Methods Minor Exam – 2024

August 19, 2024

Read the following instructions carefully:

- You will have 3 hours to answer all 3 parts of the exam. It is recommended that you spend around an hour on each part of the exam.
- This exam is closed book, but you may bring a reading list (as per the American/Comparative/International Relations/Political Theory exams).
- If you are uncertain about a question or believe that further information is required to answer a question, start by stating the assumptions you will make to answer the question.
- Grading: Please show every step of your derivations. We grade steps of derivations as well as your final answers. So, even if you cannot solve the problem entirely, we can give you partial points for your derivations. Even if your final answer is correct, you might not get full points if your derivations are incomplete.

1 Broad knowledge short questions

Provide your reasoning when answering the questions below. Formal derivations may be helpful but not strictly necessary for this section. Your answers need not be lengthy (a paragraph will suffice), but be sure to answer the sub-parts of each question.

Question 1.1

Imagine a bivariate regression model in which a dependent variable is regressed on an independent variable. For simplicity, assume that the independent variable is a randomly assigned continuous variable (such as a dosage) and that the parameter of interest is its slope.

(a) What are the consequences of adding random measurement error to the dependent variable for the unbiasedness of the OLS estimator?

(b) What about adding random measurement error to the independent variable? Is OLS unbiased in this instance?

(c) What about the following scenario: for a random subset of the observations, the dependent variable is replaced with a normal random number with mean zero and standard deviation of one. Is OLS unbiased in this instance?

Question 1.2

Suppose you have a dependent variable Y_i, an independent variable X_i, and a disturbance term u_i, as in the following regression equation:

$$Y_i = a + bX_i + u_i.$$

Assume that X_i and u_i are related, rendering the OLS estimator biased. Now suppose that X_i could be modeled as follows:

$$X_i = c + dZ_i + e_i.$$

A common approach is to estimate b using instrumental variables regression.

(a) Is the IV estimator equivalent to the slope from a regression of Y_i on Z_i divided by the slope from a regression of X_i on Z_i?

(b) Assuming that Z_i is independent of u_i, explain why the IV estimator is consistent but not unbiased.

(c) Sometimes researchers test the exclusion restriction required for consistent IV estimates by regressing Y_i on both X_i and Z_i. Explain why this diagnostic approach can produce misleading results.

Question 1.3

Suppose you conduct an experiment in which the outcome is a binary variable. The binary treatment is assigned via block random assignment: 100 of the 200 women are assigned to treatment, and 100 of the 1,000 men are assigned to treatment. Among women in the treatment group, 55 have $Y_i = 1$ outcomes and 45 have $Y_i = 0$ outcomes. Among women in the control group 40 have $Y_i = 1$ outcomes and 60 have $Y_i = 0$ outcomes. Among men in the treatment group 50 have $Y_i = 1$ outcomes and 50 have $Y_i = 0$ outcomes. Among men in the control group, 300 have $Y_i = 1$ outcomes and 700 have $Y_i = 0$ outcomes.

(a) The difference-in-means estimator in this example is $52.5 - 30.91 = 21.59$. Is the difference-in-means an unbiased estimator of the ATE? Why or why not?

(b) Another estimator is a regression of Y_i on treatment and a "fixed effect" for block (i.e., an indicator variable scored 1 if the subject is a man and 0 if the subject is a woman). The slope estimate using this estimator is 18.23 percentage points. Does this regression estimator generate unbiased estimates of the ATE? Why or why not?

(c) Yet another estimator regresses Y_i on treatment within each block. The two block-level estimates of the slope are then pooled by weighting each block by its N of cases. In this case, the estimate is

$$\frac{200}{1300}(15) + \frac{1100}{1300}(20) = 19.23$$

Does this estimator generate unbiased estimates of the ATE? Why or why not? Why is this estimator called the "inverse probability weighted" (IPW) regression estimator?

Question 1.4

Consider the staggered adoption design (also known as a stepped-wedge design) where different units can receive the treatment in different time periods, but once units receive the treatment, they remain exposed to the treatment. Suppose we assume that the parallel trends assumption holds for all groups for all time periods.

(a) Is the two-way fixed effects estimator (with a full set of fixed effects for both unit and period) unbiased for the ATT? Why or why not?

(b) If we additionally assume that treatment effects are constant across units, is the two-way fixed effects estimator unbiased for the ATT?

Question 1.5

A common regression modeling approach is to include product (or "interaction") terms as regressors. In this case, the dependent variable is Y_i and the independent variables are X_i, Z_i, and the product of X_i and Z_i:

$$Y_i = a + bZ_i + cX_i + d(Z_iX_i) + u_i.$$

(a) The inclusion of a product term changes the interpretation of the parameters b and c. What is the proper interpretation of the three slope parameters b, c, and d?

