Final Exam

Instructions: This exam is due by midnight on Sunday, June 8. There are 58 points.

You should submit your exam as a single PDF document on Canvas. My strong preference would be that you write your answers directly on the exam sheet (i.e., using a tablet or with pen and paper), then convert your exam sheet into a single PDF and upload it. If you are working with pen and paper, note that Apple or Android phones have features that allow you to scan documents as a PDF (this typically works much better than taking photos of your exam sheet and then converting those photos into PDF's).

If working directly on the exam sheet does not work for you, you can write your answers on a separate sheet, but in this case, please make sure it is very clear which exam question a given answer corresponds to. If you are handwriting, please make sure your answers are legible. Please also make sure you leave yourself some time before the deadline to do any PDF conversion/joining that you need to upload your exam as a single PDF.

As usual, you should abide by length restrictions on short answer questions. If a question asks for a one sentence answer, that means I will stop reading after one sentence when grading. The length restrictions are looser on this exam than on the midterms because I want to give you an opportunity to demonstrate your knowledge. But, as always, succinct and on-target answers will be graded more generously and I will deduct points for rambling, unfocused answers. Writing two incorrect sentences and then one correct sentences will, for example, typically result in no credit.

There are several short answer questions that ask you to explain things in the context of the question. This means I am looking for words-only explanations focusing on the specific variables, etc., being discussed (rather than generic statements like $cov(x, y) = 0$).

This is an open-book, open-everything exam. There are two rules. First, you are not allowed to work together with other students. Second, you must provide detailed documentation of any and all use of generative AI on the exam. The last page of the exam is reserved for you to disclose your gen-AI usage. Any generative AI use which is not acknowledged will be considered a violation of the honor code. I also reserve the right to deduct two points for over-reliance on AI on the exam at my discretion.

Question 1. Healthcare productivity (28 points)
Health economists are often interested in the returns to medical spending in terms of health outcomes. Suppose that you are a researcher interested in the causal effect of postnatal care provided to newborns on early-life health.

You have data on all Medicaid-covered births in California during 2015-2017. Your dataset includes the dummy variable $survive_i$, which equals one if newborn i survives to their first birthday and zero otherwise. Your dataset also includes the variable $spending_i$, which is the dollar value of all postnatal medical services provided to the newborn. This variable, $spending_i$, is your measure of the intensity of care. Finally, your dataset includes the infant's birthweight in grams, bw_i.

You should take note of two important institutional details. First, all the families in your dataset are covered by Medicaid, which is health insurance provided by the government for low-income families. Second, postnatal care decisions are made entirely by the supervising physician. Families themselves have no influence over medical care decisions.

In this context, the central challenge to estimating the causal effect of medical services is that the infant's underlying health at birth, denoted by h_i, is observed by the physician but is not captured in your dataset. Physicians provide more medical services to less healthy newborns (lower h_i), but these newborns are also less likely to survive. The infant's birthweight, bw_i, is a proxy measure for the infant's underlying health h_i; that is, $cov(bw_i, h_i) > 0$. On average, heavier newborns have higher survival rates, while lighter newborns tend to receive more postnatal medical care.

1. You start your analysis by estimating the bivariate regression $survive_i = \beta_0 + \beta_1 spending_i + u_i$. In what direction do you think your estimated $\hat{\beta}_1$ is biased? (2 pts)

   (a) $\hat{\beta}_1$ is biased downward.

   (b) $\hat{\beta}_1$ is not biased.

   (c) $\hat{\beta}_1$ is biased upward.

   (d) Insufficient information.

2. Briefly explain your answer above (1-2 sentences; words only). (2 pts)

3. Now suppose you add a control for birthweight to the above regression. Specifically, you estimate the regression $survive_i = \beta_0 + \beta_1 spending_i + \beta_2 bw_i + u_i$. How do you expect your estimated $\hat{\beta}_1$ to change from the above regression? (2 points)

   (a) $\hat{\beta}_1$ will decrease.

   (b) $\hat{\beta}_1$ will not change.

   (c) $\hat{\beta}_1$ will increase.

   (d) Insufficient information.

