Introductory Econometrics I – Final Exam

Yingjie Feng

School of Economics and Management

Tsinghua University

June 12, 2022

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.
- On-campus students: please turn in the exam question book and your answer book at the end of the exam.
- Online students: please upload your answers to the Web Learning Page no later than **21:15** on Jun 12.

1 Question 1 (20 points)

Consider the population model

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

$$x = \pi z + v, \qquad \pi \neq 0$$

where $\mathbb{E}[u|x] \neq 0$, $\mathbb{E}[u|z] = 0$, $\mathbb{E}[v|z] = 0$. In addition, assume $\mathbb{E}[z] = 0$, $\mathbb{E}[z^2] > 0$ and $\mathbb{E}[z^3] \neq 0$. There is a random sample $\{(y_i, x_i, z_i) : 1 \leq i \leq n\}$.

1. [5 points] Show that $\mathbb{E}[u] = 0$, $\mathbb{E}[zu] = 0$ and $\mathbb{C}ov[z, x] \neq 0$. Use these facts to write down the sample-analogue estimating equations from which you can obtain the (method-of-moment) IV estimators $\hat{\beta}_0^{IV}$ and $\hat{\beta}_1^{IV}$.

   - **Solution**: By law of iterated expectation, $\mathbb{E}[u] = \mathbb{E}[\mathbb{E}[u|z]] = 0$, $\mathbb{E}[zu] = \mathbb{E}[\mathbb{E}[zu|z]] = \mathbb{E}[z\mathbb{E}[u|z]] = 0$ and $\mathbb{C}ov[x, z] = \pi\mathbb{V}[z] + \mathbb{C}ov[v, z] \neq 0$.

     The estimating equations are

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_i) = 0$$

$$\sum_{i=1}^{n} z_i(y_i - \hat{\beta}_0^{IV} - \hat{\beta}_1^{IV} x_i) = 0$$

2. [5 points] Do you think z^2 is also a valid IV for x? [Hint: relate your discussion to the exogeneity and relevance conditions, and apply law of iterated expectation.]

   - **Solution**: By assumption that $\mathbb{E}[u|z] = 0$, $\mathbb{E}[z^2u] = \mathbb{E}[z^2\mathbb{E}[u|z]] = 0$, so the exogeneity condition holds. Regarding the relevance, $\mathbb{C}ov[z^2, x] = \mathbb{E}[z^2x] = \pi\mathbb{E}[z^3] + \mathbb{E}[z^2v] = \pi\mathbb{E}[z^3] + 0$. So it is a valid IV (since we have assumed $\pi \neq 0$ and $\mathbb{E}[z^3] \neq 0$).

3. [10 points] Consider the IV estimator $\hat{\beta}_1^{IV}$ you defined in **part 1**. In fact, it can be written as

$$\hat{\beta}_1^{IV} = \beta_1 + \sum_{i=1}^{n} w_i u_i, \quad w_i = \frac{z_i - \bar{z}}{\sum_{i=1}^{n}(z_i - \bar{z})x_i}, \quad \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i.$$

   In this part, we **do not** assume $\mathbb{E}[u|z] = 0$. Instead, suppose now that $u = \gamma z + e$ with $\mathbb{C}ov[z, e] = 0$ and $\mathbb{E}[e] = 0$. Use $\text{plim}(\hat{\beta}_1^{IV})$ to denote the probability limit of the IV estimator

$\hat{\beta}_1^{IV}$.

(i) Show that in this case $\text{plim}(\hat{\beta}_1^{IV}) = \beta_1 + \frac{\mathbb{V}[z]\gamma}{\mathbb{C}ov[z,x]}$. [Hint: Use the expression of $\hat{\beta}_1^{IV}$ given above and apply Law of Large Numbers.]

- **Solution**: $\text{plim}(\hat{\beta}_1^{IV}) - \beta_1 = \frac{\mathbb{C}ov[u,z]}{\mathbb{C}ov[x,z]} = \frac{\gamma\mathbb{C}ov[z,z]+\mathbb{C}ov[e,z]}{\mathbb{C}ov[x,z]} = \frac{\mathbb{V}[z]\gamma}{\mathbb{C}ov[z,x]}$.

(ii) Complete the missing entries (a), (b), (c) and (d) in the following table, which describes the consistency/inconsistency pattern of the estimator $\hat{\beta}_1^{IV}$ in different scenarios (two entries have been completed as an example).