(b) In the context of sharp regression discontinuity designs, X_i may be thought of as the "running variable," i.e., the covariate that predicts outcomes, and Z_i refers to the receipt of the treatment. To fix ideas, suppose that X_i is a party's vote share in a two-party election, and Z_i is whether a candidate from the left-leaning party is elected (i.e., receives at least 50% of the vote). Y_i is the budget allocated to guns versus butter. Assume that X_i has been "centered" such that X_i is zero when the vote share is 50%. Interpret the slope b in this context.

2 Analytical Question

Question 2.1

In this question, we start with the basic difference-in-differences (DID) design where there are two groups (treatment and control groups) and two time periods (one time period before the treatment assignment and one time period after the treatment). We use the following notation.

- We observe units over two time periods $t = 1$ (pre-treatment) and $t = 2$ (post-treatment).
- Define D_{it} to be a binary treatment variable taking the value of one when unit i is in the treatment condition at time t, and taking zero, otherwise.
- Define G_i to be a binary variable that defines whether unit i is in the treatment group. Formally,

$$\begin{cases} D_{i1} = 0 \text{ and } D_{i2} = 1 \text{ if } G_i = 1 \\ D_{i1} = D_{i2} = 0 \text{ if } G_i = 0 \end{cases} \tag{1}$$

- Define $Y_{it}(d)$ to be a potential outcome for unit i when $D_{it} = d$.
- We assume consistency of potential and observed outcomes: $Y_{it} = Y_{it}(D_{it})$.
- We assume that $\{Y_{i1}(1), Y_{i1}(0), Y_{i2}(1), Y_{i2}(0), D_{i1}, D_{i2}, G_i\}_{i=1}^n$ are i.i.d. samples.

We are interested in the average treatment effect on the treated (ATT), defined as:

$$\text{ATT} = \mathbb{E}\{Y_{i2}(1) - Y_{i2}(0) \mid G_i = 1\}.$$

Researchers often rely on the following DID estimator to estimate the ATT.

$$\widehat{\mathsf{DID}}_{\mathsf{basic}} \;=\; \left\{ \frac{1}{n_1} \sum_{i=1}^{n} G_i (Y_{i2} - Y_{i1}) - \frac{1}{n_0} \sum_{i=1}^{n} (1 - G_i)(Y_{i2} - Y_{i1}) \right\},$$

where $n_g = \sum_{i=1}^n \mathbf{1}\{G_i = g\}$ denotes the number of units in each group where $g \in \{0, 1\}$. Throughout Question 2, you can assume n_g/n is constant and is equal to $\Pr(G_i = g)$ for $g \in \{0, 1\}$.

Question 2.1.1

Prove that the estimator $\widehat{\mathsf{DID}}_{\mathsf{basic}}$ is unbiased for the ATT under the following parallel trend assumption:

$$\mathbb{E}\{Y_{i2}(0) - Y_{i1}(0) \mid G_i = 1\} \;=\; \mathbb{E}\{Y_{i2}(0) - Y_{i1}(0) \mid G_i = 0\}. \tag{2}$$

Question 2.1.2

Suppose the treatment variable D_{it} is randomized. Note that we only assume this randomized treatment assignment and we do not impose the parallel trend assumption. In this scenario, is the estimator $\widehat{\mathsf{DID}}_{\mathsf{basic}}$ unbiased for the ATT? If so, show the step-by-step proof. Otherwise, derive the bias of the basic DID estimator.

Question 2.2

Propose a 95% confidence interval for the estimator $\widehat{\mathsf{DID}}_{\mathsf{basic}}$.

Question 2.3

In this question, we examine the DID design with multiple pre- and post-treatment periods where there are two groups (treatment and control groups) and four time periods (two time periods before the treatment assignment and two time periods after the treatment). We use the following notation; note that some notations are the same as above, but we repeat everything for comprehensiveness.

- Index units with $i \in \{1, \ldots, n\}$, where n is the total number of unique units.
- We observe the same n units over four time periods $t = \{0, 1\}$ (pre-treatment) and $t = \{2, 3\}$ (post-treatment).
- Define D_{it} to be a binary treatment variable taking the value of one when unit i is in the treatment condition at time t, and taking zero, otherwise.
- Define G_i to be a binary variable that defines whether unit i is in the treatment group. Formally,

$$\begin{cases} D_{i0} = 0, \ D_{i1} = 0, \ \text{ and } \ D_{i2} = 1, D_{i3} = 1 \ \text{ if } \ G_i = 1 \\ D_{i0} = D_{i1} = 0, \ \text{ and } \ D_{i2} = 0, D_{i3} = 0 \ \text{ if } \ G_i = 0 \end{cases} \tag{3}$$

- Define $Y_{it}(d)$ to be a potential outcome for unit i when $D_{it} = d$.
- We assume consistency of potential and observed outcomes: $Y_{it} = Y_{it}(D_{it})$.
- We assume that $\{Y_{i0}(1), Y_{i0}(0), Y_{i1}(1), Y_{i1}(0), Y_{i2}(1), Y_{i2}(0), Y_{i3}(1), Y_{i3}(0), D_{i0}, D_{i1}, D_{i2}, D_{i3}, G_i\}_{i=1}^n$ are i.i.d. samples.