4. Briefly explain your answer above (1-2 sentences; words only). (2 pts)

5. Which of the following statements must be true for your estimated $\hat{\beta}_1$ from the regression $survive_i = \beta_0 + \beta_1 spending_i + \beta_2 bw_i + u_i$ to be an unbiased estimate of the true causal effect of spending on survival? (2 pts)

   (a) $cov(h_i, spending_i) = 0$

   (b) $cov(h_i, spending_i | bw_i) = 0$

   (c) $var(h_i | bw_i) = 0$

   (d) $corr(h_i, bw_i) = 1$

6. Briefly explain your answer above. First, explain precisely what the answer you chose means in words. Then, explain why you chose that answer. (2-3 sentences; words only) (2 pts)

7. Infants with birthweights below 1500 grams are thought to be at especially high risk of mortality. These newborns receive an official designation of Very Low Birthweight (VLBW). The American Medical Association recommends that significant additional postnatal care be provided to VLBW newborns.

   Given this, you create a new dummy variable, $VLBW_i$, that equals one if $bw_i < 1500$ and equals zero if $bw_i \geq 1500$. You estimate the regression $spending_i = \pi_0 + \pi_1 VLBW_i + \eta_i$. What is the expected sign of $\hat{\pi}_1$? (2 pts)

   (a) $\hat{\pi}_1 < 0$

   (b) $\hat{\pi}_1 > 0$

8. A colleague sees the results of the above regression. Noticing that your estimated $\hat{\pi}_1$ is highly statistically significant ($|t| > 10$), she suggests using the $VLBW$ dummy as an instrumental variable for spending. Specifically, she recommends estimating the following two regressions:

$$spending_i = \pi_0 + \pi_1 VLBW_i + \eta_i$$

$$survive_i = \gamma_0 + \gamma_1 VLBW_i + \nu_i$$

   And then computing $\hat{\beta}_{IV} = \hat{\gamma}_1 / \hat{\pi}_1$ as your estimate of the causal effect of health spending on survival.

   In what direction do you expect this estimated $\hat{\beta}_{IV}$ is biased as an estimate of the true causal effect of medical spending on survival? (2 pts)

   (a) $\hat{\beta}_{IV}$ is biased upwards

   (b) $\hat{\beta}_{IV}$ is biased downwards

9. Explain the rationale for your previous answer. (2-4 sentences; words only) (2 pts)

10. But your colleague was on the right track. What should you do instead? Explain clearly what regression(s) you would run and how you would compute an estimate of the causal effect of medical spending on survival. (2-4 sentences; some notation allowed) (3 pts)

11. What assumption(s) are required for the strategy you outlined in your previous answer to deliver a valid estimate of the causal effect of medical spending on survival? Provide some intuition for why these assumption(s) allow your strategy to obtain an estimate of the causal effect of interest (2-4 sentences; words only) (3 pts)

12. Are there any tests you could do to assess the plausibility of the assumption(s) you answered above? If so, explain clearly (i) what test(s) you would suggest; (ii) what outcome(s) of the test(s) would increase or decrease your confidence in the assumptions(s); and (iii) why those outcome(s) would increase your confidence in the assumption(s) (2-4 sentences; some notation allowed). (3 pts)

13. Revisiting part 8 above, there is a very specific condition under which your colleague's strategy will work. What is it? Provide an intuitive explanation clearly in words. (1-2 sentences) (1 pt).

    Hint: you may find a graph helpful.

Question 2. Saving incentives (30 points)
Economists and policymakers have long been troubled by the lack of retirement savings among lower-income households in the United States. Recent surveys have found that half of all Americans accumulate no savings each year and that the typical, low-income household has only about $2,000 in savings. Researchers have recently begun to explore whether policy interventions can effectively increase retirement savings rates.

One research team conducted the following experiment. The team stationed a researcher in each H&R Block location in the city of St. Louis, MO during the tax season in 2006 (H&R Block is a tax preparation company that helps individuals file their taxes; in St. Louis, there are about 40 locations spread across different neighborhoods throughout the city). When a customer entered an H&R Block location and began the tax preparation process, the researcher stationed at that location used a random number generator to randomly classify the customer as *treatment* or *control.*

If the customer was categorized as *control*, the researcher provided the customer with a pamphlet explaining the benefits of saving for retirement with a dedicated retirement savings account (often called an IRA). If the customer was categorized as *treatment*, the researcher provided the customer an identical pamphlet except that the pamphlet also included the following offer: if the customer opened a new retirement account in the next week and made a deposit into the new account, the research team would make a 50 percent matching contribution to the customer's new account. In other words, if the customer opened a new account and deposited 100 dollars, the research team would contribute an additional 50 dollars to the customer's new account.

