[This question paper contains 12 printed pages.]

Your Roll No_____

Sr. No. of Question Paper : 6870

Unique Paper Code : 2273100027

Name of the Paper : Introduction to Causal Inference

Name of the Course : Common Pool of DSE

Semester : VI/VIII

Duration : 3 hours **Maximum Marks :** 90

Instructions

There are two Sections, A & B with a total of 6 questions.

Question 1 from Section A is compulsory.

Answer any 4 out of the 5 questions from Section B.

Marks are given at the end of each question.

Use of simple calculators is allowed.

Section A

1. (a) Define the Potential Outcomes Framework. Using this framework, explain the fundamental problem of causal inference and why the Average Treatment Effect (ATE) cannot be directly observed at the individual level.

(b) Express the Simple Difference in Outcomes (SDO) as a function of the Average Treatment Effect (ATE) and a selection bias term. What does the selection bias term capture?

(c) How does randomization of treatment assignment eliminate both selection bias and heterogeneous treatment effect bias from the SDO?

(d) In a randomized experiment, a researcher discovers post-randomization that the treatment group has significantly higher baseline education levels than the control group. What does this indicate, and how should the researcher address it?

(e) A government run vaccination drive is rolled out in rural schools. Every student in a treated school gets vaccinated (treatment), while no student in a control school does. However, students in treated schools interact with students from control schools on their way home, potentially spreading immunity via herd effects.

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 2021, State X implemented a new conditional cash transfer scheme for rural households, providing monthly payments to households who enrolled their children in school. State Y, with a similar socioeconomic profile, did not implement any such scheme. You have panel data from both states for 2020 (pre-treatment) and 2022 (post-treatment). The outcome variable is the average monthly per capita consumption expenditure (in ₹). The regression output from a DiD model is shown below.

Variable	Coefficient	Std. Error	p-value
Intercept	3200	110	0.000
Post (2022 = 1)	180	130	0.165
Treated (State X = 1)	-90	120	0.450
Treated × Post	640	195	0.001

(a) Provide the estimating equation for a DiD estimation of this kind. What is the estimated effect of the conditional cash transfer scheme on monthly per capita consumption? Interpret the DiD coefficient.

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

(c) What is the predicted average per capita consumption in 2022 for households in State X?

(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 drought occurred in State X between 2020 and 2022, independently affecting household expenditure. How could this confound the DiD estimate?

(4 + 2 + 2 + 3 + 3 + 4)

3. The following table is based on a study evaluating the effect of a microfinance program on annual household income via Propensity Score Matching (PSM) using observational data from a developing country.

Estimation Method	ATT Estimate (₹)	Standard Error	t-statistic
PSM (Nearest Neighbor)	8,450	3,200	2.64
Radius Matching (0.02)	8,100	3,400	2.38
Kernel Matching	7,950	3,100	2.56
OLS	12,400	4,200	2.95

(a) What do the estimates say about the program's impact on annual household income?

(b) Explain the concept of common support in the context of propensity score matching. Why is it important?

(c) What is a Linear Probability Model (LPM)? Why is logit/probit preferred over an LPM for estimating the propensity score?

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

(e) What is the Conditional Independence Assumption (CIA)? Under what conditions will matching fail to give unbiased estimates of treatment effects?

(3 + 4 + 5 + 3 + 3)

4. Researchers use distance to the nearest district hospital (a continuous variable) as an instrument for health insurance enrollment to estimate its effect on out-of-pocket medical expenditure. The results below show the impact of distance on health insurance enrollment and medical expenditure.

Dependent Variable	Independent Variables	Coefficient	Std. Error	t-stat
First Stage (Insurance Enrollment)	Distance to hospital (km)	-0.012	0.004	-3.00
	Constant	0.65	0.05	
Reduced Form (Log Medical Expenditure)	Distance to hospital (km)	0.009	0.003	3.00
	Constant	7.2	0.15	
IV Estimate (2SLS)	Insurance Enrollment	-0.75	0.28	-2.68

(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 control variable. Can a control variable solve the endogeneity issue?

