

# Methods Minor Exam – August, 2025

**Read the following instructions carefully:**

- You will have 4 hours to answer all 3 parts of the exam. It is recommended that you spend approximately equal time 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 – Causal Estimands

Imagine a study – not necessarily an experiment – with  $N$  subjects, exactly half of which are treated. Assume that SUTVA holds.

- (a) Using potential outcomes notation, define the average treatment effect (ATE) and the average treatment effect on the treated (ATT).
- (b) Imagine a dataset with extensive background information  $\mathbf{X}$  about every subject. How does an “exact matching” estimator of the ATT work? (Describe the procedure used to generate estimates.) Under what assumptions could a researcher use exact matching to obtain unbiased estimates of the ATT? (You don’t need to show a proof; just state the key assumptions).
- (c) If this were a randomized trial, the difference-in-means estimator would generate an unbiased estimate of the ATE. Would it also generate an unbiased estimate of the ATT? Why or why not?
- (d) Exact matching is often praised as an estimation strategy for analysis of observational data, but it is rarely used to analyze experimental data. Suppose we applied exact matching to an experiment. What are the advantages or disadvantages of exact matching in terms of the bias or precision with which the ATT or ATE is estimated?

## Question 1.2 – Experiments with One-sided Noncompliance

Many door-to-door voter mobilization experiments encounter noncompliance when they attempt to talk to voters at their homes. Suppose  $N$  subjects’ names appear on a voter registration database for a given location. Let  $Z \in \{0, 1\}$  represent the random assignment of each subject:  $Z = 1$  if a subject is on the target list to be contacted by door-to-door canvassers, and  $Z = 0$  if a subject is assigned to the control group. Let  $D \in \{0, 1\}$  represent whether the subject is actually contacted:  $D = 1$  when the subject answers the door when the canvasser visits, and  $D = 0$  when they do not. (Assume that this study is an instance of one-sided noncompliance, in that the assigned control group is never visited by canvassers: i.e.,  $D = 0$  when  $Z = 0$ .) Let  $Y$  be the official voter turnout record for each voter. (Subscripts for units are omitted throughout this question.)

- (a) Briefly list the assumptions required to identify the average causal effect of speaking with a canvasser among “compliers” (i.e., those who would answer the door if a canvasser visits)?
- (b) For each assumption, briefly explain why this complier average causal effect (CACE) could not be identified if that assumption is violated.
- (c) What estimator would you use to estimate the CACE, and why? Is that estimator unbiased and/or consistent?
- (d) Sometimes in the face of one-sided noncompliance, researchers opt for a placebo-controlled design. In this context, a placebo condition might involve sending canvassers to doors but randomly varying whether they deliver a script encouraging voting or encouraging some unrelated activity, such as recycling. What are the advantages and disadvantages of a placebo-controlled design in terms of bias and precision?

### **Question 1.3 – Statistical Knowledge**

Evaluate each of the statements below. Are they true or false? Why?

- (a) “A 95% confidence interval states that there is a 95% chance that the true parameter lies within the estimated interval.”
- (b) “Suppose that you observe  $n$  i.i.d. realizations of a random variable  $X$  (with finite variance). Even when  $X$  is not normally distributed, the central limit theorem nevertheless proves that the sample mean of  $X$  converges in distribution to a normally-distributed random variable.”
- (c) “For a randomized experiment conducted on subjects that are drawn at random from a large population, the Neyman estimator of the standard error for a difference in means is conservative for the population average treatment effect.”

### **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

Assume that you have a dataset of  $n$  iid observations, denoted by  $i = 1, \dots, n$ . For each unit, you observe the following variables: an outcome  $Y_i \in \mathbb{R}$ ; a binary treatment  $D_i \in \{0, 1\}$ ; and a covariate  $R_i \in [0, 1]$ . The treatment is known to switch on for  $R_i \geq c$ , where  $c \in (0, 1)$  is a scalar cutoff. Assume throughout this question that SUTVA holds, so observed outcomes are related to potential outcomes  $Y_i(D_i)$  by  $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$ .

### Question 2.1

Suppose that you decide to implement a regression discontinuity design to estimate a causal effect of  $D_i$  on  $Y_i$ . Formally define a target estimand in terms of potential outcomes.

### Question 2.2

Formally state a continuity assumption, and then use it to prove that your estimand from question 2.1 can be non-parametrically identified by observed conditional expectations.

