[This question paper contains 12 printed pages.]

Your Roll No...............

Sr. No. of Question Paper : 6224 **J**

Unique Paper Code : 2273100027

Name of the Paper : Introduction to Causal Inference

Name of the Course : **Common Pool of DSE**

Semester : VI

Duration : 3 Hours Maximum Marks : 90

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. There are **two Sections, A & B** with a total of **6** questions.
3. Question **1** from **Section A** is compulsory.
4. Answer **any 4** out of the **5** questions from **Section B**.
5. Marks are given at the end of each question.
6. Use of simple calculators is allowed.
7. Answers may be written either in English or Hindi; but the same medium should be used throughout the paper.

छात्रों के लिए निर्देश

1. इस प्रश्न-पत्र के मिलते ही ऊपर दिए गए निर्धारित स्थान पर अपना अनुक्रमांक लिखिए।
2. इसमें दो खंड ए और बी हैं, जिनमें कुल 6 प्रश्न हैं।
3. खंड ए से प्रश्न 1 अनिवार्य है।
4. खंड बी से 5 प्रश्नों में से किन्हीं 4 के उत्तर दीजिए।
5. प्रत्येक प्रश्न के अंत में अंक दिए गए हैं।
6. साधारण कैलकुलेटर का उपयोग करने की अनुमति है।
7. इस प्रश्न-पत्र का उत्तर अंग्रेजी या हिंदी किसी एक भाषा में दीजिए, लेकिन सभी उत्तरों का माध्यम एक ही होना चाहिए।

Section A

1. (a) Show how the Simple Difference in Outcomes (SDO) can be written as a function of the Average Treatment Effect on the Treated (ATT/ATET) and a selection bias term.

(b) How does unit randomization help in getting an unbiased estimate of the ATT?

(c) With regard to part (b), what if selection into treatment is independent only of Y^0 and not of Y^1.

(d) Explain why it might still be beneficial to include covariates in a regression analysis of a randomized controlled trial even if there is good balance in the covariates across the treatment and control groups.

(e) A nonprofit organization runs a financial literacy program to improve household savings. The intervention is randomized at the household level. However, due to logistical constraints, the content and delivery of the program vary across households :

- Some households receive in-person training sessions,
- Others receive online modules with the same core content,
- A few receive only printed brochures summarizing the material.

At the end of the study, the researcher estimates a single average treatment effect by comparing all treated households to controls.

What aspect of SUTVA is violated here? Briefly explain one strategy the researcher could use to address this issue analytically or in study design.

(6+3+2+4+3)

Section B

2. In 2020, City A introduced a new public transportation subsidy aimed at reducing commuting costs for low-income residents. City B, with a similar demographic and economic profile, did not implement any such subsidy. You have data from both cities for the years 2019 (pre-treatment) and 2021 (post-treatment). The outcome variable is the average monthly commuting cost (in USD) for low-income individuals. The regression output for a simple DiD regression is given below.

Variable	**Coefficient**	**Std. Error**	**p-value**
Intercept	85.0	1.2	0.000
Post (2021 = 1)	1.5	1.5	0.320
Treated (City A = 1)	-2.0	1.3	0.120
Treated × Post	-8.0	2.1	0.001

(a) Provide the estimating equation for a DiD estimation of this kind. What is the estimated effect of the public transportation subsidy on monthly commuting costs? Interpret the DiD coefficient.

(b) Is the effect statistically significant at the 5% level? Explain your reasoning.

(c) What is the predicted average monthly commuting cost in 2021 for individuals in City A?

(d) Describe the key identifying assumption behind the DiD approach in this context.

(e) Suggest a method or data feature you could use to test the plausibility of the parallel trends assumption in this setting.

(f) Suppose a major employer relocated out of City A between 2019 and 2021, disproportionately affecting commuting patterns. How could this confound the DiD estimate? (4+2+2+3+3+4)

3. The following table is based on the Dehejia & Wahba (2002) study evaluating the effect of job training on earnings via Propensity Score Matching (PSM) using observational data.

Estimation Method	ATT Estimate ($)	Standard Error	t-statistic
PSM	1794	875	2.05
Radius (0.01)	1850	880	2.10
Mahalanobis	1680	900	1.87
OLS	2250	950	2.37

(a) What do the estimates say about the program's impact?

(b) Explain the concept of common support in the context of matching.

(c) What is a Linear Probability Model (LPM)? Why is logit/probit preferred over a linear probability model for creating the propensity score?

(d) Why use matching over OLS in this context? What does consistency in the results across matching methods suggest?

