Econ 140
Summer 2022
Instructor: Fernando Hoces de la Guardia
GSIs: Elena Stacy & Yige Wang

Midterm Exam 2

Tuesday July 26, 2022

Student Name:

Student ID Number:

Exam Instructions:

- **You have 80 minutes to answer this exam**
- **This exams has a total of 80 points (suggesting the length of time to spent in each question). Each question indicates the number of points, and indicates a maximum length for its answer**
- **Most questions ask for short answers (from a couple of words, to one or two sentence maximum)**
- **Explanation in black or blue ink is recommended as these often scan the best.**
- **You must submit your solutions using the exam packet provided.**
- **Do not write your solutions on pages that say "Do not write solutions on this page"**. Answers written on these pages will not be graded. You may use these pages as scratch paper.
- **When time is called, STOP** writing, immediately **CLOSE** your exam packet and hold it up until it is collected.
- **Show your work**. Credit will only be awarded on the basis of what is written on the exam.
- **Sign the academic honesty pledge**. Cheating will be punished.

Affirm the academic honesty pledge below. For those writing on a non-printed copy, please just write "Academic Honesty Pledge as on exam", and sign your name.
If you do not affirm this pledge, your exam will be marked invalid.

0. ACADEMIC HONESTY PLEDGE

I confirm that I have abided by all academic honesty rules for UC Berkeley and Economics 140. I confirm that I did not see this exam before my official exam start time. I confirm that I have not shared and will not share this exam with anyone else. I confirm that I haven't copied from anybody else's exam.

Signature: _____

Formula for the SE of $\widehat{\beta}$ in a bi-variate regression:

$$SE(\widehat{\beta}) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_X}$$

Model	β_1 **Interpretation**
Level-level $Y_i = \beta_0 + \beta_1 X_i + e_i$	$\Delta Y = \beta_1 \cdot \Delta X$ *A one-unit increase in X leads to a β_1-unit increase in Y*
Log-level $\log(Y_i) = \beta_0 + \beta_1 X_i + e_i$	$\%\Delta Y = 100 \cdot \beta_1 \cdot \Delta X$ *A one-unit increase in X leads to a $\beta_1 \cdot 100\%$ increase in Y*
Log-log $\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + e_i$	$\%\Delta Y = \beta_1 \cdot \%\Delta X$ *A one-percent increase in X leads to a $\beta_1\%$ increase in Y*
Level-log $Y_i = \beta_0 + \beta_1 \log(X_i) + e_i$	$\Delta Y = (\beta_1 \div 100) \cdot \%\Delta X$ *A one-percent increase in X leads to a $\beta_1 \div 100$-unit increase in Y*

1. For the Oregon Health Plan Experiment (OHP), describe in one sentence, one of the two key policy questions that the experiment shed light on (considered only as relevant information from Ch1).[2pts, 1 sentence]

Solution: Two possible solutions:

- OHP studies studies the effect of receiving an offer of free health insurance, among a vulnerable population in Oregon, on their (mental and physical) health.
- OHP studies studies the effect of receiving an offer of free health insurance, among a vulnerable population in Oregon, on their financial security (or financial health).

2. RCTs are simple in logic, but difficult in logistics. In the RAND Health Insurance Experiment (HIE) discussed in class, the original design had a total of 14 treatments, but when preforming the analysis, this 14 were groups into 4 broader categories. The researchers did this to address which of the following concepts discussed in class [2pts]

   a) Increase sample size to address the fundamental problem of identification
   b) Reduce the logistical burden of tracking individuals over 14 different insurance services.
   c) Increase sample size so the estimated coefficients can be approximated by well-known distributions.
   d) Reduce variation in the error term.

Solution: C

3. When conducting a hypothesis test in a regression analysis our main goal is [2pts]:

   a) Check if the estimated coefficient is consistent with some population parameter
   b) Check if an observed statistic is large in magnitude
   c) Check if the t-statistic is larger than 2 or smaller than -2
   d) Check if the population coefficient is consistent with a range around our estimated parameter.

Solution: A

4. Explain in one or two sentences how assigning an intervention on **the basis of who needs it most would NOT solve** the selection bias problem (an example of need-based assignment would be to assign health insurance to the people that have the worst health, or are poorer, first). [3pts, 2 sentences]

Solution: "All else equal" is violated as those who needs the intervention are different from those who do not need one. Those who receive the intervention are, by construction, more likely to have poor (potential) outcomes without the intervention (Y_{i0})

5. Consider the following table displaying average undergraduate GPA from a random sample of Penn State University students. [6pts, 2pts each]

	Mean	SD	N
GPA	3.10	.4	400

a) Construct the 95% confidence interval for the sample mean GPA at Penn State and interpret.