	$\mathbb{C}ov[z,x] > 0$	$\mathbb{C}ov[z,x] < 0$
$\gamma > 0$	$\text{plim}(\hat{\beta}_1^{IV}) > \beta_1$	(a)
$\gamma = 0$	(b)	$\text{plim}(\hat{\beta}_1^{IV}) = \beta_1$
$\gamma < 0$	(c)	(d)

- **Solution**:

	$\mathbb{C}ov[z,x] > 0$	$\mathbb{C}ov[z,x] < 0$
$\gamma > 0$	$\text{plim}(\hat{\beta}_1^{IV}) > \beta_1$	$\text{plim}(\hat{\beta}_1^{IV}) < \beta_1$
$\gamma = 0$	$\text{plim}(\hat{\beta}_1^{IV}) = \beta_1$	$\text{plim}(\hat{\beta}_1^{IV}) = \beta_1$
$\gamma < 0$	$\text{plim}(\hat{\beta}_1^{IV}) < \beta_1$	$\text{plim}(\hat{\beta}_1^{IV}) > \beta_1$

2 Question 2 (30 points)

Consider the regression of log wage ($lwage$) on years of schooling ($educ$). Suppose we have twins data. Then, the regression model can be written as

$$lwage_{i,1} = \beta_0 + \beta_1 educ_{i,1} + f_i + v_{i,1}$$

$$lwage_{i,2} = \beta_0 + \beta_1 educ_{i,2} + f_i + v_{i,2}$$

where the subscript i denotes the ith family, and the subscripts 1 and 2 denote the two twin children in a family. For example, $lwage_{i,1}$ denotes the log wage of the first twin child in family i. f_i denotes an *unobserved* "family effect" which is common to the two twin children in family i. So the composite error is $u_{i,e} = f_i + v_{i,e}$, $e = 1, 2$. The variables $(v_{i,e}, educ_{i,e})$ are i.i.d. over both $i = 1, 2, \cdots, n$ and $e = 1, 2$, and f_i are i.i.d. over $i = 1, \cdots, n$. Assume

$$\mathbb{E}[v_{i,e}|\, educ_{i,1},\, educ_{i,2}] = 0, \quad \mathbb{V}[f_i] > 0, \quad \mathbb{C}ov[f_i, v_{i,e}] = 0.$$

1. Suppose that $\mathbb{E}[f_i|educ_{i,1}, educ_{i,2}] = 0$. Consider the OLS estimator $\hat{\beta}_1$ from regressing $lwage_{i,e}$ on $educ_{i,e}$, for $i = 1, \cdots, n$, $e = 1, 2$.

   (a) [5 points] Do you think $\hat{\beta}_1$ is unbiased? Is it consistent? [Hint: what is $\mathbb{E}[u_{i,e}|educ_{i,1}, educ_{i,2}]$ in this case?]

   - **Solution**: The zero conditional mean assumption is implied by the assumptions imposed. Thus, the OLS estimate is unbiased and consistent (consistency only requires uncorrelatedness between $educ_{i,e}$ and $u_{i,e}$).

   (b) [5 points] Which kind of standard errors do you prefer to report for $\hat{\beta}_1$? Explain why.

   - **Solution**: $\mathbb{V}[u_{i,e}] = \mathbb{V}[f_i] + \mathbb{V}[v_{i,e}]$, and $\mathbb{C}ov[u_{i,1}, u_{i,2}] = \mathbb{V}[f_i] > 0$. So it is better to report standard errors that are robust to intra-cluster correlation where each family is viewed as a cluster.

   (c) [10 points] Suppose only in this part that we *cannot* directly observe $educ_{i,e}$. Instead we only have the years of schooling reported by the respondents, $\widetilde{educ}_{i,e}$, which satisfies

   $$\widetilde{educ}_{i,e} = educ_{i,e} + \varepsilon_{i,e}, \quad \mathbb{E}[\varepsilon_{i,e}] = 0,$$

   where the measurement error $\varepsilon_{i,e}$ satisfies the classical error-in-variable assumptions: $\mathbb{C}ov[\varepsilon_{i,e}, u_{i,e}] = 0$, $\mathbb{C}ov[\varepsilon_{i,e}, educ_{i,e}] = 0$. Let $\widetilde{\beta}_1$ be the OLS estimator from regressing $lwage_{i,e}$ on $\widetilde{educ}_{i,e}$.