(e) When will matching fail to give unbiased estimates of treatment effects?

(3+4+5+3+3)

4. Researchers use proximity to a 4-year college (a binary variable) as an instrument for college attendance to estimate its effect on earnings. The results below show the impact of being close to a college on years of schooling/education and income. The study has an adequately powered sample.

Dependent Variable	**Independent Variables**	**Coefficient**	**Std. Error**	**t-stat**
First Stage (Years of Education)	Proximity to college	0.9	0.3	3.0
	Constant	12.5	0.8	
Reduced Form (Log Income)	Proximity to college	0.18	0.07	2.57
	Constant	9.8	0.2	
IV Estimate (2SLS)	Years of Education	0.20	0.09	2.22

(a) Interpret the coefficients from the first-stage and reduced-form regressions.

(b) Use the Wald estimator to manually compute the IV estimate.

(c) In the context of a valid IV estimation what can explain a larger IV estimate compared to OLS?

(d) Differentiate between an instrumental variable and a proxy variable. Can a proxy variable solve the endogeneity issue

(e) Explain briefly the assumptions under which the treatment assignment (Z) can serve as an instrument to estimate the local average treatment effect (LATE) of a treatment (D) on an outcome (Y). (4+2+3+4+5)

5. A state government in India introduced a home energy efficiency subsidy program in 2023. Households with annual income below ₹5,00,000 were eligible to receive a subsidy of up to ₹20,000 for installing energy-efficient appliances. Households earning ₹5,00,000 or more were not eligible.

You are interested in estimating the causal effect of subsidy eligibility on annual household electricity consumption, measured in kilowatt-hours (kWh). To do this, you use data from households earning between ₹4,50,000 and ₹5,50,000, and run a local linear regression:

$$Electricity_i = \alpha + \tau . D_i + f(Income - 5{,}00{,}000) + \varepsilon_i$$

Where :

– $D_t = 1$ if the household earns less than ₹5,00,000 (eligible), and 0 otherwise.

– f() is a linear function of income centered at Rs. 5,00,000.

Variable	Coefficient	Std. Error	p-value
Intercept	9500	100	0.000
Eligibility (D = 1)	-600	200	0.003
Income - ₹5,00,000	-0.03	0.01	0.010

(a) What is the estimated causal effect of subsidy eligibility on electricity consumption?

(b) Is this effect statistically significant at the 5% level? Provide your reasoning.

(c) What is the predicted electricity consumption at the income threshold for a household that just qualifies for the subsidy?

(d) What is the key identifying assumption for the RDD design to yield a valid causal estimate in this context?

(e) How might income misreporting around the ₹5,00,000 threshold affect the validity of the RDD? What type of bias could it introduce?

(f) Describe a diagnostic test or visual analysis that could help assess whether the RDD assumptions are plausible. (3+2+3+4+3+3)

6. In 2020, a city government introduced a health insurance subsidy for informal sector workers. Workers who enrolled in the program received a 50% subsidy on their insurance premiums. Participation was voluntary.
To estimate the program's effect on out-of-pocket health expenditure in 2022, a researcher collected data on workers from 2020 (before the program) and 2022 (after). The researcher first used propensity score matching (PSM) to compare outcomes in 2022 between treated and control workers matched on pre-treatment characteristics. Then, they used a Difference-in-Differences (DiD) model with the full panel data.

Group	**Mean Expenditure**
Treated	6200
Matched Control	7100
ATT	-900

Variable	**Coefficient**	**St. Error**	**p-value**
Intercept	7300	200	0.000
Post (Year 2022=1)	100	100	0.320
Treated	-150	250	0.550
Treated.Post	-1000	200	0.000

(a) Explain the curse of dimensionality and discuss how Propensity Score Matching (PSM) solves the same.

(b) Using an appropriate Directed Acyclic Graph (DAG) briefly explain the covariate balancing property of the propensity score.

(c) Which estimate (PSM/DiD) do you think is more credible in this context and why?

(d) What assumption does DiD rely on that Matching does not? Explain how you would test its plausibility in this context?

(e) Briefly explain how would the DiD estimation differ if the program was rolled out in say 3 phases? What are assumptions that would be needed to enable an unbiased estimation in this case? (4+3+4+4+3)

खण्ड ए

1. (क) प्रतिफल में सरल अंतर (एसडीओ) को ट्रीटेड व्यक्ति पर एवरेज ट्रीटमेंट इफेक्ट (एटीटी/एटीईटी) और चयन पक्षपात पद के एक फलन के रूप में कैसे व्यक्त किया जा सकता है?