Solution:

$$SE = \frac{.4}{\sqrt{400}} = 0.02$$

$$CI = [3.10 - 2SE, 3.10 + 2SE] = [3.06, 3.14]$$

Full credit even if there is a computation error from $\frac{.4}{\sqrt{400}}$ to 0.02 (or if expressed as is) as long as the error is consistent with the definition of CI above.

b) Calculate the t-statistic and interpret significance for the null hypothesis that population average GPA is 4.1. Repeat for a population average GPA of 3.09.

Solution:

$$t_1 = \frac{\hat{\mu} - \mu_0}{SE(\hat{\mu})} = \frac{3.1 - 4.1}{0.02} = -50$$

statistically significant at 95% level, so that we reject the null hypothesis that the sample mean is different from the population mean.

$$t_2 = \frac{\hat{\mu} - \mu_0}{SE(\hat{\mu})} = \frac{3.1 - 3.09}{0.02} = -0.5$$

it is statistically insignificant at 95% level so that we fail to reject the null hypothesis that the sample mean is not different from the population mean.

c) Provide approximate p-values for the two t-statistics in (b)

Solution: When population mean is 4.1, t-statistic is -50, p-value is very close to 0(1pt); when population mean is 3.09, t-statistic is -0.5, p-value is greater than 0.3 (as a the probability of a t-statistic of $|t| > 1$ is about 30%) (1pt, or 0.5pt if it says "greater than 0.05").

d) Compare these two results.

Solution: When population mean is 4.1, it is very different from sample mean 3.1, as a result, a t-statistic of -50 and a p-value close to 0 means that there is close to 0% of the chance that the sample mean is the same as the population mean. On the other hand, when population mean is 3.09, very close to 3.1, sample mean is likely not different from population mean.

A short answer along the lines of "the sample mean is almost impossible to have come from (or being compatible with) a population of 4.1, and quite likely to have come from a population of 3.1" also gets full credit.

6. A t-statistic, that tests that an intervention had an effect of 1.5, is -3.2. The standard deviation for the outcome variable is 5, and the study had 100 observations. What is the estimated effect of the intervention (using SE for sample mean is fine)? [3pts]

Solution:

$$SE = \frac{SD}{\sqrt{n}} = \frac{5}{\sqrt{100}} = 0.5$$

$$t = \frac{\hat{\mu} - \mu_0}{SE(\hat{\mu})} \Rightarrow \hat{\mu} - 1.5 = -3.2 \times 0.5$$

$$\Rightarrow \hat{\mu} = -1.6 + 1.5 = -0.1$$

7. The 95% confidence interval of an estimated coefficient is [1.34, 5.74]. What is the corresponding standard error of this coefficient? (hint: approximately how many times standard error can fit in this 95% confidence interval?) [2pts]

Solution: A confidence interval ranges for 4 SE, so

$$(5.74 - 1.34)/4 = 1.1$$

8. The 95% confidence interval for a sample mean [-2, 4]: [2pts]

   a) Will contain the estimated sample mean 95% of the time

   b) Is less likely to contain the population value in the [-1,1] range than in the [1, 4] range

   c) Will contain the true mean 95% of the time

   d) Is equally likely to contain the population value in the [0,1] range than in the [0, 4] range

Solution: Here was I typo in C, making C an B correct

9. P-Hacking: [4pts, 2pts each]

   a) Define the problem of p-hacking.

Solution: "p-hacking" happens when flexibility in data analysis allows portrayal of almost anything as below an arbitrary p-value threshold. Any definition that is close in spirit the one above is fine.

b) In the Dale and Krueger example covered in class we have seen many regression (18 if you count tables 1-3!). Could you think of additional methodological choices that could generate another set of 18 (or more) regressions (there are many ways in you can answer this, for example you can use material from nonlinearities, regression inference an even from before, if it helps you can also assume that you have other variables that you can include in your regression)?

Solution: Some possible solutions:

- Define the outcome variable in levels instead of logs and repeat the previous 18 regressions.
- Define of the covariates that was in logs (levels) in levels (logs) and repeat all 18
- Remove or add some regresssors
- Repeat all 18 regression but adjusting for robust standard errors (assuming original regressions used old fashioned SEs)
- Repeat all 18 regression but adjusting for old fashioned standard errors (assuming original regressions used robust SEs)
- redefine the 150 selective groups into a smaller subset and repeat everything else.

