UGBA 96 Data and Decisions, Spring 2018

Conrad Miller
The University of California at Berkeley
Haas School of Business

Final Exam Solutions

Name: _____

Section: 4

On my honor, I swear that:

- *I did not/will not discuss the exam with anyone prior, during or after the exam;*
- *I did all of the work on the exam on my own;*
- *I did not take more than the allotted time to finish the exam;*
- *I followed the instructions as stated above.*

_____ _____

Signature *Date*

TIME ALLOWED: 2 Hours

TOTAL POINTS: There are 3 problems worth 100 total points.

MATERIALS ALLOWED: The exam is closed book, closed note, and no laptops or tablets are allowed. You are allowed to use a single 8½ x 11 sheet of paper (double-sided) with ***handwritten*** *notes.*

NOTE: I have indicated how many points each part is worth. Allocate your time accordingly. Keep your answers short and to the point.

1. Testing Course Requirements for Haas Students (45 points)

This question has 12 parts (A) through (L).

The Haas Undergraduate Program requires a statistics course as a prerequisite to admission. Currently, there are multiple courses on campus that satisfy this prerequisite (e.g., Stats 20, 21, 131A). The program is considering whether to **require** that all students take **Data 8**, Foundations of Data Science, to satisfy the statistics requirement. To determine whether to institute this requirement, Haas administrators want to know the causal effect of requiring Data 8 on student performance in Haas core courses. In particular, they want to know what causal effect completing Data 8 *would have had* on Haas students that *did not* take Data 8 but took some other statistics course instead. We will assume throughout that enforcing a Data 8 requirement will have no effect on who is ultimately admitted to the Haas Undergraduate Program, and neither will any other treatment considered below.

To summarize: the unit is a student. The treatment of interest is completing Data 8. The control condition is taking another statistics course. The outcome is performance in core Haas classes.

Note: we will assume that we have some absolute measure of performance, like scores on exams, rather than some relative or 'curved' measure of performance, like grades. In other words, for the metric we are considering, the performance of one student does not depend directly on the performance of their classmates.

A) [3 points] In the next few problems, we will apply the potential outcomes framework to our research question. Explain the meaning of Y_{i1} in this context.

Yi1 is the potential outcome (Haas core performance) for student i if he or she completes Data 8.

B) [3 points] Explain the meaning of $E[Y_{i1}]$.

E[Yi1] is the population average of the potential outcome (Haas core performance) for each student if he or she completes Data 8.

C) [3 points] The university is interested in measuring the average causal effect of completing Data 8 among students that did not take Data 8 but took some other statistics course instead. Write down this causal effect using the potential outcomes notation.

E[Yi1 – Yi0 | Di = 0]

Give partial credit if they identify as treatment on the untreated, but do not correctly use potential outcomes notation.

D) [4 points] An administrator argues that one important advantage of requiring all students to take Data 8 is that it will allow instructors of Haas core courses to tailor the instruction to the students' skills (e.g. including Python labs). This improved course design may improve student performance. However, instructors can only tailor the course in this way when a sufficient number of students in the course have taken Data 8.

What implication does this side benefit have for using the potential outcomes framework in this setting?

This introduces interference. Now potential outcomes depend on whether a sufficient number of classmates have also taken Data 8, rather than just an individual student's treatment.

Ignore this issue for the remainder of the problem.

E) [3 points] In order to measure the causal effect of interest, an administrator proposes comparing outcomes for Haas students that completed Data 8 and Haas students that did not complete Data 8.

Write down this comparison using the potential outcomes notation.

E[Y_i1 | D_i = 1] – E[Y_i0 | D_i = 0]

F) [3 points] Will this comparison identify the causal effect of interest? Why or why not?

In general, no, this comparison will not identify the causal effect of interest. The concern is *selection bias*: students that have completed Data 8 differ from students that have not for reasons unrelated to Data 8 itself. For example, students that complete Data 8 may have more baseline exposure to programming, and so may be more successful in Haas core courses that involve programming, on average.