(ख) एटीटी का निष्पक्ष अनुमान प्राप्त करने में इकाई यादृच्छिकरण कैसे मदद करता है?

(ग) भाग (ख) के संबंध में, क्या होगा यदि ट्रीटमेंट में चयन केवल Y^0 से स्वतंत्र है, Y^1 से नहीं?

(घ) एक यादृच्छिक नियंत्रित परीक्षण के प्रतीपगमन विश्लेषण में सहचरों को शामिल करना लाभदायक क्यों हो सकता है, भले ही ट्रीटमेंट और नियंत्रण समूहों में सहचरों में अच्छा संतुलन हो।

(ङ) एक गैर-लाभकारी संगठन घरेलू बचत में सुधार के लिए वित्तीय साक्षरता कार्यक्रम चलाता है। हस्तक्षेप घरेलू स्तर पर यादृच्छिक है। हालाँकि, तार्किक बाधाओं के कारण, कार्यक्रम की सामग्री और वितरण घरों में अलग-अलग होते हैं:

- कुछ घरों को व्यक्तिगत प्रशिक्षण सत्र मिलते हैं,
- अन्य को समान मूल सामग्री के साथ ऑनलाइन मॉड्यूल मिलते हैं,
- कुछ को केवल सामग्री का सारांश देने वाले मुद्रित ब्रोशर मिलते हैं।

अध्ययन के अंत में, शोधकर्ता सभी ट्रीटेड परिवारों की तुलना नियंत्रण से करके एकल एवरेज ट्रीटमेंट इफेक्ट का अनुमान लगाता है।

यहाँ SUTVA के किस पहलू का उल्लंघन किया गया है? संक्षेप में एक रणनीति की व्याख्या कीजिए जिसका उपयोग शोधकर्ता विश्लेषणात्मक रूप से या अध्ययन डिजाइन में इस मुद्दे को संबोधित करने के लिए कर सकता है। (6 + 3 + 2 + 4 + 3)

खण्ड बी

2. 2020 में, शहर A ने कम आय वाले निवासियों के लिए आवागमन लागत को कम करने के उद्देश्य से एक नई सार्वजनिक परिवहन सब्सिडी शुरू की। समान जनसांख्यिकीय और आर्थिक प्रोफाइल वाले शहर B ने ऐसी कोई सब्सिडी लागू नहीं की। आपके पास वर्ष 2019 (पूर्व-उपचार) और 2021 (उपचार के बाद) के लिए दोनों शहरों के डेटा हैं। परिणाम चर कम आय वाले व्यक्तियों के लिए औसत मासिक आवागमन लागत (यूएसडी में) है। एक सरल DiD प्रतीपगमन के लिए प्रतीपगमन आउटपुट नीचे दिया गया है।

चर	गुणांक	प्रमाप त्रुटि	p-मान
अंतःखंड	85.0	1.2	0.000
पोस्ट (2021 = 1)	1.5	1.5	0.320
ट्रीटेड(शहर A = 1)	-2.0	1.3	0.120
ट्रीटेड × पोस्ट	-8.0	2.1	0.001

(क) इस तरह के डीआईडी अनुमान के लिए अनुमान समीकरण लिखिए। सार्वजनिक परिवहन सब्सिडी का मासिक आवागमन लागत पर अनुमानित प्रभाव क्या है? डीआईडी गुणांक की व्याख्या कीजिए।

(ख) क्या प्रभाव 5% के स्तर पर सांख्यिकीय रूप से महत्वपूर्ण है? अपने तर्क की व्याख्या कीजिए।

(ग) शहर A में व्यक्तियों के लिए 2021 में अनुमानित औसत मासिक आवागमन लागत क्या है?

(घ) इस संदर्भ में डीआईडी दृष्टिकोण के पीछे प्रमुख पहचान धारणा का वर्णन कीजिए।

(ङ) इस सेटिंग में समानांतर प्रवृत्ति धारणा की व्यवहार्यता का परीक्षण करने के लिए आप जिस विधि या डेटा सुविधा का उपयोग कर सकते हैं, उसका सुझाव दीजिए।

(च) मान लीजिए कि एक प्रमुख नियोक्ता 2019 और 2021 के बीच शहर A से बाहर चला गया, जिससे आवागमन पैटर्न पर असंगत रूप से असर पड़ा। यह डीआईडी अनुमान को कैसे भ्रमित कर सकता है? (4 + 2 + 2 + 3 + 3 + 4)

3. निम्नलिखित तालिका देहेजिया और वहबा (2002) के अध्ययन पर आधारित है, जिसमें अवलोकन संबंधी डेटा का उपयोग करके प्रोपेन्सिटी स्कोर मैचिंग (पीएसएम) के माध्यम से कमाई पर नौकरी प्रशिक्षण के प्रभाव का मूल्यांकन किया गया है।

अनुमान विधि	एटीटी अनुमान ($)	प्रमाप त्रुटि	t-सांख्यिकी
पीएसएम	1794	875	2.05
रेडियस (0.01)	1850	880	2.10
महालनोबिस	1680	900	1.87
ओएलस	2250	950	2.37

(क) कार्यक्रम के प्रभाव के बारे में अनुमान क्या कहते हैं?