10. Under which conditions you could convincingly argue that regression is a good research design tool to answer a causal question? [3pts, 1-3 sentences]

Solution: When we can argue that we can control for all potential observable characteristics.

(Alternatively, a slightly more refined answer could be:) When we have control for anything that might be correlated with the treatment and the outcome, hence preventing OVB.

11. We discuss the Table below while learning about RCTs, but now we can look with more detail at the notes: is says that this are regression coefficients. [6pts, 2pts each]

TABLE 1.6
OHP effects on health indicators and financial health

Outcome	Oregon: Control mean (1)	Oregon: Treatment effect (2)	Portland area: Control mean (3)	Portland area: Treatment effect (4)
A. Health indicators				
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	−.13 (.30)
B. Financial health				
Medical expenditures >30% of income			.055	−.011 (.005)
Any medical debt?			.568	−.032 (.010)
Sample size	23,741		12,229	

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on health indicators and financial health. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

(a) How many regressions are in panel A of this table?

Solution: Five (2pts). Seven (1pt)

(b) Write down the regression equation for the Physical Health index (assume that the regression uses only the treatment variable as a regressor).

Solution: Physical Health index $= 45.5 + 0.29 * \text{OHP treatment}_i + e_i$. Given that the question did not explicitly said 'regression estimate' the following solution is also fine:
$Y_i = \alpha + \beta T_i + e_i$ With Y_i Physical health index and T_i OHP Treatement

(c) Is it statistically significant? Why?

Solution: $0.29/0.21 < 2$, not significant at 5% level.

12. Recall this table from section 7 ("The gender gap in Canadian federal election...") [8pts, 2pts each]:

	(Model 1)	(Model 2)
Woman	-9.877^{***}	-0.450^{***}
	(0.366)	(0.135)
Vote share lag		0.276^{***}
		(0.006)
Party performance		0.666^{***}
		(0.004)
Incumbent Party		6.783^{***}
		(0.148)
Distance from contention		-0.015^{**}
		(0.005)
Constant	27.682^{***}	0.281
	(0.139)	(0.246)
Observations /N	23903	23903
R^2	0.030	0.872

Standard errors in parentheses
$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

a) What is the corresponding equation for Model 2?, What is the estimated equation for Model 2?

Solution: Regression equation:

$$\text{Outcome}_i = \alpha + \beta_1 \times \text{Women}_i + \beta_2 \times \text{Vote share lag}_i + \beta_3 \times \text{Party performance}_i +$$

$\beta_4 \times \text{Incumbent Party}_i + \beta_5 \times \text{Distance from contention}_i + e_i$

Full credit if the refer to the outcome in general terms (Y or outcome is fine)

Estimated equation:

$$\text{Outcome} = 0.281 - 0.450 \times \text{Women} + 0.276 \times \text{Vote share lag} + 0.666 \times$$

Party performance$+6.783 \times$Incumbent Party $- 0.015 \times$ Distance from contention $+ e_i$

b) Provide an interpretation of this coefficient using the idea of regression at matching.

Solution:

On average women have -0.45 units less units (of the outcome) relative to men within **when comparing within groups (cells) of similar covariates/regressors (voter share, party performance, party incumbency, and distance from contention)**. Given that the question does not specify a specific variable, any variable is fine as long as it has the interpretation in bold (parenthesis is optional)

c) Provide an interpretation in plain English for the coefficient for the variable Women in model 2

Solution: Women have, on average, -0.45 of (outcome) when comparing individuals with similar characteristics (of voter share, party performance, incumbency and distance from contention)

d) Compute the t-statistic, for standard the null of a zero coefficient, for the coefficient of the variable "Distance from contention" in model 2. Give a rough estimate of the corresponding p-value.

Solution:

$$t = \frac{-0.015 - 0}{0.005} = 3$$

(1pt) The p-value is smaller than 0.01 (1pt, with 0.5pt if they say smaller than 0.05).

13. The regression residuals [2pts]

    a) are aggregated in a sum to solve the regression minimization problem.

    b) are unknown since we do not know the population regression function.

    c) are aggregated in a squared sum to solve the regression minimization problem.

    d) should not be used in practice since they indicate that your regression does not run through all your observations.

Solution: C

14. Describe the OVB formula in one sentence (all English, no symbols) [3pts, 1 sentence]

Solution:

Short equals long plus relationship between omitted and treatment times relationship between outcome and omitted in long (fine if instead of "relationship it say effect").

