Introductory Econometrics I – Final Exam

Yingjie Feng

School of Economics and Management

Tsinghua University

June 11, 2023

Contents

Notes:

- Duration of examination: **120 minutes**.
- Please write your name and student ID clearly on the first page of the answer book.
- Use the last page of this exam question book as the scratch paper.
- Please do not open the exam paper until the proctors ask you to do so.
- Please answer *all* questions. Feel free to use either English or Chinese.
- Answers without proper justification will *not* receive (partial) credit.
- Please turn in the exam question book and your answer book at the end of the exam.

1 Question 1 (25 points)

Consider the population model

$$y = \beta_0 + \beta_1 x^* + u, \quad \mathbb{E}[u|x^*] = 0, \ \mathbb{V}[x^*] = \sigma^2_{x^*}.$$

x^* is *unobserved*, but we have k measurements of x^*:

$$x_j = x^* + e_j, \quad \mathbb{C}ov[u, e_j] = \mathbb{E}[e_j] = 0, \ \mathbb{V}[e_j] = \sigma^2_e, \qquad 1 \le j \le k.$$

There is a random sample $\{(y_i, x_{i1}, \cdots, x_{ik}) : 1 \le i \le n\}$.

1. Let $\hat{\beta}_1$ be the OLS estimator from regressing y on x_1 (the first measurement of x^*).

   (a) [5 points] Show that $\text{plim}(\hat{\beta}_1) = \beta_1 + \frac{\mathbb{C}ov[x_1, u-\beta_1 e_1]}{\mathbb{V}[x_1]}$ ("plim" denotes the probability limit).

   - **Solution**: $\text{plim}(\hat{\beta}_1) = \beta_1 + \frac{\text{plim}\frac{1}{n}\sum_{i=1}^n (x_{i1}-\bar{x}_1)(v_i-\bar{v})}{\text{plim}\frac{1}{n}\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2} = \beta_1 + \frac{\mathbb{C}ov[x_1,v]}{\mathbb{V}[x_1]}$ for $v = u - \beta_1 e_1$.

   (b) [5 points] Assume that $\mathbb{C}ov[e_1, x^*] = -\sigma^2_e$. Give the explicit expression of $\text{plim}(\hat{\beta}_1)$.

   - **Solution**: Note that $\mathbb{C}ov[x_1, u - \beta_1 e_1] = -\beta_1 \mathbb{C}ov[x_1, e_1] = -\beta_1(\mathbb{C}ov[x^*, e_1] + \mathbb{V}[e_1]) = 0$. So $\hat{\beta}_1$ is consistent for β_1.

   (c) [5 points] Assume that $\mathbb{C}ov[e_1, x^*] = 0$. Give the explicit expression of $\text{plim}(\hat{\beta}_1)$.

   - **Solution**: $\mathbb{V}[x_1] = \sigma^2_{x^*} + \sigma^2_e$ and $\mathbb{C}ov[x_1, u - \beta_1 e_1] = 0 - \beta_1 \sigma^2_e$. So $\text{plim}(\hat{\beta}_1) = \beta_1 \sigma^2_{x^*}/(\sigma^2_{x^*} + \sigma^2_e)$.

   (d) [5 points] Suppose that x^* refers to the *exact* years of schooling, while x_1 is the highest grade *completed*. For example, if $x^* = 12.5$, then $x_1 = 12$. Do you think the assumption $\mathbb{C}ov[e_1, x^*] = 0$ in part (c) is appropriate in this context? Explain why.

   - **Solution**: No. The measurement error $e_1 = x_1 - x^* = \lfloor x^* \rfloor - x^*$, which is correlated with x^*.

2. [5 points] Assume $\mathbb{C}ov[e_j, x^*] = 0$ for all $j = 1, \cdots, k$, and $\mathbb{C}ov[e_j, e_l] = 0$ for any $j \ne l$. Let $\tilde{\beta}_1$ be the OLS estimator of β_1 from regressing y on z where $z = \frac{1}{k}\sum_{j=1}^k x_j$ is the average of the k measurements of x^*.

   Give the explicit expression of $\text{plim}(\tilde{\beta}_1)$. How does $\text{plim}(\tilde{\beta}_1)$ change as k grows? [Hint: Derive the variance of z and do similar calculations as for part 1(c).]