   (i) Do you think $\widetilde{\beta}_1$ is downward biased or upward biased? (You do not have to give a proof of your answer.)

   (ii) If you can design this survey, could you come up with a question the answer to which can be used as an instrumental variable for $\widetilde{educ}_{i,e}$? Explain why you think it could be a valid IV.

   - **Solution**: The measurement error leads to attenuation bias, so the estimated return to education may be smaller than the truth (we believe the return to education is positive).

     The second question is open-ended. For example, we can ask each twin child to report the years of schooling of himself/herself *and* his twin brother/sister's educa-

tion. So we have a repeated measurement that can be used as an IV for self-reported education $\widetilde{educ}_{i,e}$. Since they are measurements of the same person's education, they are likely to be correlated (*relevance condition*), and hopefully, they are two independent measurements so that the measurement errors are uncorrelated (*exogeneity condition*).

2. Now suppose that we understand f_i as the unobserved ability of the twin children in each family i (assuming twins have the same ability).

   (a) [5 points] Do you think the assumption $\mathbb{E}[f_i|educ_{i,1}, educ_{i,2}] = 0$ in part 1 is plausible? Do you think the OLS estimate $\hat{\beta}_1$ in part 1 is still unbiased? Explain your answer.

   - **Solution**: The zero conditional mean assumption is no longer plausible since ability is very likely to be correlated with the education level. For example, people with higher ability may be better at exams and thus achieve higher level of education. Then, the OLS estimate $\hat{\beta}_1$ would be biased.

   (b) [5 points] Now, consider the transformed regression model

   $$\Delta lwage_i = \beta_1 \Delta educ_i + \Delta v_i, \quad i = 1, \cdots, n$$

   where $\Delta lwage_i = lwage_{i,1} - lwage_{i,2}$, $\Delta educ_i = educ_{i,1} - educ_{i,2}$ and $\Delta v_i = v_{i,1} - v_{i,2}$. Let $\check{\beta}_1$ be the OLS estimate of β_1 from regressing $\Delta lwage_i$ on $\Delta educ_i$, $i = 1, \cdots, n$. Do you think $\check{\beta}_1$ is unbiased? (You can assume there is some variation in $\Delta educ_i$ so this estimator exists.)

   - **Solution**: By assumption that $\mathbb{E}[v_{i,e}|educ_{i,1}, educ_{i,2}] = 0$, $\mathbb{E}[v_{i,1} - v_{i,2}|educ_{i,1}, educ_{i,2}] = 0$, and then the estimator $\check{\beta}_1$ is unbiased.

3 Question 3 (50 points)

Consider a dataset on women's labor supply and other features (from Angrist and Evans, 1998):

- $hours$ = weekly hours worked
- $kids$ = number of children

- $educ$ = years of schooling
- $nonmomi$ = family income - mom's income
- age = age of mom
- $black$: a race dummy, equals 1 if the mom is black, equals 0 if the mom is not black

All women in this dataset are married and have at least two children.

1. Consider the regression model

$$hours = \beta_0 + \beta_1 kids + \beta_2 educ + \beta_3 nonmomi + \beta_4 age + \beta_5 age^2 + \beta_6 black + u$$

(a) [5 points] The estimated equation given by OLS regression is

$$\widehat{hours} = -15.6 - 2.32\, kids + 0.59\, educ - 0.06\, nonmomi + 2.05\, age - 0.03\, age^2 + 6.09\, black$$

How do you interpret the coefficients of $kids$ and $black$?

- **Solution**: Holding other factors fixed, on average having one more kid reduces the mom's weekly labor supply by 2.32 hours. Holding other factors fixed, on average black women work 6.09 more hours weekly than non-black women.

(b) [5 points] When we add the interaction term between $black$ and $kids$, the estimated equation becomes

$$\widehat{hours} = -16.01 - 1.80\, kids + 9.80\, black - 1.38\, black \times kids + 0.61\, educ - 0.06\, nonmomi + 1.97\, age - 0.03\, age^2$$

What are the effects of having one more kid on the weekly working hours for black and non-black women respectively?

- **Solution**: For non-black women, having one more kid reduces labor supply by 1.8 hours while for black women, having one more kid reduces labor supply by 3.18 hours.