15. In a video watched on lecture one, an interviewer brings up the point that the gender pay gap in the UK is 9% as evidence that modern society is still primarily dominated by men. Her point is that gender causes a wage differential.

    a) Write down the underlying regression that corresponds to interviewers' claim that women earn 9% less than men on average in the UK?[2pts]

Solution:

$$log(Wage) = \alpha - 0.09 * Gender + e_i$$

where gender = 1 means women, =0 means men.

-0.5 if outcome variable is not logged.

Another possible solution :

$$log(Wage) = \alpha + \beta * Gender + e$$

With and estimated $\beta = 0.09$.

Let's assume that another (new) commentator jumps into the conversation and says "That's probably an underestimate once you take into account that women stay away from industries that have a bad record in the treatment of women, and that those industries tend to have lower earnings on average"

    b) Write down the underlying regression that corresponds to the new commentator's argument (hint: think of an Industry Machismo Index (IMI), where 0 means no machismo in the industry and 100 means maximum machismo). [2pts]

Solution:

$$log(Wage) = \alpha + \beta_1 * Gender + \beta_2 * IMI + e$$

The new commentator suggests there is OVB (but in a different direction than the previous commentator). If we control for IMI the gap would be larger (β_1 more negative) than the estimated coefficient of -0.09 from part (a).

c) Discuss how the OVB formula could help us understand the effect of including a variable like "an industry machismo index" on the effect of gender on wages. Write down the long, short and auxiliary equations and interpret the OVB formula (the right hand side, not the $\beta_l - \beta_s$ part) [3pts, 2-5 lines]

Solution:
long: $log(Wage) = \alpha_l + \beta_l * Gender + \lambda * IMI + e_l$
short: $log(Wage) = \alpha_s + \beta_s * Gender + e_s$
auxiliary: $IMI = \pi_0 + \pi_1 * Gender + u$
$OVB = \beta_s - \beta_l = \lambda * \pi_1$
"Women stay away from industries that have a bad record in the treatment of women" means $\pi_1 < 0$. "Those industries (with high IMI) tend to have lower earnings on average" means $\lambda < 0$. As a result, $\beta_s - \beta_l > 0$, so $\beta_s > \beta_l$ and since they are negative, the true effect (β_l) is even more negative (a larger gap).

16. High variation in _____ increases the t-statistics regression coefficients (for 0 null hypothesis), while high variation in _____ reduces it.[3pts]

a) the sample size; the residuals

b) the residual; the sample size

c) the regressor; the residual

d) the residual; the regressor.

Solution: C

(Table 4 is used in questions 18 and 19 below)

Table 4: Regression results for a representative sample of 100,000 individuals ages between 40 and 50 years old in the US.

		Log(Salaries)			
	(1)	(2)	(3)	(4)	(5)
Years of education (S_i)	0.145 (0.051)	0.143 (0.05)	0.122 (0.064)	0.092 (0.09)	0.097 (0.065)
Entered labor marked in recession year (R_i)		-0.05 (0.02)	-0.052 (0.022)	-0.045 (0.018)	-0.051 (0.01)
Age (A_i)			0.02 (0.05)	0.012 (0.002)	0.014 (0.0015)
Distance to wealthiest zip code in county (D_i)				-0.005 (0.0001)	-0.0046 (0.12)
Rural = 1/ Urban =0					-0.05 (0.06)

Note: Distance to wealthiest zip code in county is a variable that measure the distance (in miles) of the zip code of birth to the wealthiest zip code in the county. So, if an individual was born far away from wealthy zip code, D will take a large vale (20-40 miles), on the other hand if the individual was born very close to the wealthiest zip code in their county D, would take small values (say 0-5). Enter labor market, takes the value of 1 if individual enter the labor market during a recession year, and zero otherwise. **Rural** is a binary variable that takes the value of 1 if the individual lives in a rural area and 0 if they live in an urban area.

17. Based on OVB and Regression as Matching. Answer the following questions related to Table 4: [12pts]

a) Interpret the coefficient for years of education in the **second** column using the idea of regression as matching. [3pts]

> **Solution:** One additional year of education is associated with a 14.3% increase in salary when comparing individuals that enter the labor force in similar recession years (or years of economics activity, or something similar).

b) Write down the regression equation for **column (3)**[2pts]

> **Solution:**
>
> $$log(Salaries) = \alpha + \beta_1 S_i + \beta_2 R_i + \beta_3 Age_i + e$$
>
> Alternatively:
>
> $$log(Salaries) = \alpha + 0.122 S_i - 0.052 R_i + 0.02 Age_i + e_i$$

c) Using the OVB formula, explain what is going on with the coefficient on years of education, when we move from 1 to 2, and from 3 to 4. Be explicit about the auxiliary equation that you would need in each case. (Note: by OVB formula, we mean the right-hand side, not "beta long – beta short) [3pts, 2-4 lines]