- **Solution**: By assumption, $z = x^* + \frac{1}{k}\sum_{j=1}^{k} e_j$ and $\mathbb{V}[z] = \sigma^2_{x^*} + \frac{\sigma^2_e}{k}$. $\text{plim}(\tilde{\beta}_1) = \beta_1 \sigma^2_{x^*}/(\sigma^2_{x^*} + \sigma^2_e/k)$. As k grows, it goes to β_1.

2 Question 2 (25 points)

There is a dataset for young men who have at least one arrest prior to 1986. The following variables are available:

- *arr*86: a binary variable equal to 1 if a man was arrested during 1986, and 0 otherwise.
- *pcnv*: the proportion of prior arrests that led to a conviction
- *tottime*: months spent in prison since age 18 prior to 1986
- *qemp*86: number of quarters (0 to 4) that the man was legally employed in 1986
- *black*: a binary variable equal to 1 if a man is black, and 0 otherwise

We obtain the following OLS regression results:

$$\widehat{arr86} = 0.384 - 0.158\, pcnv - 0.001\, tottime - 0.032\, qemp86 + 0.143\, black$$

Throughout this question, you can give answers that may involve sums, products, quotients or square root of known values (i.e., you do not have to actually calculate a value).

1. [5 points] How do you interpret the coefficient of *black*?

   - **Solution**: Holding other factors fixed, the probability of arrest for black young men is 14.3 percentage points higher than that for non-black people.

2. [5 points] What is the effect of increasing *pcnv* from 0.25 to 0.75?

   - **Solution**: With other factors fixed, increasing the proportion of conviction decreases the probability of arrest by 7.9 percentage points.

3. [5 points] Now, we include the quadratic term $pcnv^2$ into the regression and obtain

   $$\widehat{arr86} = 0.359 + 0.150\, pcnv - 0.324\, pcnv^2 - 0.002\, tottime - 0.028\, qemp86 + 0.140\, black$$

   Again, what is the effect of increasing *pcnv* from 0.25 to 0.75?

- **Solution**: With other factors fixed, the effect of increasing *pcnv* from 0.25 to 0.75 is $(0.15 * 0.75 - 0.324 * 0.75^2) - (0.15 * 0.25 - 0.324 * 0.25^2) = -0.087$, i.e., decreasing the probability of arrest by 8.7 percentage points.

4. [5 points] Still consider the regression in part 3. We want to test whether the effect of *pcnv* on *arr*86 is constant across different levels of *pcnv*. Explain how to do this test.

   - **Solution**: The marginal effect is heterogeneous across *pcnv* if the quadratic term has a nonzero coefficient. So do a t-test for the coefficient of $pcnv^2$. If the t statistic is greater than the critical value for a desired significance level (say, 5%), then we reject the null hypothesis that the coefficient of $pcnv^2$ equals 0, and conclude the effect of *pcnv* is heterogeneous.

5. [5 points] Which standard errors do you prefer to report in this analysis? Explain why.

   - **Solution**: The conditional variance in this model is $p(x)(1-p(x))$ where $p(x)$ denotes the response probability. Thus, it is better to report heteroskedasticity-robust standard error.

3 Question 3 (50 points)

Wei and Zhang (2011) provide one possible explanation for the high saving rate in China: as the sex ratio rises, parents with a son raise their savings to improve their son's relative attractiveness for marriage. So consider the following household-level regression:

$$\log(sav_{i,g}) = \beta_0 + \beta_1 sexrat_g + \beta_2 inc_{i,g} + u_{i,g}, \quad i = 1, \cdots, n_g,\ g = 1, \cdots, G, \tag{1}$$

where

- $sav_{i,g}$: saving rate (i.e., savings/income) of household i in county g, measured in decimals
- $sexrat_g$: local sex ratio in county g (the number of men per woman in the premarital cohort)
- $inc_{i,g}$: income of household i in county g

Here g indexes counties and i indexes households. There are n_g households in county g, and in total we have G counties. The sample consists of three-person families only (parents with one child).

Let $\hat{\beta}_j$ be the OLS estimate of β_j for $j = 0, 1, 2$.

1. [5 points] If we measure $sav_{i,g}$ in percentage rather than in decimals (i.e., multiply the current $sav_{i,g}$ by 100), how does this change the OLS estimates $\hat{\beta}_j$'s? Does it change the standard error of $\hat{\beta}_1$?

   - **Solution**: This transformation only adds log(100) to the intercept estimate, and all other slope estimates do not change. The standard error of $\hat{\beta}_1$ does not change as well.