(e) Explain briefly the assumptions under which the instrument (distance to hospital) can be used to estimate a Local Average Treatment Effect (LATE), and characterize the complier population in this context.

(4 + 2 + 3 + 4 + 5)

5. A municipal government in Maharashtra introduced a free mid-day meal enhancement program in 2022. Primary schools with an average daily attendance below 70% were eligible for the enhanced program. Schools with attendance of 70% or above were not eligible.

You are interested in estimating the causal effect of the program on average student learning test scores. You use data from schools with attendance between 60% and 80%, and run a local linear regression:

$Score_i = \alpha + \tau \cdot D_i + f(Attendance - 70) + \varepsilon_i$

Where:

$D_i = 1$ if the school has attendance below 70% (eligible), and 0 otherwise.

f() is a linear function of attendance centered at 70%.

Variable	Coefficient	Std. Error	p-value
Intercept	52.0	1.8	0.000
Eligibility (D = 1)	5.8	2.2	0.010
Attendance − 70	0.6	0.15	0.001

(a) What is the estimated causal effect of the enhanced meal program on student test scores?

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

(c) What is the predicted test score at the attendance threshold for a school that just qualifies for the program?

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

(e) How might schools strategically manipulating their reported attendance around the 70% 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 RDD assumptions are plausible.

(3 + 2 + 3 + 4 + 3 + 3)

6. In 2021, a state government launched a subsidized vocational skills training program for unemployed youth. Enrollment was voluntary. To estimate the program's effect on monthly income in 2023, a researcher collected panel data from 2021 (before) and 2023 (after). The researcher first applied Propensity Score Matching (PSM) to compare 2023 outcomes between participants and matched non-participants. Then, they used a Difference-in-Differences (DiD) model on the full panel.

Group	Mean Monthly Income (₹)
Treated (Participants)	18,500
Matched Control	15,200
ATT	3,300

Variable	Coefficient	Std. Error	p-value
Intercept	14,800	400	0.000
Post (Year 2023 = 1)	350	280	0.211
Treated	500	500	0.315
Treated × Post	2,900	420	0.000

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

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

(c) Which estimate (PSM or 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 DiD estimation would differ if the program was rolled out in three phases across different districts. What assumptions would be needed for unbiased estimation in this case? (4 + 3 + 4 + 4 + 3)

निर्देश

इसमें दो खंड हैं, ए और बी, जिनमें कुल 6 प्रश्न हैं।

खंड क का प्रश्न 1 अनिवार्य है।

खंड ख के 5 प्रश्नों में से किन्हीं 4 प्रश्नों के उत्तर दीजिए।

प्रत्येक प्रश्न के अंत में अंक दिए गए हैं।

साधारण कैलकुलेटर का उपयोग करने की अनुमति है।

खंड क

1.(क) संभावित परिणाम ढाँचे को परिभाषित कीजिए। इस ढाँचे का उपयोग करते हुए, कारण-कार्य संबंध अनुमान की मूलभूत समस्या और यह स्पष्ट कीजिए कि औसत उपचार प्रभाव (एटीई) को व्यक्तिगत स्तर पर प्रत्यक्ष रूप से क्यों नहीं देखा जा सकता है।

(ख) परिणामों में सरल अंतर (एसडीओ) को औसत उपचार प्रभाव (एटीई) और चयन पूर्वाग्रह पद के फलन के रूप में व्यक्त कीजिए। चयन पूर्वाग्रह पद क्या दर्शाता है?

(ग) उपचार आवंटन का यादृच्छिकीकरण एसडीओ से चयन पूर्वाग्रह और विषम उपचार प्रभाव पूर्वाग्रह दोनों को कैसे समाप्त करता है?

(घ) एक यादृच्छिक प्रयोग में, एक शोधकर्ता को यादृच्छिकीकरण के बाद पता चलता है कि उपचार समूह में नियंत्रण समूह की तुलना में आधारभूत शिक्षा स्तर काफी अधिक है। यह क्या दर्शाता है, और शोधकर्ता को इसका समाधान कैसे करना चाहिए?

