# ECON 20: Final Exam

Professor Mello
Dartmouth College

June 4, 2021

## Instructions

This exam is due on Monday, June 7 at 11pm. This exam is open-note, open-book, open-whatever, but you are not allowed to work together. There are 60 total points.

Upload your answers to all questions as a single PDF file. Make it clear which answer corresponds to which question on the exam. Please do not write your answers directly on this exam sheet. Your are free to hand-write or type your answers, but I would recommend typing your answers to ensure legibility.

Abide by length recommendations on short answers. If a question asks for a one sentence answer, that means I will stop reading after one sentence when grading. The length recommendations on this exam are looser than on the midterms because I want to give you opportunities to display your knowledge. However, I will deduct points for rambling, unfocused answers. Writing two incorrect sentences and then one correct sentence, for example, will receive no or almost no credit.

# 1 Health Care Productivity (30 Points)

Health economics 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 the intensity 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. The variable $spending_i$ is your measure of the intensity of care. Your dataset includes the infant's birthweight in grams, $bw_i$, as well.

You should take note of two things. First, postnatal care decisions are made entirely by the supervising physician; parents and families have no influence over care decisions. Second, all families in your sample are covered by Medicaid, which is publically-provided health insurance for lower-income families. That is, all families in your sample have generous health insurance through the government.

The central challenge in 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 less likely to survive. The infant's birthweight, $bw_i$, is a proxy measure for the infant's underlying health, $h_i$. Birthweight is positively correlated with the infant's unobserved 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 points)

(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 points)

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 points)

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 points)

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

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

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

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

6. Briefly explain your answer above. First, explain what the answer you selected means in words. Then, explain why you chose that answer, again in words (2 sentences; words only). (2 points)

7. Infants with birthweights below 1500 grams are thought to be at especially high risk of mortality. This group of newborns receives 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 sign of $\hat{\pi}_1$? (2 points)

   (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:

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

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

   And then computing $\hat{\beta}_{IV} = \hat{\gamma}_1 / \hat{\pi}_1$ as your estimate of the causal effect of health spending on survival. What assumption(s) are needed for $\hat{\beta}_{IV}$ to be an unbiased estimator for the causal effect of health spending on survival? (1-2 sentences; words only) (2 points)

9. First, explain in words why the strategy above is unlikely to deliver an unbiased estimate of the causal effect. Then, explain in what direction the estimated $\hat{\beta}_{IV}$ is likely to be biased. Please explain how you arrived at your conclusion about the direction of bias (2-4 sentences; words only; be specific). (4 points)
   *Hints: (i) Use the definition of $VLBW_i$; (ii) you may find it useful to draw a graph.*

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

11. What assumption(s) are needed for your answer above to provide a valid estimate of the causal effect of medical spending on survival? Provide some intuition for why these assumption(s) allow your strategy to ontain an estimate of the causal effect (2-4 sentences; words only; be specific). (3 points).

12. Are there any tests you could do to assess the plausibiilty of the assumption(s) you answered above? Explain clearly (i) what test(s) you would recommend; (ii) what outcome(s) of the test(s) would increase your confidence in the assumption(s); and (iii) why those outcome(s) would increase or decrease your confidence in the assumption(s) (2-4 sentences; some notation allowed). (3 points)

13. Returning to part 8, 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 points)
*Hint: explain how a graph would look.*

## 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; there are about 40 locations spread across many different neighborhoods in St. Louis). 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 with 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 account.

The research team has hired you to consult on their econometric analysis of the experiment. Your 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 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 Check

| | (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 points)

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

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

(c) Customers visting 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 Results

| | (1) | (2) |
|---|---|---|
| | Control Mean | 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 interpretations of the regression coefficient $\hat{\beta}_1$ in Table 2 above? (2 points)

(a) A pamplet 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 pamplet 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 points)

   (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 points)

   (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. One member of the research team has argued that, when estimating the regression shown in Table 2, the standard errors should be adjusted for clustering within H&R Block locations. However, the team member cannot seem to justify their argument. What would be a correct rationale for using their suggested cluster correction? (1-2 sentences; words only; be specific) (2 points)

6. 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. He starts by showing you Table 3:

Table 3: Regression Results from Follow-Up Survey

| | (1)<br>Depression | (2)<br>Depression | (3)<br>Any Retirement Account |
|---|---|---|---|
| Any Retirement Account | $-0.22^{***}$ | | |
| | (0.032) | | |
| Treatment | | $-0.04^{*}$ | $0.24^{***}$ |
| | | (0.024) | (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 dependant 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.

Interpret the coefficient in column 1 (2 points).

7. The researcher wants to use the coefficient in column 1 as his estimate of the causal effect of having a retirement account on depression. You argue that he should be using an instrumental variables estimate instead. Compute that instrumental variables estimate (2 points).

8. What can you learn by comparing the $-0.22$ in Column 1 with the IV estimate you just computed? (1-3 sentences; words only; be specific) (2 points)

9. The researcher just examined Table 1 above and is now concerned that your IV estimator is biased because whether an individual is in the treatment or control group appears correlated with their socioeconomic status. How do you respond? (1-3 sentences; words only; be specific) (2 points)

10. The researcher is also skeptical of your IV estimate because its validity depends on an untestable assumption, namely the exclusion restriction. How do you respond? (1-3 sentences; words only; be specific) (2 points)

11. The researcher is also skeptical of your IV estimate because it has a larger standard error than his preferred estimate in Column 1 of Table 3. How do you respond? (1-3 sentences; words only; be specific) (2 points)

12. As a well-trained econometrician, you know that you should interpret your IV estimate as a local average treatment effect (LATE) as long as the *no defiers* (or *monotonicity*) assumption holds. Focusing on the specific application being explored in this question, explain the *no defiers* assumption in words. (1-3 sentences; words only; be specific). (2 points)

13. Your fellow researcher is concerned that the *no defiers* assumption is not satisfied. In particular, he shows you that a significant share of customers in the treatment group do not have retirement accounts. How do you respond? Be as specific as possible (1-3 sentences; words only). (2 points)

14. There are also many customers in the control group who do have retirement accounts. Select all of the following that are true. (2 points)

    (a) These customers might be never-takers.

    (b) These customers might be compliers.

    (c) These customers might be always-takers.

    (d) These customers might be defiers.

15. Assuming the *no defiers* assumption holds, you should interpret your IV estimate as the causal effect for the subgroup of compliers. Who are the compliers in this case? Be as specific as possible (1-2 sentences; words only). (2 points).

16. BONUS. In this particular case, make an argument for why the LATE *is* a useful parameter (1-3 sentences; words only). (1 point)

17. BONUS. In this particular case, make an argument for why the LATE *is not* a useful parameter (1-3 sentences; words only). (1 point)