(ख) मिलान के संदर्भ में सामान्य समर्थन की अवधारणा की व्याख्या कीजिए।

(ग) रैखिक संभाव्यता मॉडल (एलपीएम) क्या है? प्रवृत्ति स्कोर बनाने के लिए रैखिक संभाव्यता मॉडल की तुलना में लॉगिटध्प्रोबिट को क्यों प्राथमिकता दी जाती है?

(घ) इस संदर्भ में ओएलएस की तुलना में मिलान का उपयोग क्यों किया जाता है? मिलान विधियों में परिणामों में स्थिरता क्या बताती है?

(ङ) मिलान कब उपचार प्रभावों के निष्पक्ष अनुमान देने में असफल हो जाएगा? (3+4+5+3+3)

4. शोधकर्ता कॉलेज में उपस्थिति के लिए एक उपकरण के रूप में 4 साल के कॉलेज (बाइनरी चर) की निकटता का उपयोग करते हैं ताकि आय पर इसके प्रभाव का अनुमान लगाया जा सके। नीचे दिए गए परिणाम स्कूली शिक्षाधशिक्षा और आय के वर्षों पर कॉलेज के करीब होने के प्रभाव को दर्शाते हैं। अध्ययन में पर्याप्त रूप से संचालित नमूना है।

आश्रित चर	स्वतंत्र चर	गुणांक	प्रमाप त्रुटि	t-सांख्यिकी
पहला चरण (शिक्षा के वर्ष)	कॉलेज से सामीप्य	0.9	0.3	3.0
	स्थिरांक	12.5	0.8	
लघु रुप (लंबी आय)	कॉलेज से सामीप्य	0.18	0.07	2.57
	स्थिरांक	9.8	0.2	
IV अनुमान (2एसएलएस)	शिक्षा के वर्ष	0.20	0.09	2.22

(क) प्रथम चरण और लघु-रूप प्रतीपगमन से गुणांकों की व्याख्या कीजिए।

(ख) IV अनुमान की मैन्युअल रूप से गणना करने के लिए वाल्ड अनुमानक का उपयोग कीजिए।

(ग) वैध उपकरण चर (IV) अनुमान के संदर्भ में, ओएलएस की तुलना में बड़ा IV अनुमान होने का क्या कारण हो सकता है?

(घ) इंस्ट्रूमेंटल चर और प्रॉक्सी चर के बीच अंतर स्पष्ट कीजिए। क्या एक प्रॉक्सी चर एंडोजेनिटी समस्या को हल कर सकता है ?

(ङ) उन मान्यताओं को संक्षेप में समझाइए जिनके तहत ट्रीटमेंट असाइनमेंट (Z), प्रतिफल (Y) पर ट्रीटमेंट (D) के लोकल एवरेज ट्रीटमेंट इफेक्ट (LATE) का अनुमान लगाने के लिए एक उपकरण के रूप में काम कर सकता है। (4+2+3+4+5)

5. भारत में एक राज्य सरकार ने 2023 में घरेलू ऊर्जा दक्षता सब्सिडी कार्यक्रम शुरू किया। ₹5,00,000 से कम वार्षिक आय वाले परिवार ऊर्जा-कुशल उपकरण लगाने के लिए ₹20,000 तक की सब्सिडी प्राप्त करने के पात्र थे। ₹5,00,000 या उससे अधिक कमाने वाले परिवार पात्र नहीं थे।

आप किलोवाट-घंटे (kWh) में मापी गई वार्षिक घरेलू बिजली उपभोग पर सब्सिडी पात्रता के कारणात्मक प्रभाव का अनुमान लगाने में रुचि रखते हैं। ऐसा करने के लिए, आप ₹4,50,000 और ₹5,50,000 के बीच कमाने वाले परिवारों के डेटा का उपयोग करते हैं, और एक स्थानीय रैखिक प्रतीपगमन दर्शाते हैं :

$$\text{विद्युत} = \alpha + \tau . D_i + f(\text{आय} - 5{,}00{,}000) + \varepsilon_i$$

जहां :

– $D_i = 1$ यदि घर ₹5,00,000 (पात्र) से कम कमाता है, और अन्यथा 0.

