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.** $\textbf{Wage}_i = (\beta_1 + \beta_2\textbf{Experience}_i)u_i$
- 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:

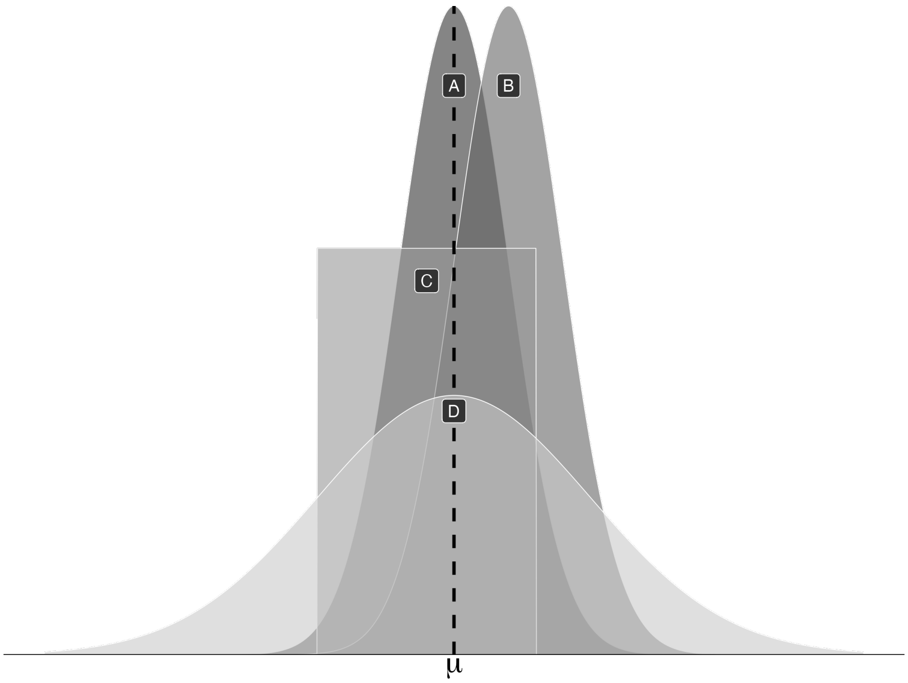
- A. Coefficients
- B. Standard errors
- **C. Residuals**
- D. Standard deviations
- E. Variance

[04] Which of the following equations describe homoskedasticity.

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

Solution:

$$\begin{aligned} E[Y_i] &= E[12 + 5X_i + u_i] \\ &= 12 + 5 \times E[X_i] + E[u_i] \\ &= 12 + 5 \times 5 + 0 = 37 \end{aligned}$$

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

Solution:

$$\begin{aligned} Var(Y_i) &= Var(12 + 5X_i + u_i) \\ &= 5^2 \cdot Var(X_i) + Var(u_i) \\ &= 25 \cdot 2^2 + 1^2 = 25 \cdot 4 + 1 \\ &= 101 \end{aligned}$$

[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.

Solution: In an OLS regression, the true value of the slope parameter β_1 cannot be guaranteed to be exactly equal to the estimated value $\hat{\beta}_1$ in a finite sample. If OLS is unbiased, $\hat{\beta}_1$ might not be exactly equal to 5, it is expected to be close within a certain amount of error.

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

Solution: The OLS estimator $\hat{\beta}_1$ is unbiased, meaning that its expected value is equal to the true value of the slope parameter. Therefore, $E[\hat{\beta}_1] = \beta_1 = 5$.

[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*.

Solution:

$$\begin{aligned} Avg\left(y_i \mid D_i = 1\right) - Avg\left(y_i \mid D_i = 0\right) & \\ &= Avg\left(y_{1,i} \mid D_i = 1\right) - Avg\left(y_{0,i} \mid D_i = 0\right) \\ &= Avg\left(\tau \ + \ y_{0,i} \mid D_i = 1\right) - Avg\left(y_{0,i} \mid D_i = 0\right) \\ &= \tau + Avg\left(y_{0,i} \mid D_i = 1\right) - Avg\left(y_{0,i} \mid D_i = 0\right) \\ &= \text{Average causal effect} + \text{Selection bias} \end{aligned}$$

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

Solution: The fundamental problem of causal inference is that we cannot observe the counterfactual outcome for each individual, i.e., we cannot observe both the treated and untreated outcomes for the same individual simultaneously. In the context of work training programs, this means we cannot observe the earnings of an individual with and without the training program at the same time, which makes it difficult to determine the causal effect of the program on earnings.