Solution: From 1 to 2:
long: $Log(Salaries) = \alpha_{l1} + \beta_{l1} \times S_i + \lambda_1 * R_i + e_{l1}$
short: $Log(Salaries) = \alpha_{s1} + \beta_{s1} \times S_i + e_{s1}$
auxiliary: $R_i = \pi_0 + \pi_1 \times S_i + u_1$
$OVB = \beta_{s1} - \beta_{l1} = \lambda_1 * \pi_1 = 0.145 - 0.143 = -.002$
The lack of substantial OVB is due to no relationship between entering in a recession and years of education (π_1 close to 0).

From 3 to 4:
long: $Log(Salaries) = \alpha_{l2} + \beta_{l2} \times S_i + \beta_{l3} \times R_i + \beta_{l4} \times A_i + \lambda_2 \times D_i + e_{l2}$
short: $Log(Salaries) = \alpha_{s2} + \beta_{s2} \times S_i + \beta_{s3} \times R_i + \beta_{s4} \times A_i + e_{s2}$
auxiliary: $D_i = \pi_2 + \pi_3 \times A_i + u_2$
$OVB = \beta_{s2} - \beta_{l2} = \lambda_2 * \pi_3 = 0.122 - 0.092 = 0.03 > 0$
The presence of substantial bias reflects implies that there must be a significant (negative) relationship between education and distance to wealthy areas ($\pi_3 < 0$).

d) In order to capture the true causal effect, it would be helpful to have some measure of "ability" (think something like general score of "skills and talents") and of "privilege" (think something like a general score for additional support received during upbringing and in labor market) both unobservable. Using the OVB formula (again, not the difference between beta long and beta short, but the other side of the equation), argue about the sign of the bias each of this two unobservable will generate. [4pts, 2-5 lines]

For simplicity lets use the columns 1 as the short equation (answer is fine if it uses any column as short):

Solution:

$$\begin{aligned}
\text{Short:} \quad & Y_i = \alpha^s + \beta^s S_i + e_i^s \\
\text{Long Ability:} \quad & Y_i = \alpha^s + \beta^s S_i + \gamma Ability_i + e_i^l \\
\text{Long Privilege:} \quad & Y_i = \alpha^s + \beta^s S_i + \lambda Privilege_i + e_i^l \\
\text{Aux Ability:} \quad & Ability_i = \pi_{a,0} + \pi_{a,1} S_i + e_i^a \\
\text{Aux Privilege:} \quad & Privilege_i = \pi_{p,0} + \pi_{p,1} S_i + e_i^p
\end{aligned}$$

(optional: Assuming higher ability, and higher privilege are capture by a higher score) The traditional story is that more ability correlates positively with both income and schooling. Similarly more privilege correlates positively with both income and schooling. Hence, OVB tells us that.

$$OVB_{Ability} = \pi_{a,1}\gamma > 0$$
$$OVB_{Privilege} = \pi_{p,1}\lambda > 0$$

18. Based on All Things Regression (Anatomy and others): Answer the following questions related to Table 4: [8pts]

a) Write down the two-regressions required to generate the coefficient of **distance to wealthiest zip code** in column 5 as the coefficient of a bi-variate regression equation. [2pts]

Solution:

$$Y_i = \alpha + \beta_4 \widetilde{D}_i + e^i$$
$$D_i = \pi_0 + \pi_1 S_i + \pi_2 R_i + \pi_3 A_i + \pi_4 Rural_i + \widetilde{D}_i$$

b) What is the t-statistic, for a null of zero association, for the coefficient on distance to wealthiest zip code on columns (4) and (5)? What happened with the coefficient themselves? Can you think of a possible explanation (hint: think of regression [2pts] anatomy)? [4pts]

Solution: t-statistics for columns 4 and 5 are -50 and about 0.04 (enough if the answer says much smaller than 2). The coefficient did not change substantially (2pts). This happens when the new variable included (Rural) is highly collinear with the one that dropped in significance (the formula for the SE using regression anatomy shows that lower variation in the residualized regressor makes the regressors increase drastically).

c) Assume that the person that gives you the regression outputs tells you "I forgot to mentioned that the variable for years of education was actually in logs", how should you re-interpret this coefficient? [2pts]

Solution: In this case the coefficient becomes an elasticity (or in what percent Y changes when S changes 1 percent). Hence the interpretation would be that salary increases in 0.13% when education increases in 1%.