The Haas administration decides to run a randomized experiment to help answer their research question. In this experiment, administrators take a list of students that expressed interest in applying to Haas, and randomly select half of those students for the treatment group. For each randomly selected student, Haas sends personalized materials promoting Data 8 as a great course for completing the Haas statistics prerequisite. The objective is to increase enrollment in Data 8 among treated students.

Below is a table with the results of the experiment with means calculated separately for the treatment and control groups. Performance is measured in standardized units--a value of 1 indicates a 1 standard deviation higher performance than the average control student.

Experiment 1 results:

	Treatment	**Control**
Completed Data 8	0.5	0.4
Haas Core Performance	0.05	0

G) [4 points] An administrator remarks that the performance of the treatment and control students is very similar, with only a difference of 0.05 standardized units. He takes this as evidence that completing Data 8 has a negligible causal effect on student performance. Do you agree? If so, explain how the evidence supports this view. If not, explain to the administrator in layman terms why the evidence does not support this view.

No, this conclusion is incorrect. The difference in performance between the treatment and control groups is small, but the difference in Data 8 completion rates is also relatively small! If we want to measure the causal effect of completing Data 8 on performance, we need to account for this 'non-compliance'—the fact that may *treated* students *do not* complete Data 8 and many *untreated* students *do* complete Data 8.

H) [4 points] Assume no defiers exist. Approximately what share of students are compliers in this experiment? Approximately what share of students are 'never-takers'?

As in problem set 2, we can estimate the share of students that are compliers using the first stage. This is the difference in Data 8 completion rates between the two groups: 0.5 – 0.4 = *0.1*

We can estimate the share of students that are never takers by looking at the share of treated students that do no complete Data 8. With no defiers, the share of never-takers is approximately the share of treated students that do not complete Data 8: *0.5*.

I) [4 points] Based on the table above, construct a LATE estimate for the causal effect of completing Data 8 on student performance.

LATE = (Reduced Form/First Stage)
= (0.05 – 0.0)/(0.5 – 0.4) = 0.5 standardized unites of performance

J) [4 points] Suppose that in addition to increasing Data 8 enrollment, the promotional materials also cause treated students to take a statistics course earlier in their academic career. Some instructors suspect that students that have taken statistics more recently perform better in core courses. What implications does this have for your answer to part (I)?

This would violate the exclusion restriction required for the LATE formula to be valid. The exclusion restriction requires that our instrument (receiving the promotional materials), only affects performance by changing the probability of completing Data 8. The channel described here—causing students to take a statistics course earlier in their career and so potentially leading to worse performance in Haas core courses—is a distinct one, and so would be a violation.

For the rest of this problem, assume the promotional materials do not cause treated students to take a statistics class earlier in their academic career.

Haas administrators decide to run a second randomized experiment. Again, administrators take a list of students that expressed interest in applying to Haas, and randomly select half of those students. In this experiment, those randomly selected students are **required** to take Data 8. As before, control students can take Data 8 if they like, but are not required to do so.

The results of the experiment are presented in the table below with means calculated separately for the treatment and control groups.

Experiment 2 results:

	Treatment	**Control**
Completed Data 8	1.0	0.4
Performance	0.10	0

K) [6 points] How do compliers differ in the first and second experiment? Compare both their behavior and treatment effects. Which experiment is more relevant for the original causal question of interest: what is the causal effect of completing Data 8 among Haas students that, in the absence of a Data 8 requirement, would take some other statistics course instead?

In the first experiment, compliers are students that will complete Data 8 if they receive the promotional materials but will not complete Data 8 if they do not. In the second experiment, compliers are students that will complete Data 8 only if they are required to do so. The LATE for the second experiment is 0.10/(0.6) = 1/6, which is substantially smaller than the LATE for the first experiment. Hence, the treatment effect is larger for compliers in the first experiment than compliers in the second.

The second experiment is more relevant, as compliers in this experiment map exactly to the treatment effect of interest that is described in the problem.

L) [4 points] The same administrator suspects that if taking Data 8 improves students' performance in core Haas classes, it will also help Haas students **persist** in the business major--that is, it will reduce the number of students that change majors to something other than business. Suppose that for students that change majors we **do not have data** on Haas core performance.