[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)?

Solution: The average treatment effect (ATE) can be calculated as the average of treated outcomes minus the average of untreated outcomes:

ATE = $\frac{26.9+25.2+25.7+25.0}{4} - \frac{29.9+30.0+31.0+29.8}{4} = -4.475$

So, the average treatment effect is -4.475.

[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*)

Solution: We cannot interpret the calculated treatment effect as causal without further information. Potential sources of selection that individuals who participated in the work training program were more likely to have lower initial earnings. Thus the negative treatment effect is likely indicative of this

[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:

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

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.

Solution: Someone who has no education has an expected log earnings if approximately 1.342. *Go easy on this one.*

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

Solution: For every additional year of education, an individual's earnings are associated with an increase of 0.1147 percent, holding all else constant.

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

Solution: The $R^2 = 0.72154$ represents the proportion of the total variation in log earnings that is explained by the model. In this case, the model explains approximately 72.15

[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.

Solution: To interpret these results causally, we must make the following assumptions:

1. Linearity: The relationship between education and log earnings is linear in the population.

2. Sample variation in X_i: There must exist some variation in the independent variable.

3. Exogeniety: X_i is exogenous.

[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.

Solution: TSS measures the total variation in the dependent variable. ESS represents the portion of the total variability that can be attributed to the model, capturing the difference between the predicted values and the mean of the dependent variable. RSS measures the unexplained variation, which is the difference between the actual values and the predicted values from the model.

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

Solution:

$$\begin{aligned}
\text{TSS} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\
&= \sum_{i=1}^{n}([\hat{Y}_i + \hat{u}_i] - [\bar{\hat{Y}} + \bar{\hat{u}}])^2 \\
&= \sum_{i=1}^{n}\left([\hat{Y}_i - \bar{Y}] + \hat{u}_i\right)^2 \qquad \text{Since } \bar{\hat{u}} = 0 \text{ which implies } \bar{Y} = \bar{\hat{Y}} \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\left((\hat{Y}_i - \bar{Y})\hat{u}_i\right) \\
&= \text{ESS} + \text{RSS} + 2\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i - 2\bar{Y}\sum_{i=1}^{n}\hat{u}_i
\end{aligned}$$

Now we must show that $2\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i - 2\bar{Y}\sum_{i=1}^{n}\hat{u}_i = 0$. We must do so by using the following properties of OLS, $\sum_{i=1}^{n}\hat{u}_i = 0$ and $\sum_{i=1}^{n}X_i\hat{u}_i = 0$. Thus, notice that by the first property, $-2\bar{Y}\sum_{i=1}^{n}\hat{u}_i = 0$ since $\sum_{i=1}^{n}\hat{u}_i = 0$. Now substitute for $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

$$\begin{aligned}
2\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i - 2\bar{Y}\sum_{i=1}^{n}\hat{u}_i &= 2\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i)\hat{u}_i \\
&= 2\hat{\beta}_0\sum_{i=1}^{n}\hat{u}_i + 2\hat{\beta}_1\sum_{i=1}^{n}X_i\hat{u}_i = 0
\end{aligned}$$

(*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?

Solution:

$$R^2 = \frac{ESS}{TSS} \quad \text{or} \quad R^2 = 1 - \frac{RSS}{TSS}$$

The reason we express goodness of fit with R^2 rather than TSS, ESS, and RSS is for ease of interpretation. It can be difficult to be informed about sum of squares of variables with different units, so we standardize them with R^2.

[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 $\hat{\beta}$. Feel free to change it to $\hat{\beta}$ 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.

> **Solution:** The objective of the OLS estimator, in the context of no intercept, is to find the β that minimizes the sum of the squared residuals. The intuition behind the FOC is that a minimum will occur when the derivative of the RSS with respect to β will be equal to zero. This is smooth as a function of β, so all minima (or maxima) occur when the derivative is zero.

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

> **Solution:**
>
> To set up the first order condition, we pick a β that minimizes the sum of squared errors. Setting up the RSS
>
> $$\text{RSS}(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$
>
> We can differentiate the RSS function with respect to β and set the resulting expression to zero. This will help us find the minimum of the function.
>
> $$\frac{\partial \text{SSE}(\beta)}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - \beta x_i) = 0$$

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

Solution: Solving for β:

$$\sum_{i=1}^{n} x_i y_i - \beta \sum_{i=1}^{n} x_i^2 = 0$$

Rearranging to isolate β:

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Thus, the OLS estimator for a simple linear regression without an intercept is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

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:

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

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