The research team has hired you to perform an econometric analysis of the experiment. You have been provided with all the data from the experiment, which includes information on all customers entering H&R Block locations during the 2006 tax season as well as information on these customers' savings accounts measured first at the time they entered the H&R Block location and then again one week later.

1. After collecting the data from the experiment, the research team first assessed the randomization by regressing a treatment group dummy variable on customer characteristics. Results from these regressions are shown below in table 1. The F-statistic and associated p-value for a test a of the joint significance of the customer characteristics are reported at the bottom of the table. In column 2, the regression also controls for fixed effects for which H&R Block location the customer entered (Location FE).

Table 1: Balance test

	(1)	(2)
	Treatment	Treatment
Annual Income	542^{***}	111
	(96)	(83)
Female	0.011	0.008
	(0.012)	(0.009)
Age	0.22	0.18
	(0.19)	(0.16)
Married	0.022^{**}	0.016
	(0.098)	(0.087)
Homeowner	0.025^{*}	0.017
	(0.012)	(0.011)
Retirement Savings	912^{***}	84
	(111)	(96)
Any Retirement Account	0.001	0.001
	(0.011)	(0.009)
Location FE	No	Yes
F-stat	17.22	1.03
p-value	0.009	0.375
Observations	14,011	14,011

Based on this table, select all of the following that are true. (2 pts)

(a) The socioeconomic status of customers visiting H&R Block varies across locations.

(b) Customers visiting H&R Block locations that attract customers of higher socioeconomic status were more likely to be randomized into the treatment group.

(c) Customers visiting H&R Block locations that attract customers of higher socioeconomic status were less likely to be randomized into the treatment group.

(d) Within locations, the treatment appears randomly assigned.

2. Table 2 below shows the main result of the experiment. This table uses data on customer's savings accounts measured one week after they entered H&R Block. Column 1 shows the control group mean. Column 2 shows the estimated $\hat{\beta}_1$ (and associated standard error) from the regression:

$$Any_{ih} = \beta_0 + \beta_1 Treatment_{ih} + a_h + u_{ih}$$

Here, Any_{ih} is a dummy variable that equals one if customer i, who visited H&R Block location h, has a dedicated retirement savings account. $Treatment_{ih}$ is a dummy variable that equals one if the customer is in the treatment group. The regression also includes H&R Block location fixed effects (a_h).

Table 2: Regression estimates

	(1) Control Mean	(2) T−C
Any Retirement Account	0.0311	0.136*** (0.009)
Location FE	-	Yes
Controls	-	No
Observations	-	14,011

Which of the following is the correct interpretation of the regression coefficient $\hat{\beta}_1$ in table 2 above? (2 pts)

(a) A pamphlet about the benefits of retirement accounts increases the probability that an individual has a retirement account by 13 percent.

(b) An offer of a 50 percent matching contribution increases the probability that an individual has a retirement account by 13 percent.

(c) A pamphlet about the benefits of retirement accounts increases the probability that an individual has a retirement account by 13 percentage points.

(d) An offer of a 50 percent matching contribution increases the probability that an individual has a retirement account by 13 percentage points.

3. If you added all the controls from Table 1 to the regression in column 2 of table 2, how would you expect the estimated $\hat{\beta}_1$ to change? (2 pts)

(a) $\hat{\beta}_1$ would decrease.

(b) $\hat{\beta}_1$ would not change.

(c) $\hat{\beta}_1$ would increase.

(d) Insufficient information.

4. If you added all the controls from Table 1 to the regression in column 2 of table 2, how would you expect the standard error of the estimated $\hat{\beta}_1$ to change? (2 pts)

   (a) $se(\hat{\beta}_1)$ would decrease.
   (b) $se(\hat{\beta}_1)$ would not change.
   (c) $se(\hat{\beta}_1)$ would increase.
   (d) Insufficient information.