(ङ) ग्रामीण विद्यालयों में सरकार द्वारा संचालित टीकाकरण अभियान चलाया जाता है। उपचारित विद्यालय के प्रत्येक छात्र को टीका लगाया जाता है (उपचार), जबकि नियंत्रण विद्यालय के किसी भी छात्र को टीका नहीं लगाया जाता है। हालांकि, उपचारित विद्यालयों के छात्र घर जाते समय नियंत्रण विद्यालयों के छात्रों के साथ संपर्क में आते हैं, जिससे सामूहिक प्रभाव के माध्यम से प्रतिरक्षा फैलने की संभावना रहती है।

यहां SUTVA के किस पहलू का उल्लंघन हो रहा है? इस समस्या का विश्लेषणात्मक रूप से या अध्ययन डिजाइन में समाधान करने के लिए शोधकर्ता द्वारा अपनाई जा सकने वाली एक रणनीति का संक्षेप में वर्णन कीजिए। (6 + 3 + 2 + 4 + 3)

खंड ख

2. 2021 में, राज्य X ने ग्रामीण परिवारों के लिए एक नई सशर्त नकद हस्तांतरण योजना लागू की, जिसके तहत अपने बच्चों को स्कूल में दाखिला दिलाने वाले परिवारों को मासिक भुगतान किया जाता है। समान सामाजिक-आर्थिक पृष्ठभूमि वाले राज्य Y ने ऐसी कोई योजना लागू नहीं की। आपके पास दोनों राज्यों के 2020 (उपचार-पूर्व) और 2022 (उपचार-पश्चात)

के पैनल डेटा उपलब्ध हैं। परिणाम चर प्रति व्यक्ति औसत मासिक उपभोग व्यय (₹ में) है। DiD मॉडल से प्राप्त प्रतिगमन परिणाम नीचे दिखाया गया है।

चर	गुणक	मानक त्रुटि	पी-मान
अवरोधन	3200	110	0.000
बाद में (2022 = 1)	180	130	0.165
उपचारित (अवस्था X = 1)	-90	120	0.450
उपचारित × बाद में	640	195	0.001

(a) इस प्रकार के DiD अनुमान के लिए अनुमान समीकरण प्रदान कीजिए। सशर्त नकद हस्तांतरण योजना का प्रति व्यक्ति मासिक उपभोग पर अनुमानित प्रभाव क्या है? DiD गुणांक की व्याख्या कीजिए।

(ख) क्या यह प्रभाव 5% स्तर पर सांख्यिकीय रूप से महत्वपूर्ण है? अपने तर्क को स्पष्ट कीजिए।

(ग) राज्य X में परिवारों के लिए 2022 में अनुमानित औसत प्रति व्यक्ति उपभोग क्या है?

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

(ङ) इस संदर्भ में समानांतर रुझानों की धारणा की विश्वसनीयता का परीक्षण करने के लिए आप जिस विधि या डेटा विशेषता का उपयोग कर सकते हैं, उसका सुझाव दें।

(च) मान लीजिए कि 2020 और 2022 के बीच राज्य X में एक भीषण सूखा पड़ा, जिससे घरेलू व्यय स्वतंत्र रूप से प्रभावित हुआ। इससे DiD अनुमान किस प्रकार भ्रमित हो सकता है? (4 + 2 + 2 + 3 + 3 + 4)

3. निम्नलिखित तालिका एक विकासशील देश से प्राप्त अवलोकन डेटा का उपयोग करते हुए प्रोपेन्सिटी स्कोर मैचिंग (पीएसएम) के माध्यम से वार्षिक घरेलू आय पर एक सूक्ष्म वित्त कार्यक्रम के प्रभाव का मूल्यांकन करने वाले अध्ययन पर आधारित है।

अनुमान विधि	एटीटी अनुमान (₹)	मानक त्रुटि	टी आँकड़ा
पीएसएम (निकटतम पड़ोसी)	8,450	3,200	2.64
त्रिज्या मिलान (0.02)	8,100	3,400	2.38
कर्नेल मिलान	7,950	3,100	2.56
ओएलएस	12,400	4,200	2.95

(क) अनुमानों से वार्षिक घरेलू आय पर कार्यक्रम के प्रभाव के बारे में क्या पता चलता है?