(c) [15 points] Angrist and Evan (1998) propose an IV, $samesex$, for $kids$. $samesex = 1$ if the first two kids are of the same sex, and $samesex = 0$ if the first two kids are of

different sex. Answer the following three questions concisely:

(i) What do you think is the argument for why *samesex* is a relevant instrument for *kids*?

(ii) Can you think of a mechanism by which *samesex* is correlated with u in this regression (so the exogeneity condition is violated)?

(iii) Suppose we believe *samesex* is a valid IV. Do you think the effect of *kids* given by the IV estimation is applicable to all women? Explain why or why not. [Hint: think about the "local average treatment effect" interpretation discussed in class.]

- **Solution**: Open-ended questions. Any reasonable answers are acceptable.

  In many cultures, parents have a preference for mixed sex composition of kids. If the first two kids are of the same sex, they might want to have more kids in order to get a kid of different sex.

  It is likely that if the two kids are of the same sex, they can share the same room, toy, clothes, etc., and thus relax the financial budget constraint of the family. Probably, this would make the woman in such a family less motivated to work.

  Without further analysis, we want to be cautious about applying the IV results to the entire population. The result of IV is driven by those women who would really have more kids if the first two kids are of the same sex (their behavior is affected by IV). The effect on parents who would never have more than two kids (or those who will always have more than 2 kids) may be very different. Also, the analysis uses data on women with at least two kids, and may not be representative of women with less than 2 kids.

(d) [10 points] Describe how to test for relevance of the instrument *samesex* for *kids*.

- **Solution**: Run a regression of *kids* on *samesex* and all other exogeneous variables $educ, nomomi, age, age^2, black$:

$$kids = \gamma_0 + \gamma_1 samesex + \gamma_2 educ + \gamma_3 nonmomi + \gamma_4 age + \gamma_5 age^2 + \gamma_6 black + e$$

  The null hypothesis of interest is $H_0 : \gamma_1 = 0$. Use a t or F test. A very small p value means strong evidence for rejecting the null hypothesis, i.e., the *kids* are strongly

partially correlated with *samesex*. (Not required: as a rule-of-thumb, $t^2 = F > 10$).

2. Now, define a dummy variable *work*, which equals 1 if $hours > 0$ and equals 0 if $hours = 0$.

   (a) [10 points] By running an **OLS regression**, we obtain the following estimated equation that explains the decision to work:

   $$\widehat{work} = -0.45 - 0.06\, kids + 0.02\, educ - 0.001\, nonmomi + 0.06\, age - 0.001\, age^2 + 0.15\, black$$

   (i) How do you interpret the coefficient of *kids* in this case?

   - **Solution**: Holding other factors fixed, with one more kid, the women's probability of working decreases by 6 percentage points.

   (ii) Which kind of standard errors do you prefer to report? Explain why.

   - **Solution**: For LPM, we have shown that $\mathbb{V}[u|\mathbf{x}] = p(\mathbf{x})(1 - p(\mathbf{x}))$. Unless the response probability $p(\mathbf{x})$ is a constant, the model is heteroskedastic. So it is better to report the heteroskedasticity-robust standard errors for this regression.

   (b) [5 points] Using **probit** estimation, we obtain the following result:

   $$\hat{\mathbb{P}}(work = 1|\mathbf{x}) = \Phi(-2.47 - 0.17\, kids + 0.05\, educ - 0.003\, nonmomi + 0.14\, age - 0.002\, age^2 + 0.40\, black)$$

   where $\hat{\mathbb{P}}(work = 1|\mathbf{x})$ denotes the estimated probability of $work = 1$ given all explanatory variables $\mathbf{x} = (kids, educ, \cdots, black)$. $\Phi(\cdot)$ is the standard normal cumulative distribution function.

   How do you calculate the average partial effect of *kids* in this case? (Assume the sample size is n. You only need to give an expression.)

   - **Solution**:

   $$\widehat{APE}_{kids} = -0.17 \times \frac{1}{n}\sum_{i=1}^{n} \phi\Big(-2.47 - 0.17\, kids_i + 0.05\, educ_i - 0.003\, nonmomi_i + 0.14\, age_i - 0.002\, age_i^2 + 0.40\, black_i \Big)$$

   where $\phi(\cdot)$ is the standard normal probability density function.