5. It turns out that another researcher, who is interested in the effect of retirement savings on mental health, conducted a follow-up survey of the study population in 2016 (ten years after the initial experiment). The researcher tracked down all members of the original study population and interviewed them, asking them a series of questions about (a) their current retirement savings and (b) their mental health. The researcher asks for you for your help with the data analysis. She starts by showing you table 3:

Table 3: Regression results from follow-up survey

	(1) Depression	(2) Depression	(3) Any Retirement Account
Any Retirement Account	-0.22^{***} (0.032)		
Treatment		-0.04^{*} (0.024)	0.24^{***} (0.053)
Control Mean	0.31	0.31	0.14
Location FE	Yes	Yes	Yes
Controls	No	No	No
Observations	14,011	14,011	14,011

Each column of the table is from a separate regression, where the left-hand side variable is labeled at the top of each column and the right-hand side variables are listed vertically on the left. If there is no regression coefficient reported, that right-hand side variable was not included in the regression. The variable $Depression_i$ is a dummy variable that equals one if the individual has experienced depression in the past thirty days. *Any Retirement Account*$_i$ is a dummy variable that equals one if the individual has a dedicated retirement account. Both variables were measured in 2016 as part of the follow-up survey. All regressions include H&R Block location fixed effects, as noted in the table footer.

Which of the following is the correct interpretation of the coefficient in column 1 of the above table? (2 pts)

(a) Individuals with retirements accounts are 22 percent less likely to have recently experienced depression.

(b) Individuals with retirement accounts are 22 percentage points less likely to have recently experienced depression.

(c) Individuals who received an offer of a matching retirement account contribution are 22 percent less likely to have recently experience depression.

(d) Individuals who received an offer of a matching retirement account contribution are 22 percentage points less likely to have recently experience depression.

6. Compute the instrumental variables estimate of the effect of having a dedicated retirement account on depression, using an offer of a 50 percent matching contribution ten years ago as an instrument for having a dedicated retirement account and controlling for H&R Block location fixed effects. (2 pts)

$\hat{\beta}_{IV} =$

7. As a well-trained econometrician, you know that the validity of the instrumental variables estimate you computed in part 6 depends on an assumption sometimes known as the relevance assumption. In the specific context of this question, explain the relevance assumption. Then assert whether or not you think the assumption is satisfied and explain your reasoning. (2-3 sentences; words only) (2 pts)

8. As a well-trained econometrician, you know that the validity of the instrumental variables estimate you computed in part 6 depends on an assumption sometimes known as the exogeneity assumption. In the specific context of this question, explain the exogeneity assumption. Then assert whether or not you think the assumption is satisfied and explain your reasoning. (2-3 sentences; words only) (2 pts)

9. As a well-trained econometrician, you know that the validity of the instrumental variables estimate you computed in part 6 depends on an assumption sometimes known as the exclusion restriction. In the specific context of this question, explain the exclusion restriction. Then assert whether or not you think the assumption is satisfied and explain your reasoning. (2-3 sentences; words only) (2 pts)

10. As a well-trained econometrician, you know that the validity of the instrumental variables estimate you computed in part 6 depends on an assumption sometimes known as the monotonicity assumption. In the specific context of this question, explain the monotonicity assumption. Then assert whether or not you think the assumption is satisfied and explain your reasoning. (2-3 sentences; words only) (2 pts)

11. Another member of the research team thinks the monotonicity assumption you discussed in part 10 is not satisfied. Specifically, they note that there are many individuals in the control group who have retirement accounts and many members of the treatment group who do not have retirement accounts and are arguing that this violates the monotonicity assumption. How would you respond to this team member? (1-3 sentences; words only) (2 pts)

12. As a well-trained econometrician, you know that you should interpret your IV estimate from part 6 as a local average treatment effect (LATE) as long as the assumptions discussed above are satisfied. In the specific context of this question, explain what this local average treatment effect is. Use words only and be as precise as possible. (1-3 sentences; words only) (2 pts)

13. A different member of the research team prefers the OLS estimate reported in table 3, column 1, to your IV estimate from part 6. Specifically, they note that the OLS estimate has a smaller standard error than your IV estimate. How would you respond to this team member? (1-3 sentences; words only) (2 pts)

14. Another of your fellow researchers argues that the LATE captured by your IV estimate (i.e., you answer to part 12) is not a very useful estimate. How would you respond to this team member? (2-3 sentences; words only) (2 pts)

15. Based on the IV estimate you computed in part 6, make a judgement about the direction and magnitude of the degree of omitted variables bias in the OLS estimate report in table 3, column 1. Clearly justify your answer, explaining what assumptions you need to make and why you think these are reasonable assumptions. (2-4 sentences; words only) (2 pts)

Gen-AI disclosure
Please list every question on the exam for which you used generative AI and provide a clear description of how it was used in each case.