What implications would this pattern have for the above analysis?

This pattern of behavior will likely introduction *attrition bias*. While randomization ensures balance between the treatment and control groups at baseline, with attrition we are no longer assured balance at endline. The pattern described in the question suggests that we will see more attrition in the control group than the treatment group. Moreover, this attrition is selective—students that perform better in core classes are less likely to leave the data. This means we are likely to understate the causal effect of completing Data 8, because low-performing students (who will disproportionately come from the control group) will be less likely to remain in the data.

This is similar to the attrition bias we saw in the Work from Home lab.

2. Online Advertising Campaigns at the Gap (30 points)

This question has 8 parts (A) through (H).

The clothing retailer Gap has hired you as a data scientist to help them determine how much to spend on their online advertising campaigns. Gap pays for online advertising on various websites, including Google, Facebook, and Twitter. In order to decide how much to spend on these campaigns, Gap must estimate the causal effect of an online campaign on sales. The higher the return, the more Gap will be willing to spend on placing advertisements.

In this question, our unit of observation will be the customer, an individual browsing online that will potentially see a Gap advertisement and potentially make a purchase. In this question, we will estimate the causal effect of seeing a Gap advertisement on spending on Gap merchandise.

For each customer, we have historical data on whether the customer has recently been shown an advertisement on any of a variety of websites (`shown_ad`, an indicator), and that customer's recent spending on Gap merchandise (`gap_spending`). Historically, customers are more likely to be shown Gap advertisements if their browsing history suggests an interest in Gap or related products, or if they fall into Gap's targeted demographic groups (18- to 34-year-olds).

A) [3 points] Explain the meaning of the following potential outcomes notation in this context: $E[Y_{i1} | D_i = 1]$.

This is the potential outcome (spending on Gap products) if customers are shown an advertisement, average across all customer that are shown an adverstiement.

As a first step towards estimating the effect of advertising, Gap has estimated the following regression model using their customer data:

$$\text{gap\_spending} = \alpha + \beta\ \text{shown\_ad} + \varepsilon$$

where `gap_spending` is measured in dollars and `shown_ad` is an indicator for whether the customer has been shown a Gap advertisement.

The estimated β coefficient is **0.50**.

B) [4 points] Interpret the β coefficient estimate in a sentence. Does this reflect the causal effect of advertising on customer spending? Why or why not?

Being shown an advertisement is associated with a $0.50 increase in customer spending on Gap products. That is, customers that are shown an advertisement spend $0.50 more on Gap products than customers that are not shown an advertisement, on average.

C) [4 points] Suppose we also had data on whether a customer had shopped for Gap in the past. The variable `prior_experience` measures the amount of time the customer has previously spent browsing the Gap website. Suppose we added this covariate as a control to the model above. How would you expect the β coefficient to change: would it become more positive or more negative? Why?

We can use the *omitted variable bias* formula to answer this. How the β coefficient to changes depends on two relationships. First, it depends on the sign relationship between the omitted variable (prior_experience) and the included variable (shown_ad). The description of how advertisements were targeted historically tells us this relationship is *positive*. Second, it depends on the sign of coefficient on prior_experience in the 'long regression', a regression where both prior_experience and shown_ad are included as covariates. We don't know this sign from the problem, but it's reasonable to assume that coefficient is positive—controlling for advertising, customers that have spent more time shopping on the Gap website spend more on Gap merchandise.

Hence, the omitted variable bias is positive, so we expect the β coefficient to be smaller/more negative in the long regression.

Suppose instead that we use `age` as an additional control and estimate the following regression model:

$$\text{gap\_spending} = \alpha + \beta\ \text{shown\_ad} + \gamma\ \text{age} + \varepsilon$$

where `age` is measured in years.

In this expanded regression model, the estimated β coefficient is **0.75** and the estimated γ coefficient is **0.01**.

D) [3 points] Interpret the estimated coefficient γ in a sentence.

