Introductory Econometrics I – Spring 2024

Problem Set 4 – Due date: June 2

Last updated: May 20, 2024

Notes: Please submit a single PDF file containing your answers to all questions on Web-learning. For empirical questions, original codes and complete results need to be attached.

1. (**Heterskedasticity and Intra-cluster Correlation**) Consider a model at the employee level

$$y_{i,e} = \beta_0 + \beta_1 x_{i,e} + f_i + v_{i,e}$$

where the unobserved variable f_i is a "firm effect" to each employee at a given firm i. The error term $v_{i,e}$ is specific to employee e at firm i. The composite error is $u_{i,e} = f_i + v_{i,e}$. The variables $(v_{i,e}, x_{i,e})$ are i.i.d. over both i and e, and f_i are i.i.d. over i. Let n_i denote the number of employees in firm i (viewed as fixed). Assume that $\mathbb{V}[f_i] = \sigma_f^2$, $\mathbb{V}[v_{i,e}] = \sigma_v^2$, f_i and $v_{i,e}$ are uncorrelated, and $\mathbb{E}[u_{i,e}|x_{i,1}, \cdots, x_{i,n_i}] = 0$.

(a) Suppose that we only have data averaged at the "firm level", that is,

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \bar{u}_i,$$

where $\bar{y}_i = \frac{1}{n_i}\sum_{e=1}^{n_i} y_{i,e}$, $\bar{x}_i = \frac{1}{n_i}\sum_{e=1}^{n_i} x_{i,e}$, and $\bar{u}_i = \frac{1}{n_i}\sum_{e=1}^{n_i} u_{i,e}$. Let $\hat{\beta}_1$ be the OLS estimator of β_1 from regressing $\bar{y}_i$ on $\bar{x}_i$.

- Do you think $\hat{\beta}_1$ is unbiased? Is it consistent? (Hint: by law of iterated expectation, $\mathbb{E}[\bar{u}_i|\bar{x}_i] = \mathbb{E}[\mathbb{E}[\bar{u}_i|x_{i,1}, \cdots, x_{i,n_i}]|\bar{x}_i]$.)
- Show that $\mathbb{V}[\bar{u}_i] = \sigma_f^2 + \sigma_v^2/n_i$.
- Suppose σ_f^2 and σ_v^2 are known. Propose a weighted least squares estimator $\tilde{\beta}_1$ of β_1 that corrects for heteroskedasticity. Which estimator do you think is more efficient, $\hat{\beta}_1$ or $\tilde{\beta}_1$?

(b) Now, suppose that we can observe the employee-level data $(y_{i,e}, x_{i,e})$, but recall f_i is *unobserved*.

- Do you think the OLS estimators from regressing $y_{i,e}$ on $x_{i,e}$ are unbiased? Are they consistent?
- Show that $\mathbb{V}[u_{i,e}] = \sigma_f^2 + \sigma_v^2$, $Cov[u_{i,e}, u_{j,g}] = 0$ for $i \neq j$ and any $1 \leq e \leq n_i$ and $1 \leq g \leq n_j$, and $Cov[u_{i,e}, u_{i,g}] = \sigma_f^2$ for $e \neq g$.
- Given your answer to the previous question, what kind of standard errors do you prefer to report for the OLS regression of $y_{i,e}$ on $x_{i,e}$?
- Suppose that we transform the model to "eliminate" the firm effect f_i:

$$\tilde{y}_{i,e} = \beta_1 \tilde{x}_{i,e,1} + \tilde{u}_{i,e}$$

where $\tilde{y}_{i,e} = y_{i,e} - \bar{y}_i$, $\tilde{x}_{i,e} = x_{i,e} - \bar{x}_{i,e}$, and $\tilde{u}_{i,e} = u_{i,e} - \bar{u}_i$. Show that

$$\mathbb{V}[\tilde{u}_{i,e}] = (1 - n_i^{-1})\sigma_v^2, \quad Cov[\tilde{u}_{i,e}, \tilde{u}_{i,g}] = -n_i^{-1}\sigma_v^2, \quad \text{for } e \neq g.$$

2. (**Measurement error with a binary regressor**) The true population regression model is

$$y = \beta_0 + \beta_1 x^* + u.$$

Let Assumptions MLR.1-MLR.4 hold for this model, and in particular $\mathbb{E}[u|x^*] = 0$. x^* is binary (taking values of 0 and 1 only), but it is unobserved. Suppose that we have a (possibly) noisy measurement of x^*, denoted by x, which is also binary. Assume $\mathbb{E}[u|x^*, x] = \mathbb{E}[u|x^*]$.

