Winter 2019: Introduction to Data

Joel Whittier

Due: January 29, 2019

Introduction

Mayor Emanuel, having heard about your excellent training at the University of Chicago, has called on you to evaluate an important intervention that was recently conducted at Chicago Public Schools.

The dataset *student_data.csv* contains information on middle schoolers in three schools: Lincoln, Hamilton, and Washington. In 2016, the randomized educational intervention took place in all three schools with the intention of boosting scores on a standardized exam that tests math. The dataset contains exam scores from both 2015 (before the intervention) and 2016 (after the intervention). The exam scores are provided in units of σ, representing one standard deviation from the mean.

READ ME

- This p-set is worth 85 points
- You need to submit a PDF of your p-set on Canvas. If you choose to type it up in Latex (though not required), we've provided the .tex file of this p-set as a template for your writeup.
- Please name all submissions "lastname_firstname_pset1.pdf"
- No working in groups - all work must be your own.
- We will look at the code you submit to check for suspicious similarities between assignments, making sure results match with the write-up, etc. However, since this is not a programming class, your grade will only be based on the write-up (of course, if you have a coding error, it might affect the results you write here!).
- If you have a question about the pset, your first line of defense is to come to Anya or Nagisa's OH. If you have a question that someone else might have, post it on the Discussion board on Canvas. If you are sending either TA an email, please CC both of us so we can keep track of responses. If you are lost and struggling and don't want to struggle in front of others, send us an email and we can help you in private.

Question 0: Naive Regression / 10 Points

Let's say you start by approaching the data naively - you decide to simply load the dataset into Stata and run a few regressions.

Hint: In order for your regression to even run, you will need to first (1) destring any continuous variables and (2) create categorical variables from the ones that should be categorical. Do NOT do any other data cleaning - remember, you're being naive!

a. What is the estimated treatment effect without any controls? With controls?

Table 1: No Controls

	(1) Exam Score Untreated	(2) Exam Score Treated
Lagged Exam Score (Normalized)	0.0000120	0.0375
	(0.00)	(0.29)
Constant	0.0866	1.117
	(0.13)	(0.82)
Observations	204	185

t statistics in parentheses

R^2-adj (Untreated): -0.0050 R^2-adj (Treated): -0.0050

Table 2: With Controls

	(1) Exam Score Untreated	(2) Exam Score Treated
Lagged Exam Score (Normalized)	-0.00857	0.0400
	(-0.13)	(0.30)
Free Lunch	-2.675	1.285
	(-1.76)	(0.38)
Two Parents	3.786	-1.461
	(2.18)	(-0.42)
Constant	1.389	0.355
	(1.02)	(0.11)
Observations	204	185

t statistics in parentheses

R^2-adj (Untreated): .0276 R^2-adj (Treated): -.0144

b. Explain how you chose the controls you used.

We should start by determining criterion which may distort the relation between the relation between a students score in the current year and the previous year. Because the scores are comparing normalized score from year to year, we do not care for baseline differences in regression, but only in criterion that would inhibit/accelerate learning over the given year.

Because of this, it is hard to justify race as a control variable. Race is typically a barrier to a baseline difference; minorities tend to be placed in worse schools and are less likely to be placed in advanced programs, neither of which should effect an experiment controlled for school and classroom. And while discrimination within the classroom can certainly exist, in classrooms where the majority of students are People of Color, it is safe to say race variables would create more noise than value. Not to mention, we can control for race relatively easily through randomization.

The main differences within the year of learning should fall into two categories: differences in school and at home. The differences in school would mostly be accounted for by institution and classroom, both of which can be controlled via randomization. The differences at home can most be explained by free lunch and two parents. We can justify using both because they signify relatively different conditions. Free lunch means they may have less access to modern day resources which would help a middle school education, and one parent signifies their parents can offer less total time to help, which would effect the total learning that year.

c. Did the treatment effect change when you added your controls? Does that surprise you?

No, the treatment effect basically didn't change after adding controls. For the untreated sample, it decreased by 0.008582, which is incredibly insignificant. Both numbers basically signify the same thing: that the previous years score and the next year score are basically uncorrelated. This is obviously surprising, because people's standing in the previous year should certainly be correlated in the current year.

d. Do the results you obtained seem sensible?

Not at all. As mentioned in the previous question, The Lagged Exam coefficient implies not only that the treatment has no effect, but also that lagged exam scores and current exam scores are entirely uncorrelated. The controls also make some radical statements. A student of free lunch would expect their score to decrease by 2.675 standard deviations, which is obviously ridiculous. This means that a student who was average in the previous year, was on a free lunch program, and had a single parent would expect to place in bottom 1% the next year. But then, if the student had 2 parents, you would expect them to place in the top 10%. Even ignoring that the R^2 is basically 0, the answer presented is absolutely illogical and places much too much emphasis on the controls.

