Midterm

EC 320 | May 8th, 2023

Instructor: Andrew Dickinson

Name: _______________________ Student ID: _______________________

Please do not open the exam until you are told to do so. Write your full name and student ID above.

Points: The total points possible on this exam is **100**. The following table explains the point breakdown.

Section	Points
Multiple Choice	25
T/F	25
Short Answer	50
Total	**100**

You may *not* use books, notes, or outside resources during this exam. You are required to show your work on each short answer problem for full points. Partial credit will be awarded so do your best on every question and do not leave anything blank! Points may be lost due to illegible answers at the grader's discretion. Best to make sure that everything is written **clearly** so it is easy for the grader to understand. Use the space provided to answer each question. The use of scratch paper is encouraged and will be provided at request.

MC

Select one and only one of the following options. Each question is worth 5 points. 25 points total.

[01] A research group is conducting a clinical trial to evaluate a new drug for treating high blood pressure. After analyzing the trial results, the research group concludes that the drug is effective. In reality, the drug does not have any effect on blood pressure. What type of error has occurred in this situation?

A. Type I error

B. Type II error

C. Type III error

D. Sampling error

E. No error

[02] Which of the following regression specifications fail the first classical assumption–that the population relationship is *linear in parameters*

A. $\sqrt{\text{Convictions}_i} = \beta_1 + \beta_2(\text{Early Childhood Lead Exposure})_i + u_i$

B. $\text{Wage}_i = \beta_1 + \beta_2 \text{Experience}_i + u_i$

C. $\log(\text{Happiness}_i) = \beta_1 + \beta_2 \log(\text{Money}_i) + u_i$

D. $\text{Wage}_i = (\beta_1 + \beta_2 \text{Experience}_i) u_i$

E. $\log(\text{Earnings}_i) = \beta_1 + \beta_2 \text{Education}_i^2 + u_i$

[03] When conducting OLS, our fitted values generate some misses (i.e., the difference between the observed value of the dependent variable and the predicted value). We call these misses:

A. Coefficients

B. Standard errors

C. Residuals

D. Standard deviations

E. Variance

[04] Which of the following equations describe homoskedasticity.

A. $\text{Var}(u_i|X) = \sigma^2$

B. $\text{Var}(u_i|X) = \sigma_i^2$

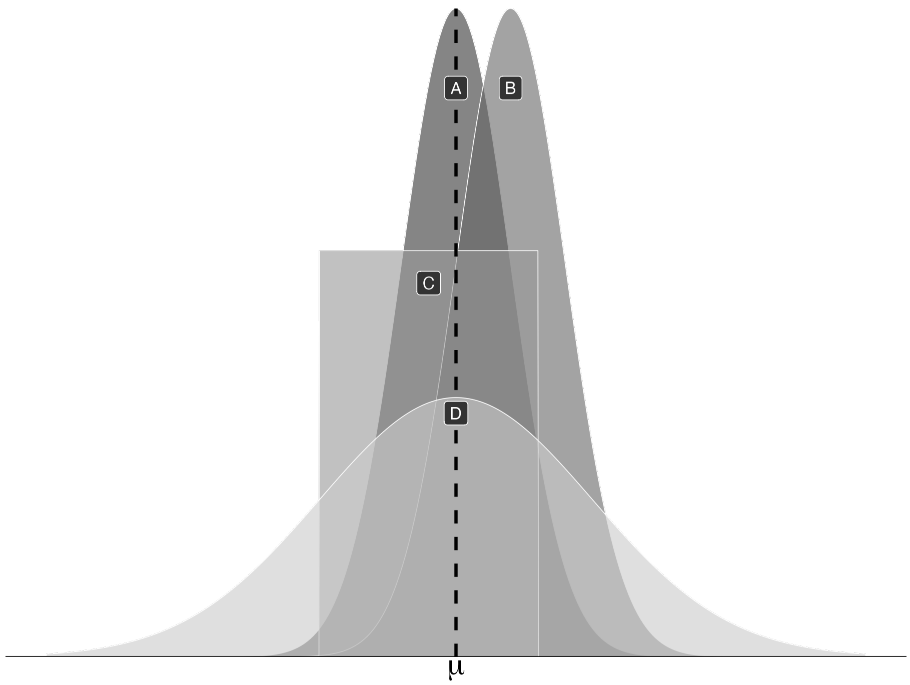
C. $\text{Var}(u_i|X) = \sigma^2 X$

D. $\text{Var}(u_i|X) = \beta X + \sigma^2$

E. $\text{Var}(u_i|X) = \beta^2$

[05] Observe the figure below. Each distribution describes an estimator. Choose the best unbiased estimator with the corresponding label. (*Please raise your hand if you find the figure difficult to read as I can cast it on the projector.*)

◯ A. ◯ B. ◯ C. ◯ D.



T/F

Circle true [T] or false [F] below. Each question is worth 5 points. 25 points total.

[06] T F A *continuous random variable* is a random variable that takes any real value with *zero probability.*

[07] T F $E[X]^2$ is equivalent to $E[X^2]$?

[08] T F In the potential outcomes framework, the *counterfactual* is a statistical method for determining the correlation between two variables by analyzing the frequency distribution of the observed data.

[09] T F It is typical, and important, of regression estimates with a causal interpretation to have a high R^2.