A one year increase in age is associated with a $0.01 increase in spending on Gap merchandise.

E) [4 points] What does the change in β estimates between the two regression models tell you about the relationship between `age` and `shown_ad`?

The β coefficient increases from 0.5 to 0.75. Based on the omitted variable bias formula, this implies that the relationship between shown_ad and age is *negative*.

You have convinced Gap to run a randomized control trial to estimate the causal effect of advertising on customer spending. For a set of customers that could potentially be shown advertisements, Gap will pay to show advertisements to a random subset of those customers. However, because Gap believes that the returns to advertising are higher for their targeted demographic, 18- to 34-year-olds, they have stipulated that the randomization in the experiment be stratified. In particular, while 20% of customers in other demographic groups will be assigned to see the ad and 80% will be assigned to the control (and so will see some non-Gap advertisement instead), 80% of target demographic customers will see the advertisement and 20% will be assigned to the control. (Just to make sure you read that right: 18- to 34-year-olds are more likely to be assigned to see the advertisement than older customers.)

F) [3 points] What is the benefit of running a randomized experiment in this context?

A randomized experiment allows us to credibly estimate the causal effect of advertising on purchase behavior, the treatment effect of interest. In the absence of an experiment, our estimates are plagued by selection bias.

Below is a table with the baseline results of the experiment with means calculated separately for the treatment and control groups.

	Treatment	**Control**
Age	24	36
Shown Advertisement	1	0
Gap Spending ($)	0.25	0.36

G) [4 points] The in-house analyst is shocked by the experiment results. She says the difference in sales between the treated and control groups is -0.11, which implies that advertising actually **decreases** sales.

Do you agree with this interpretation? Why or why not?

No, this interpretation is incorrect. We cannot simply compare treatment and control here, because the two groups are unbalanced. We can see this from the differences in average age. The two groups are unbalanced despite randomization because we made treatment more likely for 18- 34-year-olds than other customers. Hence, the treatment group may have lower spending because that group is generally younger, not because they were shown an advertisement.

H) [5 points] Describe a strategy for recovering the causal effects of advertising using data from the experiment. Assume you have the customer-level data used to create the table above and not just the aggregate statistics. (Note: there are a few potential strategies. You only need to describe one that works.)

In general, we must control for differences in age (18- to 34-year-olds versus everyone else) to recover the causal effect of interest. We can use subclassification, matching, or regression to do this. In all cases, we want to control for an indicator for 18- to 34-year-olds.

3. Discontinuities in Credit Card Limits (25 points)

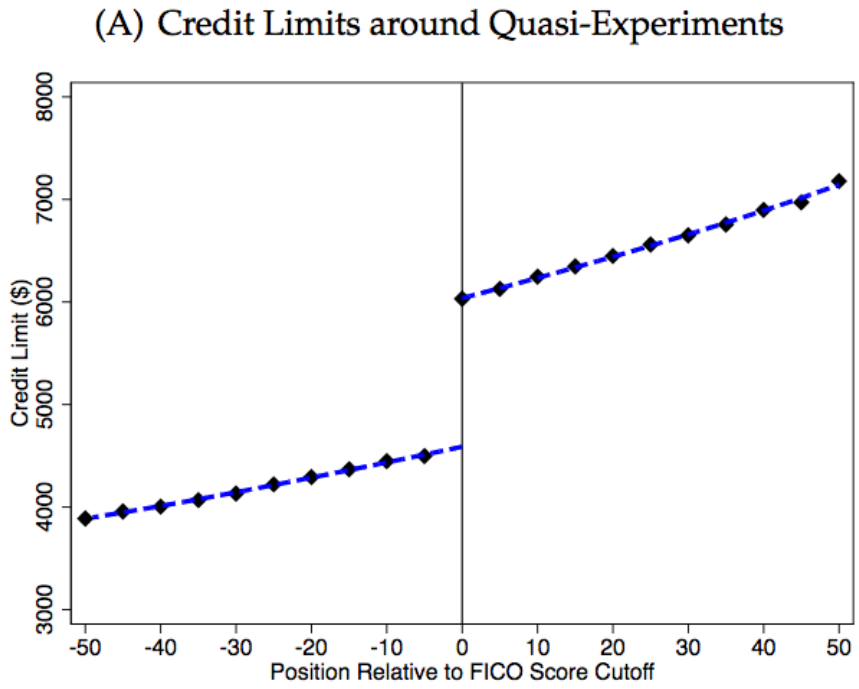
This question has 5 parts (A) through (E).