(a) Show that

$$\mathbb{E}[y|x=1] = (\beta_0+\beta_1)\mathbb{P}(x^*=1|x=1) + \beta_0\mathbb{P}(x^*=0|x=1),$$
$$\mathbb{E}[y|x=0] = (\beta_0+\beta_1)\mathbb{P}(x^*=1|x=0) + \beta_0\mathbb{P}(x^*=0|x=0).$$

Hint: Use law of iterated expectation, e.g., $\mathbb{E}[y|x=1] = \mathbb{E}[y|x=1,x^*=1]\times\mathbb{P}(x^*=1|x=1) + \mathbb{E}[y|x=1,x^*=0]\times\mathbb{P}(x^*=0|x=1)$.

(b) Let $\hat{\beta}_1$ be the OLS estimator of β_1 from regressing y on x. Show that

$$\hat{\beta}_1 \to_{\mathbb{P}} \beta_1\Big(\mathbb{P}(x^*=0|x=0) - \mathbb{P}(x^*=0|x=1)\Big)$$

Hint: Use part (a) above, law of large numbers and the conclusion of Question 1, part (b) in Problem Set 1. That is, the OLS estimator $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$ where $\bar{y}_1$ is the average of y_i over the part of the sample with $x_i = 1$, and $\bar{y}_0$ is the average of y_i over the part of the sample with $x_i = 0$.

(c) Suppose the probability of the observed measurement x being misclassified is zero. That is, $\mathbb{P}(x^*=0|x=1) = \mathbb{P}(x^*=1|x=0) = 0$. What is the probability limit of $\hat{\beta}_1$ in this case? Explain the intuition for your result.

(d) Suppose the probability of the observed measurement x being misclassified is one. That is, $\mathbb{P}(x^*=0|x=1) = \mathbb{P}(x^*=1|x=0) = 1$. What is the probability limit of $\hat{\beta}_1$ in this case? Explain the intuition for your result.

(e) Suppose the probability of the observed measurement x being misclassified is 0.5. That is, $\mathbb{P}(x^*=0|x=1) = \mathbb{P}(x^*=1|x=0) = 0.5$. What is the probability limit of $\hat{\beta}_1$ in this case? Explain the intuition for your result.

3. (**Job Training Program**) Use the data in `jtrain98.dta`. The outcome we would like to explain, $y = earn98$, is a worker's earning in 1998. The key explanatory variable (or "treatment"), $d = train$, is a binary variable indicating whether a worker was in a job training program in 1997. We want to evaluate the effectiveness of the job training program in improving labor market earnings. We consider this problem in the potential outcomes framework discussed in class.

(a) Clearly explain what the potential outcomes $y_i(1)$ and $y_i(0)$ represent in this context and how they are related to the observed outcome y_i.

(b) Run the simple regression of y on d. Precisely interpret the coefficient on d. Is it statistically significant? Do you think it gives a proper estimate of the causal effect of the job training program on workers' earnings?

(c) Run regression of d on $earn96$, $educ$, age and $married$ (the meanings of these variables are explained in corresponding variable labels in the dataset). Use results from this regression to explain your findings in part (b).

(d) Run regression of y on $earn96$, $educ$, age and $married$ using the subsample $d = 1$, and obtain the fitted value $\hat{y}_i(1)$ for every worker in the whole sample. Then, run the same regression using the subsample $d = 0$, and the obtain the fitted value $\hat{y}_i(0)$ for every worker in the whole sample. Compute

$$\hat{\tau}_{\texttt{ATE}} = \frac{1}{n}\sum_{i=1}^{n}[\hat{y}_i(1) - \hat{y}_i(0)],$$

where n denotes the full sample size. Note that this is just the average treatment effect estimator under selection on observables we discussed in class. Compare it with the coefficient on d from the regression of y on d, $earn96$, $educ$, age and $married$ using the full sample.

4. (**Data exercise: heteroskedasticity**) For this question, use the data `gpa.dta`. We also provide the `Stata` code `ps4-hr.do` for your reference. (Please feel free to write your own code if you wish.) We are interested in the factors determining college GPA. The data contain the following variables: *colgpa* is college GPA, *hsgpa* is high school GPA, *act* is college entrance exam score, *skipped* denotes the total number of lectures missed during the semester, and *pc* is a dummy variable with a value of 1 if the student owns a personal computer and 0 otherwise.

   (a) Use OLS to estimate the model:

   $$colgpa_i = \beta_0 + \beta_1 hsgpa_i + \beta_2 act_i + \beta_3 skipped_i + \beta_4 pc + u_i.$$

   Obtain the OLS residuals.