2. [5 points] Suppose that the OLS estimate $\hat{\beta}_1 = 1.34$. What is the effect of increasing the local sex ratio by 0.1?

   - **Solution**: With other factors fixed, if the sex ratio increases by 0.1, the saving rate $sav_{i,g}$ increases by 13.4%.

3. [10 points] Suppose that we also have a dummy variable $son_{i,g}$, which equals 1 if the child in the family is a boy and equals 0 otherwise. Explain how to use this information to examine whether the impact of sex ratio on saving rates for families with a son differ from that for families with a daughter. (You must provide details on estimation and testing procedures.)

   - **Solution**: Run the following regression (it is fine to add the interaction only)

   $$\log(sav_{i,g}) = \beta_0 + \beta_0' \, son_{i,g} + \beta_1 \, sexrat_g + \beta_1' \, sexrat_g \times son_{i,g} + \beta_2 \, inc_{i,g} + e_{i,c}$$

   Then, test the null hypothesis that $\beta_1' = 0$. If the t test statistic is larger than a certain critical value, reject the null and conclude there is a significant difference between families with a son and families with a daughter in terms of the impact of the sex ratio.

4. [5 points] We believe $u_{i,g} = v_g + e_{i,g}$, where v_g is some (unobserved) county-specific random shock that is i.i.d. over g, and $e_{i,g}$ is a household-specific error that is i.i.d. over i and g. Explain whether the conventional standard errors for OLS are appropriate. If not, explain precisely what standard errors should be reported.

   - **Solution**: The existence of the county-specific shock that is common to all households in the same county may lead to the within-county correlation of errors. Therefore, it is

better to report robust standard errors clustered at the county level.

5. [5 points] Still consider the setup in part 4. One concern is that the unobserved errors v_g might be correlated with $sexrat_g$, making the OLS estimate biased. A researcher proposes to consider the regression of $\log(sav_{i,g})$ on sex_g, $inc_{i,g}$ and a list of dummies $d_g(g')$'s:

$$\log(sav_{i,g}) = \beta_0 + \beta_1\, sexrat_g + \beta_2\, inc_{i,g} + \sum_{g'=1}^{G} v_{g'}\, d_g(g') + e_{i,g},$$

where each $d_g(g')$ is a dummy variable equal to 1 if the household belongs to county g' and equal to 0 otherwise, and each $v_{g'}$ is treated as a parameter on $d_g(g')$.
Do you think this regression can identify β_1 (uniquely determine β_1) ? What if we drop β_0? Explain your answers. [Hint: note that $sexrat_g$ varies across g only.]

   - **Solution**: No. The d_g's are perfectly collinear with $sexrat_g$. Note that even β_0 is dropped, this issue still exists.

6. [10 points] Still consider the regression of $\log(sav_{i,g})$ on sex_g and $inc_{i,g}$ in (1). To deal with the potential endogeneity, one proposes an IV for $sexrat_g$: monetary penalty pen_g for violating the family planning policy ("one-child policy"[1]), which was set by the local government.
What do you think is the argument for pen_g is relevant for explaining $sexrat_g$? Can you think of one argument by which pen_g is correlated with $u_{i,g}$ and thus the exogeneity requirement for IV is violated?

   - **Solution**: Open questions.

     With a larger amount of penalty, families are more likely to do child-sex selection. Boys are more preferred in tradition, which increases the sex ratio.

     In the previous regression, we did not control for the social security involvement. With better social security coverage, people save less. On the other hand, poor local government is likely to set a larger penalty on violating the policy and are usually unable to provide good social security coverage as well.

7. [10 points] Suppose that in addition to pen_g in part 6, we have another potential IV for $sexrat_g$: a dummy variable $extra_g$ equal to 1 if there is an extra penalty for having *more*

[1] For simplicity, just think of this policy as a strict birth quota: each family can only has one child.

than two children and equal to 0 otherwise.

Explain in detail how to test whether the two IVs, pen_g and $extra_g$, are relevant for $sexrat_g$.

- **Solution**: Run regression

$$sexrat_g = \pi_0 + \pi_1 pen_g + \pi_2 extra_g + \pi_3 inc_{i,g} + \epsilon_{i,g}$$

  Test $H_0 : \pi_1 = \pi_2 = 0$ by a (robust) F test. If F is very large, e.g., > 10, reject the null that IVs are weak.