A bank is targeting a new credit card at college students. They must decide what credit limit--the maximum amount a borrower is allowed to borrow on the card--to offer to students. They want to set a credit limit that will maximize revenue relative to costs. The costs depend on how likely a borrower is to default on the credit card debt (i.e., not pay their bills). Revenue largely depends on how much fees the bank can collect on the account. Both the costs and revenue will potentially depend on the credit limit the bank sets.

The bank has hired you as a data scientist to analyze their existing data. In this question, we'll focus on the revenue side: at the borrower-level, what is the causal effect of credit limit on fees collected? If the bank increases the credit limit by $X, how much will fees collected increase or decrease?

After speaking with bank officials, you learn about an important institutional quirk: the bank uses a formula for determining credit limits for card applicants, and the formula depends in part on the interval of an applicant's credit score (e.g. 500-549, 550-599, 600-649, et cetera), which measures expectations for applicant's ability to repay debt. When an applicant moves from one interval to another, there is a jump in the allowed credit limit. This creates discontinuities in credit limits by credit score, as seen below.

Note on the figure: a commonly used credit score is known as a FICO score.



This creates a natural opportunity to use the regression discontinuity research design to estimate the causal effect of credit limits.

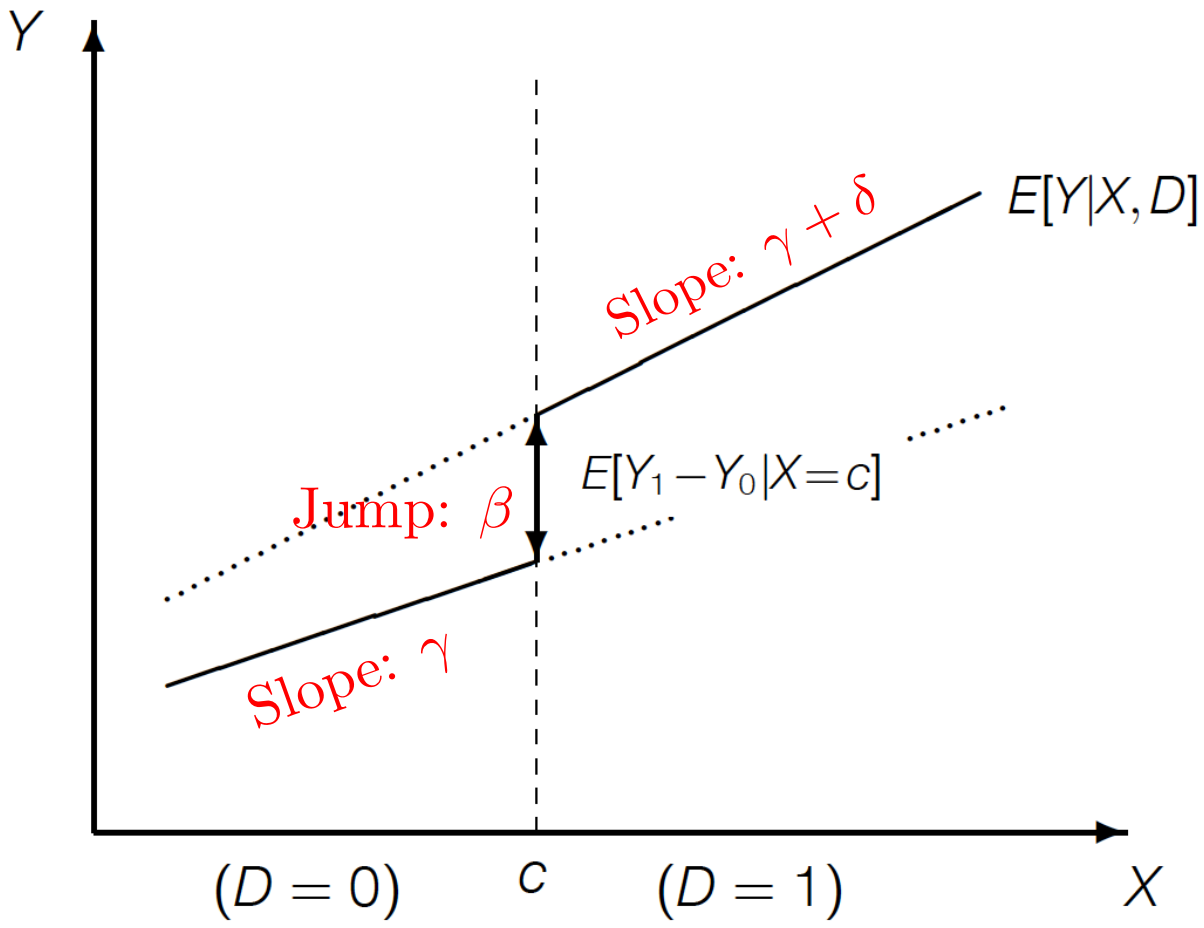
A) [5 points] In your own words, describe the logic of the regression discontinuity approach. Under what assumptions does it identify a causal relationship between the treatment and some outcome?

