Methods Minor Exam – 2024

December 2, 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).
- You are not allowed to use any form of artificial intelligence during this exam.
- 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 are 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 – Covariate Adjustment

Imagine an experiment with N subjects, exactly half of which are (completely) randomly assigned to treatment. Experimental researchers routinely use regression to estimate the average treatment effect by regressing an outcome Y on the treatment indicator T. (Subscripts designating units are omitted throughout this question.)

(a) Is this estimator unbiased? Does it remain unbiased if N is a small odd number, like 7, and the number of subjects assigned to treatment (n_1) is different from the number of subjects assigned to control (n_0) but both n_1 and n_0 are greater than zero?

(b) Sometimes researchers will add "covariates" – variables that are measured prior to treatment – to this regression model as regressors. Consider the simple case in which a single covariate X is included in the regression equation: $Y = a + bT + cX + u$. Does this regression give an unbiased estimate of the average treatment effect? Why or why not?

(c) Does the inclusion of X in the regression model always improve the precision with which the average treatment effect is estimated? If so, explain why. If not, explain why not.

(d) What should we infer if the estimate of the ATE changes quite a bit depending on whether we include X as a regressor?

(e) Explain how you would graph the estimated causal effect of T on Y that is revealed by a regression of $Y = a + bT + cX + u$ to account for the fact that regression adjustment eliminates the covariance between Y and X as well as the covariance between T and X.

Question 1.2 – Experiments with Two-sided Noncompliance

In a classic paper, Angrist (1990) used the Vietnam Draft Lottery in order to understand the effects of military service on earnings many years later among draft-eligible men. Let Z represent the lottery outcome: $Z = 1$ if an eligible man was selected for military service ("drafted") and $Z = 0$ if not selected. Let D represent whether an eligible man actually served in the military; $D = 1$ when eligible men served

in the military and $D = 0$ when they did not serve. Let Y be the annual earnings that are recorded by the government twenty years after the draft lottery. (Subscripts for units are omitted throughout this question.)

(a) What assumptions are required to identify the average causal effect of military service among "compliers" (i.e., those who would serve in the military if and only if they are drafted)?

(b) For each assumption, briefly explain why this local average treatment effect could not be identified if that assumption is violated.

(c) What estimator would you use to estimate the local average treatment effect, and why? Is that estimator unbiased? Is it consistent?

Question 1.3 – Modeling Interactions

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.

(c) Suppose that in part (b) X_i were not "centered"; instead, X_i is zero when the vote share is 0% and 1 when the vote share is 100%. Interpret the slope b in this context. Explain how you would calculate the estimated effect of the discontinuity that occurs when the party wins 50% of the vote if you were given estimates of a, b, c, and d?

Question 1.4 – Standard Errors

(a) Define the term "standard error."

(b) What is the difference between a standard error and a standard deviation?

(c) What is a clustered standard error? When is it appropriate to use a clustered standard error rather than a conventional standard error?

(d) Consider the following statement: "When analyzing data, it is best to use the estimator that has the smallest standard error." Would you agree or disagree? Why or why not?

2 Analytical Question

Setup. In this question, we examine a randomized experiment. When you provide answers, use the following notation. If you need to introduce additional notation, explain it in detail. We assume no interference across units throughout this question.

- Index experimental units with $i \in \{1, \ldots, N\}$, where N is the total number of experimental units.
- Define $T_i \in \{1, 0\}$ to be a binary treatment variable.
- Define $Y_i(t)$ to be the potential outcome when unit i receives $T_i = t$, where $t \in \{0, 1\}$.
- Define Y_i to be the observed outcome for unit i, and we assume consistency of potential outcomes, $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$.

Suppose a researcher employs a block randomized experiment. That is, treatment assignment is completely randomized within blocks. More specifically, units are partitioned into two equally sized blocks, defined by a binary pre-treatment variable $X_i \in \{0, 1\}$. There are $N/2$ units with $X_i = 1$ and $N/2$ units with $X_i = 0$.

Within block 1 (units with $X_i = 1$), 75% of the units receive the treatment and the other 25% receive the control based on complete random assignment. Use N_{11} (N_{01}) to denote the number of treated (control) units within block 1, where $N_{11} + N_{01} = N/2$. On the other hand, within block 0 (units with $X_i = 0$), 25% of the units receive the treatment and the other 75% receive the control based on complete random assignment. Use N_{10} (N_{00}) to denote the number of treated (control) units within block 0, where $N_{10} + N_{00} = N/2$.

The design can be summarized as the following table, where each cell shows the number of units for each combination of X_i and T_i.