– f() 5,00,000 रुपये पर केंद्रित आय का एक रैखिक फलन है।

चर	गुणांक	प्रमाप त्रुटि	p-मान
अंतःखंड	9500	100	0.000
पात्रता (D = 1)	-600	200	0.003
आय- ₹5,00,000	-0.03	0.01	0.010

(क) सब्सिडी पात्रता का बिजली उपभोग पर अनुमानित कारणात्मक प्रभाव क्या है?

(ख) क्या यह प्रभाव 5% के स्तर पर सांख्यिकीय रूप से महत्वपूर्ण है? अपना तर्क दीजिए।

(ग) सब्सिडी के लिए योग्य परिवार के लिए आय सीमा पर अनुमानित बिजली खपत क्या है?

(घ) इस संदर्भ में वैध कारणात्मक अनुमान प्राप्त करने के लिए आरडीडी डिजाइन के लिए मुख्य पहचान धारणा क्या है?

(ङ) ₹5,00,000 सीमा के आसपास आय की गलत रिपोर्टिंग आरडीडी की वैधता को कैसे प्रभावित कर सकती है? यह किस प्रकार का पूर्वाग्रह उत्पन्न कर सकता है?

(च) एक निदान परीक्षण या दृश्य विश्लेषण का वर्णन कीजिए जो यह आकलन करने में सहायक हो कि आरडीडी की धारणाएं उचित हैं या नहीं। (3 + 2 + 3 + 4 + 3 + 3)

6. 2020 में, एक शहर की सरकार ने अनौपचारिक क्षेत्र के श्रमिकों के लिए स्वास्थ्य बीमा सब्सिडी शुरू की। कार्यक्रम में नामांकित श्रमिकों को उनके बीमा प्रीमियम पर 50% सब्सिडी मिली। भागीदारी स्वैच्छिक थी।

2022 में कार्यक्रम के व्यक्तिगत स्वास्थ्य व्यय पर प्रभाव का आकलन करने के लिए, एक शोधकर्ता ने 2020 (कार्यक्रम से पहले) और 2022 (कार्यक्रम के बाद) में श्रमिकों पर डेटा एकत्र किया। शोधकर्ता ने पहले उपचार-पूर्व विशेषताओं के आधार पर मिलाए गए ट्रीटेड और नियंत्रण श्रमिकों के बीच 2022 के परिणामों की तुलना करने के लिए प्रोपेन्सिटी स्कोर मैचिंग (पीएसएम) का उपयोग किया। इसके बाद, उन्होंने पूर्ण पैनल डेटा के साथ एक डिफरेंस-इन-डिफरेंस (DiD) मॉडल का उपयोग किया।

समूह	औसत व्यय
ट्रीटेड	6200
मैच्ड कंट्रोल	7100
एटीटी	-900

चर	गुणांक	प्रमाप त्रुटि	p-मान
अंतःखंड	7300	200	0.000
पोस्ट (वर्ष 2022=1)	100	100	0.320
ट्रीटेड	-150	250	0.550
ट्रीटेड. पोस्ट	-1000	200	0.000

(क) कर्स ऑफ डायमेंशन की व्याख्या कीजिए। प्रोपेन्सिटी स्कोर मैचिंग (पीएसएम) इसे कैसे हल करता है।

(ख) एक उपयुक्त निर्देशित अचक्रीय ग्राफ (डीएजी) का उपयोग करके प्रोपेन्सिटी स्कोर के सहचर संतुलन गुण को संक्षेप में समझाइए।

(ग) आपको कौन सा अनुमान (पीएसएम/डीआईडी) इस संदर्भ में अधिक विश्वसनीय लगता है और क्यों?

(घ) डीआईडी किस अवधारणा पर निर्भर करता है जिस पर मैचिंग निर्भर नहीं करता? समझाइए कि आप इस संदर्भ में इसकी संभाव्यता का परीक्षण कैसे करेंगे?

(ङ) यदि कार्यक्रम को 3 चरणों में शुरू किया गया तो डीआईडी अनुमान किस प्रकार भिन्न होगा? संक्षेप में समझाइए। इस मामले में निष्पक्ष अनुमान को सक्षम करने के लिए किन धारणाओं की आवश्यकता होगी? (4 + 3 + 4 + 4 + 3)

(100)