[10] T F Greater variation in X_i increases the variance of our OLS slope parameter.

Short answer

Provide a written response to each question in the space provided. Show all work and clearly mark your answers.

[11] (10 points) **Random variables**

Suppose we have three random variables, Y_i, X_i, and u_i for which the data generating process is known—u_i is distributed normally with mean of 0 and standard deviation of 1 and is statistically independent, X_i is distributed normally with mean of 5 and standard deviation of 2, and $Y_i = 12 + 5X_i + u_i$.

[11a] (2 points) Find the expected value of Y_i

[11b] (2 points) Find the variance of Y_i

[11c] (3 points) Suppose we collected a 1,000 observation sample of these variables. If we regress Y_i on X_i and find the OLS estimates $\hat{Y}_i = \beta_0 + \beta_1 \hat{X}_i$, if OLS is unbiased, will $\hat{\beta}_1 = 5$? Explain.

[11d] (3 points) Will $E[\hat{\beta}_1] = 5$? Explain.

[12] (10 points) **Potential outcomes framework**

Suppose we are interested in measuring the effect of government sponsored work training programs. We would like to describe, in the potential outcomes framework, the treatment effect τ. Suppose that $Y_{0,i}$ is the potential hourly earnings of individual i when they are not treated, $Y_{1,i}$ is the potential hourly earnings of individual i when they are treated, and D_i is an indicator for the true treatment status for individual i. Assume that τ is constant across individuals.

[12a] (2 points) Recall the difference-in-means estimator:

$$Avg\left(y_i \mid D_i = 1\right) - Avg\left(y_i \mid D_i = 0\right)$$

From this estimator, derive an expression that describes how this estimator can explained by τ and *selection bias*.

[12b] (2 points) In a few sentences, describe the fundamental problem of causal inference? Relate it to the current context of work training programs.

[12c] (2 points) Suppose we can observe the following data on hourly earnings.

i	D_i	$Y_{0,i}$	$Y_{1,i}$
1	0	29.9	NA
2	0	30.0	NA
3	0	31.0	NA
4	0	29.8	NA
5	1	NA	26.9
6	1	NA	25.2
7	1	NA	25.7
8	1	NA	25.0

where NA means the data point is unobserved. What is the average treatment effect (ATE)?

[12d] (4 points) Can we interpret this treatment effect as causal?

- If yes, does it seem that work training programs are a good investment? (*pick one*)
- If no, what are some potential sources of selection bias? (*pick one*)

[13] (10 points) **Interpreting OLS estimates**

Suppose we regress hourly earnings on the number of years of education for a sample of 96 people, specified in the following model:

$$\log(\text{Earnings}_i) = \beta_0 + \beta_1 \text{Education}_i + u_i$$

The results of the regression are exported to the following table:

Dependent Variable:	log(Earnings)
Variables	
Intercept term	1.342***
	(0.1053)
Education	0.1147***
	(0.0074)
Fit statistics	
N	96
R^2	0.72154

Standard-errors in parentheses
*Significance Codes: ***: 0.01, **: 0.05, *: 0.1*

[13a] (2 points) Interpret the intercept term β_0 given in the table above.

[13b] (2 points) Interpret the slope parameter β_1 given in the table above.

[13c] (2 points) Interpret the R^2 given in the table above.

[13d] (4 points) What assumptions must we make in order for OLS to be unbiased? List them below with a brief explanation of each one.

[14] (10 points) **Goodness of fit.** (*The equations for TSS, ESS, and RSS are found on the last page*)

[14a] (3 points) In a few sentences, describe the intuition behind what TSS, ESS, and RSS measure.

[14b] (5 points) Show that $TSS = ESS + RSS$

(*More space provided on the back*)

[14c] (2 points) Write one formula for R^2 in terms of TSS, ESS, and RSS. Why do we typically express goodness of fit with R^2 rather than TSS, ESS, and RSS?

[15] (10 points) **Deriving OLS.**

In a simple linear regression model, the intercept term represents the expected value of the dependent variable when the independent variable is zero. However, in some cases, it might not make sense to have an intercept or assume that the dependent variable has a non-zero value when the independent variable is zero. In such situations, a simple linear regression without an intercept might be more appropriate. Suppose we have the following simple linear regression model without an intercept:

$$Y_i = \beta X_i + u_i$$

where the corresponding residuals are written as:

$$\hat{u} = Y_i - \beta X_i$$

In this question, we will derive the OLS estimate of β. (*Hint: This derivation follows the derivation of OLS with an intercept—with a lot less algebra. The result will look different yet be functionally equivalent of the standard result*)

Note: I am being loose here by naming my estimate in the residuals equation β and not $\widehat{\beta}$. Feel free to change it to $\widehat{\beta}$ or just b. It does not matter and you will not lose points so long as you are consistent.

[15a] (2 points) Describe in words what the objective of the OLS estimator is and how the first order condition reaches that objective.

[15b] (4 points) Set up and solve the first order condition. (i.e. Find $\frac{\partial \text{RSS}}{\partial \hat{\beta}} = 0$)

[15c] (4 points) Solve for the simple OLS estimator (i.e., Solve for β from the first order condition above).

Formulas

Goodness of fit:

$$\text{TSS} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$\text{RSS} = \sum_{i=1}^{n} \hat{u}_i^2$$

$$\text{ESS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

OLS formulas:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$