	$T_i = 1$	$T_i = 0$
$X_i = 1$	$N_{11} = 3N/8$	$N_{01} = N/8$
$X_i = 0$	$N_{10} = N/8$	$N_{00} = 3N/8$

Question 2.1

We formally define the conditional average treatment effects (CATE) as follows:

$$\tau_1 \quad = \quad \frac{1}{N/2} \sum_{i=1}^{N} X_i \{Y_i(1) - Y_i(0)\}, \tag{1}$$

$$\tau_0 \;=\; \frac{1}{N/2}\sum_{i=1}^{N}(1-X_i)\{Y_i(1)-Y_i(0)\}. \tag{2}$$

What does the difference between the CATEs, $\tau_1-\tau_0$, mean? Please provide a methodological explanation and state whether or not the difference in CATEs can be interpreted as a causal estimand.

Question 2.2

Propose unbiased estimators for τ_1 and τ_0. Provide a step-by-step proof of each estimator's unbiasedness.

Question 2.3

Now, we are interested in the average treatment effect (ATE), which is formally defined as follows:

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\{Y_i(1)-Y_i(0)\}. \tag{3}$$

The difference-in-means estimator, defined in equation (4), is not in general unbiased for the ATE (equation (3)) under this experimental design.

$$\widehat{\tau}^{DM} \;=\; \frac{1}{N_{11}+N_{10}}\sum_{i=1}^{N}T_iY_i - \frac{1}{N_{01}+N_{00}}\sum_{i=1}^{N}(1-T_i)Y_i. \tag{4}$$

Question 2.3.1

Provide an intuitive methodological explanation about why the difference-in-means estimator is biased by using the following two phrases: {treatment assignment probability, confounder}.

Question 2.3.2

Derive an exact expression of the bias, defined as Bias $= \mathbb{E}[\widehat{\tau}^{DM} \mid \mathcal{O}_N]-\tau$, where $\mathcal{O}_N = \{Y_i(1), Y_i(0), X_i\}_{i=1}^{N}$. To simplify the problem, we assume that $Y_i(0) = 0$ for everyone in the experiment.

Hint: You can use the following equalities in your proof.

$$\begin{aligned}
\mathbb{E}[T_iX_i \mid \mathcal{O}_N] &= \frac{3X_i}{4},\\
\mathbb{E}[(1-T_i)X_i \mid \mathcal{O}_N] &= \frac{X_i}{4},\\
\mathbb{E}[T_i(1-X_i) \mid \mathcal{O}_N] &= \frac{(1-X_i)}{4},\\
\mathbb{E}[(1-T_i)(1-X_i) \mid \mathcal{O}_N] &= \frac{3(1-X_i)}{4}.
\end{aligned}$$

Question 2.3.3

Propose an alternative estimator for the ATE that you believe is unbiased. Explain in words why you believe this estimator is unbiased.

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 newspaper coverage of U.S. House representatives on citizen knowledge of their Congresspeople, the actions of these Congresspeople in office, and voter support for incumbent representatives, in order to understand the role of the media in supporting political accountability. Newspapers can report on local politicians, but the politicians they report on are likely to depend on the location of their readership. A possible identification strategy is then to exploit changes in the incentives for newspapers to report on Congresspeople due to changes in readership composition over time.

The following data are available to you at the Congressional county level for each congressional session between 1991-1992 and 2001-2002: the share of the average local newspaper's readership that lives in each county (weighting by newspaper market share in that county), which captures the congruence between newspaper markets and Congressional districts; repeated cross-sectional survey data containing questions asking whether respondents in a given county read about the incumbent Congressperson and can accurately recall their name; electoral returns for each county over the time period; various measures of legislator activity in the U.S. House, including number of witness appearances during congressional hearings, committees served on, votes in line with party leadership, and DW-NOMINATE scores for each legislator-session; government spending per capita directed toward each district; and various county-level and district-level covariates.

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 democracy on primary school enrollment rates, in order to understand whether democracies are more likely than autocracies to provide pro-poor public goods and services. Between 1820 and 2010, many countries democratized, but did so at different times; some countries subsequently ceased to be democratic and others oscillated between regime types multiple times. Whether a country is democratic at a given moment in time is not randomly assigned. There is considerable variation in primary school enrollment rates across countries and time, although countries that democratized later generally already had higher primary school enrollment rates.

The following data is available to you at the country-year level for 109 large countries from across the world between 1820 and 2010: an annual binary measure of democracy, defined by a polity2 score between 6 and 10; the annual share of the population aged 5-14 in school; and various predetermined country characteristics. However, the pre-1945 school enrollment data relies on imputation for many countries, due to missing data, and the first year of data collection varies across countries.

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 GDP growth on civil war in Sub-Saharan Africa, in order to understand the effect of economic fluctuations on intrastate conflict. Since GDP growth is not randomly assigned, a possible identification strategy leverages variation in levels of rainfall within countries over time because the agriculture sector in many Sub-Saharan African economies requires sufficient rainfall each year to generate high crop yields. Rainfall varies significantly across years within countries, and both the current and prior year might influence GDP growth.

The following data is available to you at the country-year level between 1981 and 1999: the annual GDP growth rate; the incidence of civil war, defined by at least 25 deaths per year in armed conflicts between a government and another domestic actor (civil war are frequent, occurring in 27% of country-years); annual rainfall in millimeters; and various predetermined country 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?