In this question, we are now interested in the long-term ATT defined as:

$$\text{ATT}_{\text{long}} = \mathbb{E}\{Y_{i3}(1) - Y_{i3}(0) \mid G_i = 1\}.$$

Question 2.3.1

Propose an estimator for the long-term ATT and an assumption under which the proposed estimator is unbiased. Note that there could be many potential estimators, and you only need to provide one estimator here.

Question 2.3.2

Propose another estimator and assumption under which this second estimator is unbiased for the long-term ATT. After proposing the second estimator, compare the advantages and disadvantages of the two estimators you propose in Question 2.3.1 and Question 2.3.2.

3 Research design and critique questions

Answer ONE of the following questions. You may choose an option from outside your subfield. If further information about the study context is required, explain why this information is important and specify what assumptions you will make in your answer.

3.1 American politics option

Suppose that you are investigating the effect of Fox News on turnout and vote share for the Republican party, in order to understand the electoral effects of partisan media outlets. Fox News first became available in 1996, with different cable markets receiving access to the news channel at different points in time. The cable markets that received Fox News earlier were not randomly assigned. In 1996, only 5% of the population was in a cable market with access to Fox News. By 2000, Fox News was available in cable markets covering 50% of the population.

The following data are available to you: the date that Fox News became available in each cable market; presidential electoral returns for every county in the country; and various pre-1996 characteristics of counties.

1. Propose an empirical strategy, including a statement of your estimand, any data restrictions, the identifying assumptions, an estimating equation (including any weights), and an approach to inference.

2. What robustness checks would you conduct to convince a reviewer that your estimates are internally valid? Explain what concern each robustness check addresses.

3. To increase the external validity of this study, what data would you seek to collect?

3.2 Comparative politics option

Suppose that you are investigating the effect of independent audit reports on the vote share of Brazilian mayors, in order to understand the extent to which mayors found to be corrupt are electorally sanctioned and mayors found to be clean are electorally rewarded. Municipal mayors serve 4 year terms, and can be re-elected only once; candidate selection is finalized 3 months before elections. Around 50 municipalities with populations below 450,000 people are randomly selected (with replacement) each month for audits of prior expenditures, with audit reports released 6 months after the audit has been announced. The reports are conducted by independent auditors, although the auditors must in part rely on municipal governments to facilitate thorough audits.

The following data is available to you: the date that every audit was announced and the report was released; the number of corruption violations detected in each audit; the date of the 2004 and 2008 municipal elections and history of the electoral returns for every municipality in the country; and various predetermined municipal and mayoral characteristics.

1. Propose an empirical strategy, including a statement of your estimand, any data restrictions, the identifying assumptions, an estimating equation (including any weights), and an approach to inference.

2. What robustness checks would you conduct to convince a reviewer that your estimates are internally valid? Explain what concern each robustness check addresses.

3. To increase the external validity of this study, what data would you seek to collect?

3.3 International relations option

Suppose you are investigating the effect of democracy on the probability of (interstate or civil) war, in order to understand the effect of regime type on the onset of conflict. Since regime type is not randomly assigned, a possible identification strategy leverages assassination attempts directed at national leaders that sometimes succeed and sometimes fail. Successful assassination attempts of autocrats may be especially likely to lead to institutional changes, such as shifts toward democracy. Unsuccessful assassination attempts may instead entrench autocrats.

The following data are available to you: the date and outcome (59 leaders died) of 298 assassination attempts of national leaders between 1875 and 2004; an annual country-year panel of Polity autocracy/democracies designations and scores from 1816 to 2004; an annual country-year panel indicating whether a country is at war from 1816 to 2004; and various characteristics of countries at different moments in time.

1. Propose an empirical strategy, including a statement of your estimand, any data restrictions, the identifying assumptions, an estimating equation (including any weights), and an approach to inference.

2. What robustness checks would you conduct to convince a reviewer that your estimates are internally valid? Explain what concern each robustness check addresses.

3. To increase the external validity of this study, what data would you seek to collect?