(ख) प्रोपेंसिटी स्कोर मैचिंग के संदर्भ में कॉमन सपोर्ट की अवधारणा को स्पष्ट कीजिए। यह महत्वपूर्ण क्यों है?

(ग) लीनियर प्रोबेबिलिटी मॉडल (LPM) क्या है? प्रोपेंसिटी स्कोर का अनुमान लगाने के लिए LPM की तुलना में लॉजिट/प्रोबिट मॉडल को क्यों प्राथमिकता दी जाती है?

(घ) इस संदर्भ में ओएलएस की तुलना में मैचिंग का उपयोग क्यों किया जाए? मैचिंग विधियों में परिणामों की संगति क्या दर्शाती है?

(ङ) सशर्त स्वतंत्रता धारणा (सीआईए) क्या है? किन परिस्थितियों में मिलान विधि उपचार प्रभावों के निष्पक्ष अनुमान देने में विफल रहेगी? (3 + 4 + 5 + 3 + 3)

4.शोधकर्ता स्वास्थ्य बीमा नामांकन के लिए निकटतम जिला अस्पताल की दूरी (एक सतत चर) का उपयोग जेब से होने वाले चिकित्सा व्यय पर इसके प्रभाव का आकलन करने के लिए करते हैं। नीचे दिए गए परिणाम स्वास्थ्य बीमा नामांकन और चिकित्सा व्यय पर दूरी के प्रभाव को दर्शाते हैं।

आश्रित चर	स्वतंत्र प्रभावित करने वाली वस्तुएँ	गुणक	मानक त्रुटि	टी स्टेट
पहला चरण (बीमा नामांकन)	अस्पताल की दूरी (किमी)	-0.012	0.004	-3.00

	स्थिर	0.65	0.05	
संक्षिप्त रूप (चिकित्सा व्यय लॉग)	अस्पताल की दूरी (किमी)	0.009	0.003	3.00
	स्थिर	7.2	0.15	
IV अनुमान (2SLS)	बीमा नामांकन	-0.75	0.28	-2.68

(क) प्रथम चरण और सरलीकृत रूप प्रतिगमन से गुणांकों की व्याख्या कीजिए।

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

(ग) वैध IV अनुमान के संदर्भ में, OLS की तुलना में बड़े IV अनुमान को क्या समझा सकता है?

(घ) इंस्ट्रुमेंटल वेरिएबल और कंट्रोल वेरिएबल के बीच अंतर स्पष्ट कीजिए। क्या कंट्रोल वेरिएबल एंडोजेनिटी की समस्या का समाधान कर सकता है?

(ङ) संक्षेप में उन मान्यताओं की व्याख्या कीजिए जिनके अंतर्गत उपकरण (अस्पताल से दूरी) का उपयोग स्थानीय औसत उपचार प्रभाव (एलएटीई) का अनुमान लगाने के लिए किया जा सकता है, और इस संदर्भ में अनुपालन करने वाली आबादी की विशेषताओं का वर्णन कीजिए।

(4 + 2 + 3 + 4 + 5)

5. महाराष्ट्र की एक नगरपालिका सरकार ने 2022 में मुफ्त मध्याह्न भोजन संवर्धन कार्यक्रम शुरू किया। जिन प्राथमिक विद्यालयों में दैनिक उपस्थिति औसत 70% से कम थी, वे इस संवर्धन कार्यक्रम के लिए पात्र थे। जिन विद्यालयों में उपस्थिति 70% या उससे अधिक थी, वे पात्र नहीं थे।

आप छात्रों के औसत अधिगम परीक्षण अंकों पर कार्यक्रम के कारण होने वाले प्रभाव का अनुमान लगाने में रुचि रखते हैं। आप 60% और 80% के बीच उपस्थिति वाले विद्यालयों के डेटा का उपयोग करते हैं और एक स्थानीय रैखिक प्रतिगमन चलाते हैं:

स्कोर$_i$ = α + $\tau \cdot D_i$ + f(उपस्थिति - 70) + ε_i

जिसमें:

D_i = 1 यदि विद्यालय में उपस्थिति 70% से कम है (पात्रता), अन्यथा 0।

f() उपस्थिति का एक रैखिक फलन है जिसका केंद्र 70% है।

चर	गुणक	मानक त्रुटि	पी-मान
अवरोधन	52.0	1.8	0.000
पात्रता (डी = 1)	5.8	2.2	0.010
उपस्थिति − 70	0.6	0.15	0.001

(a) उन्नत भोजन कार्यक्रम का छात्रों के परीक्षा अंकों पर अनुमानित कारण प्रभाव क्या है?