   (b) We want to estimate the model using WLS. Estimate the weights by regressing $log(\hat{u}_i^2)$ on all independent variables. Get the fitted value $\hat{g}_i$, and $\hat{h}_i = exp(\hat{g}_i)$. Then use $\hat{h}_i$ as the weight.

   (c) In the WLS estimation in the last question, obtain heteroskedasticity-robust standard errors. In other words, allow for the fact that the variance function estimated might be *mis-specified*. Do the standard errors change much?

   (d) Finally, estimate the model using OLS and report the heteroskedasticity-robust standard errors. We want to test the null hypothesis $\beta_4 = 0$ against the alternative $\beta_4 \neq 0$. Can you reject the null hypothesis at the 5% significance level using 1) OLS estimator with robust standard error 2) WLS estimator using conventional standard error and 3) WLS estimator using robust standard error?

5. (**Numeric simulation: spatial correlation**) In this question, we illustrate the importance to clustering the standard errors when there is spatial correlation. We also provide the `Stata` code `ps4-cr.do` for your reference. (Please feel free to write your own code if you wish.)

   (a) Set the number of observations to 1000. Generate a unique *id* for each observation. Generate x from the standard uniform distribution, and u from a standard normal distribution. Generate $y_i = 3x_i + u_i$.

   (b) Regress y on 1 and x. Get the coefficients and standard errors.

   (c) Replicate each observation with 3 extra copies (so in total you have 4 copies of each observation). Regress y on 1 and x using the new sample and report the conventional standard errors. Compare the standard errors with the last question. What do you find?

   (d) Regress y on 1 and x using the new sample and cluster the standard errors at the *id* level. What do you find?

The following two questions are *optional* (just in case we may not be able to finish all discussions on instrumental variables before the due date). But remember IV methods are *required* and may be *tested* in the final exam.

6. (**Wald Estimator**) Consider the simple regression model:

$$y = \beta_0 + \beta_1 x + u.$$

We are concerned that x is endogenous, that is, $\mathbb{E}[u|x] \neq 0$. To solve the endogeneity issue, we find an instrumental variable z which is *binary* (taking values of 0 and 1 only). Suppose that we have a random sample $\{(y_i, x_i, z_i) : 1 \leq i \leq n\}$. The size of the group with $z_i = 1$ is $n_1 = \sum_{i=1}^n z_i$, and the size of the group with $z_i = 0$ is $n_0 = \sum_{i=1}^n (1 - z_i)$. Show that the instrumental variable estimator can be written as

$$\hat{\beta}^{IV} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}, \quad \text{where}$$

$$\bar{y}_1 = \frac{1}{n_1} \sum_{i:z_i=1} y_i, \quad \bar{y}_0 = \frac{1}{n_0} \sum_{i:z_i=0} y_i, \quad \bar{x}_1 = \frac{1}{n_1} \sum_{i:z_i=1} x_i, \quad \bar{x}_0 = \frac{1}{n_0} \sum_{i:z_i=0} x_i.$$

In words, $\bar{y}_0$ 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 where $\bar{y}_1$ and $\bar{x}_1$ are the sample averages of y_i and x_i over the part of the sample with $z_i = 1$.

7. (**Return on Education**) Card (1995) use a dummy variable indicating whether there is a college near one's residence as an instruments to estimate the impact of education on wage.[1] We want to replicate his results using the data set `card.dta`. The data set includes the following variables:

- `lwage`: the natural log of wage
- `educ`: years of education
- `nearc4`: a dummy variable indicating whether there is a college near one's residence (1 = yes, 0 = no)
- `exper,exper2`: work experience and its square
- `smsa`: whether lives in urban areas (1 = urban, 0 = rural)
- `motheduc`: mother's years of education

Use the data to answer the following questions (you need to write your own codes)

(a) Estimate the model using OLS (assume heteroskedasticity and report robust standard errors). What's the coefficient and standard error of `educ`?

$$\texttt{lwage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{exper} + \beta_3 \texttt{exper2} + \beta_4 \texttt{smsa} + \beta_5 \texttt{motheduc} + u.$$

(b) We want to use `nearc4` as an IV for `educ`. Use regression to determine whether `nearc4` satisfies the instrument relevance condition.

(c) Do you think `nearc4` is a valid IV? Briefly discuss your answer.

(d) Use `nearc4` as the instrument for `educ`, and estimate the model using 2SLS. Assume heteroskedasticity and report robust standard errors. What's the coefficient and standard error of `educ`?

(e) Compare the OLS estimates and the IV estimates of β_1. Which is larger? Which has a larger standard error?

[1] Card, David. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling". Published in "Aspects of Labour Economics: Essays in Honour of John Vanderkamp", edited by Louis Christofides, E. Kenneth Grant and Robert Swindinsky. University of Toronto Press, 1995.