we observe x^* where

$$x_i^* = x_i + \epsilon_i$$

where x is independent of ϵ.

Solution

(a)

The OLS estimator of β is:

$$\beta^{OLS} = \frac{\text{cov}(x_i, y_i)}{\text{var}(x_i)}$$

If y is measured with error, we observe:

$$y_i^* = y_i + \epsilon_i$$

where x is independent of ϵ.

Then:

$$\begin{aligned}\beta^{OLS} &= \frac{\text{cov}(x_i, y_i^*)}{\text{var}(x_i)} \\ &= \frac{\text{cov}(x_i, \alpha + \beta x_i + u_i + \epsilon_i)}{\text{var}(x_i)} \\ &= \beta \frac{\text{var}(x_i)}{\text{var}(x_i)} + \frac{\text{cov}(x_i, u_i)}{\text{var}(x_i)} \\ &= \beta + \frac{\text{cov}(x_i, u_i)}{\text{var}(x_i)}\end{aligned}$$

If $\text{cov}(x_i, u_i) = 0$, the estimator is consistent.
Thus, OLS is consistent despite measurement error in y if the error is independent of x.

(b) Now suppose we measure x with error:

$$x_i^* = x_i + \epsilon_i$$

where ϵ_i is independent across individuals and of x_i.

The data model becomes:

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

Then the OLS estimator is:

$$\begin{aligned}\beta^{OLS} &= \frac{\text{cov}(x_i^*, y_i)}{\text{var}(x_i^*)} \\ &= \frac{\text{cov}(x_i^*, \alpha + \beta x_i^* + (u_i - \beta \epsilon_i))}{\text{var}(x_i^*)} \\ &= \frac{\beta \text{var}(x_i^*) - \beta \text{cov}(x_i^*, \epsilon_i)}{\text{var}(x_i^*)} \\ &= \beta \left(1 - \frac{\text{cov}(x_i + \epsilon_i, \epsilon_i)}{\text{var}(x_i + \epsilon_i)}\right) \\ &= \beta \left(1 - \frac{\text{var}(\epsilon_i)}{\text{var}(x_i) + \text{var}(\epsilon_i)}\right) \\ &= \beta \left(1 - \frac{\sigma_\epsilon^2}{\sigma_x^2 + \sigma_\epsilon^2}\right)\end{aligned}$$

Conclusion: OLS is **biased toward zero** $\Rightarrow$ **attenuation bias**.
Elasticity of hours w.r.t. wages will be **underestimated** if wages are measured with error.