- **Key idea: Treatment changes discontinuously at some cutoff. Use discontinuity in $E[Y^{obs} \mid X]$ at the cutoff value $X = c$ to estimate the effect of D on Y for units near $X = c$**
- **Key assumption is continuity**
  - **$E[Y_1 \mid X]$ and $E[Y_0 \mid X]$ are continuous at $X = c$**
  - **Essentially have a 'local' randomized experiment**
  - **Baseline characteristics are the same above and below $X = c$**
  - **Only thing changing at cutoff is *treatment status***

B) [5 points] Write down a regression model to estimate the discontinuity in the figure above. You can simply label the cutoff as **c**. Write down which parameters correspond to each portion of the figure below. (There are **three** portions to label: i, ii, and iii. i and iii refer to slopes, while ii refers to the jump at the cutoff.)

$$Y_i = \alpha + \beta D_i + \gamma(X_i - c) + \delta(X_i - c) \times D_i + e_i$$

$$\text{where } D_i = \mathbf{1}_{\{X_i \geq c\}}$$



C) [5 points] Your colleague is concerned that the credit limits that the bank offers may also influence who decides to sign up for the card in the first place. Potential borrowers may learn the credit limit that the bank is offering them, and some may decide not to complete their sign-up in response. If they do not sign-up for the card, they will not show up in the data. Does this present a problem for the regression discontinuity approach? Why?

This likely violate the continuity assumption. While initial applicants with credit scores just below the cutoff will be similar to initial applicants just above the cutoff, the set of applicants that complete the sign-up process will differ systematically on one side and the other. We will no longer have a 'local' experiment at the cutoff.

D) [5 points] How would you test for the issue described in part (C)?

We can check for bunching. We can also check for discontinuity in baseline applicant covariates.

E) [5 points] Suppose you find that, at the cutoff, the average credit limit increases by $1000, and revenue increases by more than the cost. Based on your analysis, the bank is considering raising the credit limit for all borrowers with credit scores currently below the cutoff by $1000. Before doing so, the bank wants to know whether they should run a randomized experiment. What, if anything, would be gained from running an experiment? Suppose all the RD assumptions are satisfied, so that we have indeed identified a **causal** effect.

If the RD assumptions are satisfied, the RD estimates give us the causal effect of increasing credit limits by $1000 for applicants with credit scores at the cutoff. The change the bank is proposing will applying to all borrowers with credit scores below the cutoff, and so would apply to borrowers that meaningfully differ from borrowers at the cutoff. A randomized experiment would allow us to estimate the average treatment effect for this (more relevant) set of borrowers.