(ख) क्या यह प्रभाव 5% स्तर पर सांख्यिकीय रूप से महत्वपूर्ण है? अपने तर्क प्रस्तुत कीजिए।

(ग) उस विद्यालय के लिए उपस्थिति सीमा पर अनुमानित परीक्षा स्कोर क्या है जो कार्यक्रम के लिए अर्हता प्राप्त करता है?

(घ) इस संदर्भ में वैध कारण अनुमान प्राप्त करने के लिए आर.डी.डी. डिजाइन की प्रमुख पहचान संबंधी धारणा क्या है?

(ङ) स्कूलों द्वारा 70% की सीमा के आसपास अपनी रिपोर्ट की गई उपस्थिति में रणनीतिक रूप से हेरफेर करने से आरडीडी की वैधता कैसे प्रभावित हो सकती है? इससे किस प्रकार का पूर्वाग्रह उत्पन्न हो सकता है?

(च) एक नैदानिक परीक्षण या दृश्य विश्लेषण का वर्णन कीजिए जो यह आकलन करने में मदद कर सकता है कि क्या आर.डी.डी. की मान्यताएँ प्रशंसनीय हैं। (3+2+3+4+3+3)

6. 2021 में, एक राज्य सरकार ने बेरोजगार युवाओं के लिए रियायती व्यावसायिक कौशल प्रशिक्षण कार्यक्रम शुरू किया। इसमें नामांकन स्वैच्छिक था। 2023 में मासिक आय पर कार्यक्रम के प्रभाव का अनुमान लगाने के लिए, एक शोधकर्ता ने 2021 (पहले) और 2023 (बाद) के पैनल डेटा एकत्र किए। शोधकर्ता ने सबसे पहले प्रतिभागियों और मिलान किए गए गैर-प्रतिभागियों के बीच 2023 के परिणामों की तुलना करने के लिए प्रोपेंसिटी स्कोर मैचिंग (PSM) का उपयोग किया। फिर, उन्होंने पूरे पैनल पर डिफरेंस-इन-डिफरेंसेस (DiD) मॉडल का उपयोग किया।

समूह	औसत मासिक आय (₹)
उपचारित (प्रतिभागी)	18,500
मिलान नियंत्रण	15,200
एटीटी	3,300

चर	गुणक	मानक त्रुटि	पी-मान
अवरोधन	14,800	400	0.000
बाद में (वर्ष 2023 = 1)	350	280	0.211
इलाज	500	500	0.315
उपचारित × पोस्ट	2,900	420	0.000

(क) आयामी अभिशाप की व्याख्या कीजिए और चर्चा कीजिए कि प्रोपेंसिटी स्कोर मैचिंग (पीएसएम) इस समस्या का समाधान कैसे करती है।

(ख) निर्देशित अचक्रीय ग्राफ (डीएजी) का उपयोग करते हुए, प्रवृत्ति स्कोर के सहचर संतुलन गुण की संक्षेप में व्याख्या कीजिए।

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

(घ) DiD किस मान्यता पर आधारित है जिस पर मिलान आधारित नहीं है? इस संदर्भ में आप इसकी विश्वसनीयता का परीक्षण कैसे कीजिएगे, यह स्पष्ट कीजिए।

(ङ) संक्षेप में समझाइए कि यदि कार्यक्रम को विभिन्न जिलों में तीन चरणों में लागू किया जाता है तो डीआईडी अनुमान किस प्रकार भिन्न होगा। इस स्थिति में निष्पक्ष अनुमान के लिए किन मान्यताओं की आवश्यकता होगी? (4 + 3 + 4 + 4 + 3)

(200)