### Question 2.3

We now turn to estimation of these conditional expectations. Assume the true data generating process for potential outcomes is:

$$\begin{aligned}\mathbb{E}[Y_i(0)] &= \phi(R_i - c), \\ \mathbb{E}[Y_i(1)] &= \tau + \psi(R_i - c),\end{aligned}$$

where Greek letters denote constant parameters and  $\phi \neq \psi$ .

Suppose that a researcher estimates the following “globally linear” OLS regression using the entire dataset:

$$Y_i = \alpha + \beta D_i + \gamma(R_i - c) + \varepsilon_i.$$

Derive an expression for the bias of  $\hat{\beta}$  as an estimator of  $\tau$ . If  $\hat{\beta}$  is biased, what is the sign of the bias?

### Question 2.4

Suppose that a researcher instead estimates the following “piecewise linear” OLS regression using the entire dataset:

$$Y_i = \alpha + \beta D_i + \gamma(R_i - c) + \delta D_i(R_i - c) + \varepsilon_i.$$

Derive an expression for the bias of  $\hat{\beta}$  as an estimator of  $\tau$ . If  $\hat{\beta}$  is biased, what is the sign of the bias?

### Question 2.5

Derive the homoskedastic standard error of  $\hat{\beta}$  for the “piecewise linear” regression, assuming  $Var[\varepsilon_i|D_i, R_i] = \sigma^2$ . How does this standard error vary with  $n$ ?

### Question 2.6

Now let the true data generating process for potential outcomes be given by:

$$\begin{aligned}\mathbb{E}[Y_i(0)] &= \phi(R_i - c), \\ \mathbb{E}[Y_i(1)] &= \tau + \psi(R_i - c)^2.\end{aligned}$$

Explain *in words* whether the “piecewise linear” OLS regression is now biased or unbiased.

### Question 2.7

Some scholars recommend restricting estimation of the “piecewise linear” regression to observations within an interval  $[c - h, c + h]$ , where  $h > 0$  is a bandwidth. Based on your answers to questions 2.5 and 2.6, explain *in words* whether or not there is a trade-off between bias and variance due to reducing the size of the bandwidth (i.e. selecting a small  $h$ )?

### 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 Question 3.1 (American Politics option)

Suppose that you are investigating the effect of short spells in jail for first-time misdemeanors on an individual's propensity to vote, in order to understand the effect of incarceration on subsequent electoral participation. In Harris County, which contains the city of Houston, first-time misdemeanors (e.g. driving while intoxicated, marijuana possession) can be punished by fines/probation or up to a year in jail. Defendants who differ in the severity of their misdemeanor may differ in their propensity to vote. However, defendants are randomly assigned to one of 15 courtrooms, each comprising a particular judge and team of prosecutors. Judges are elected to four-year terms and vary in their sentences for misdemeanors, with the most generous judge sentencing 47% of defendants to jail time and the harshest judge sentencing 63% of defendants to jail time.

The following data is available to you at the individual level for all defendants whose cases were filed with the Harris Country Criminal Courts between November 2008 and November 2012: information about charges faced, presiding judge's courtroom, and sentence; linked Texas voter file records (wherever possible), including registration and turnout; and various demographic covariates.

1. Propose and justify 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 Question 3.2 (Comparative Politics option)

Suppose that you are investigating the effect of the number of years of exposure to progressive Catholic bishops on electoral support for the PT (Workers' Party) in Brazil between 1989 and 2002, in order to understand the electoral effects of religious leaders in the presidential elections after a transition back to democracy (in 1985). Importantly, the pope has complete discretion to appoint Catholic bishops, but

can only replace bishops when they reach the retirement age of 75 (or otherwise leave office). Paul VI was a progressive Catholic pope who died in 1978; he was replaced by a conservative pope, John Paul II, who replaced incumbent bishops with more conservative bishops when he could. There is considerable variation across Brazilian dioceses in the age of incumbent bishops in 1978, and thus across dioceses in the number of years until John Paul II could replace bishops appointed by his progressive predecessor. Since progressive bishops may not be randomly assigned across diocese, this generates plausibly exogenous variation in the total number of years that a diocese was led by a progressive bishop.

The following data is available to you at the municipal level (where Brazil's many municipalities are located strictly within 189 dioceses (or excluded archdioceses), which are in turn located within 26 states or federal district of Brasilia) for the presidential elections held in 1989, 1994, 1998, and 2002: presidential vote shares for all parties; the year of death or of the 75th birthday of the bishop in each diocese when John Paul II became pope in 1978; and various predetermined municipal characteristics.

1. Propose and justify 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 Question 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 and justify 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?