Question 1: Data Cleaning / 25 Points

Some time has passed and you've enrolled in Professor Levitt's data class. You realize that your naive regression and data cleaning from Question 0 isn't going to work.

After your Question 0 commands in your .do file, reload the same dataset into Stata using the `import delimited using student_data.csv,clear` command. This will clear your previous variables from memory, which is fine. Your .do file will now load the data twice. Your code should be organized like the stata template .do file on Canvas.

Use what you learned from class to decide how to clean the data. You will be graded on what you decide needs to be cleaned *as well as the solution you implement.* Be very careful in your code and specific in your write-up - not every cleaning step applies to every row or column in the data!

Your response should be organized in the following formats.

	Dirty data problem	Solution
1)	Inconsistent naming convention in "Free Lunch" and "Two Parent"	Changed "NA" to "No"
2)	Extra space in front of "School" variable	Remove all white-space from school entries
3)	Inconsistent capitalization in "School" variable	Make every School value proper
4)	Students have missing normalized scores	Set missing scores to 0 and create dummy equal to 1 when a score is missing
5)	"Free Lunch" and "Two Parent" are not in a regressible state	Create a dummy variable where 1 is "Yes"
6)	Periods at the end of a few "school" entries	Remove all periods from school entries

Let's say that as part of your cleaning, you decide that you need to eliminate some observations that do not make sense. If so, use the below template to organize your reasoning. Make sure to report the number of observations remaining after each step.

Filtering Step	Number of Observations Remaining
Start	500
Remove students with no scores recorded	$500 - 3$
Remove scores over three deviations	$500 - 3 - 10$
Remove scores which changed by over 2.5 deviations	$500 - 3 - 10 - 10$
Final	477

Feel free to write sentences or bullets about anything you think is important that doesn't fit into the tables.

Question 2: Estimating the Treatment Effect of the Intervention with and without Controls / 20 Points

a. What is the estimated treatment effect without any controls? With controls?

Table 3: No Controls

	(1)	(2)
	Exam Score Untreated	Exam Score Treated
Lagged Exam Score (Normalized)	0.936	0.809
	(21.63)	(15.56)
Constant	-0.0953	0.0674
	(-2.82)	(1.81)
Observations	242	235

t statistics in parentheses

$R^s q$-adj (Untreated): .6596 R^2-adj (Treated): .5074

Table 4: Controlled

	(1) Exam Score Untreated	(2) Exam Score Treated
Lagged Exam Score (Normalized)	0.928	0.814
	(21.04)	(15.49)
Free Lunch	-0.123	0.0386
	(-1.61)	(0.43)
Two Parent	0.0295	0.0273
	(0.33)	(0.28)
Exam Missing	0.00571	-0.142
	(0.04)	(-1.20)
Lagged Missing	-0.0625	0.0571
	(-0.61)	(0.53)
Constant	-0.00359	0.0407
	(-0.05)	(0.48)
Observations	242	235

t statistics in parentheses

$R^s q$-adj (Untreated): .6582 R^2-adj (Treated): .5036

b. Did the treatment effect change when you added your controls? Does that surprise you?

The treatment effect certainly had a more significant shift in this part than part 0, as part 0 had such extreme residuals it basically said nothing, however it still seems insignificant. Both of the associated slopes for the Lagged Exam Score are relatively close, as it goes from .936 to .928 for untreated and .809 to .814 for treated.

What really surprised me was how the constant term seemed to change for the controls. As I mentioned in question 1, while we expect Two Parent and Free Lunch to affect the constant term, we also expect it to create a difference in learning. Maybe it affects the constant term significantly more, however it still feels strange the β coefficient barely changed. It did show a significant disadvantage for Free Lunch students, but very little change in Two Parent. I do think the small value of Two Parent might have to do with its limited sample size however, as it seems only a few students do at 18%.

c. Do the results you obtained seem sensible?

They certainly seem plausible, but I still wouldn't call them entirely sensible. As we can see, the Treatment affect is likely to pull both higher scorers and lower scorers closer to the mean, which is pretty common for an education treatment. The Treatment also seems to remove a significant bias against low income students. However, it still deems two parents insignificant, which based off of intuition, should likely be more significant than free lunch. This is because missing time spent reading to children and helping children with homework sounds more significant than missing all necessary school supplies. However, it is still important to not shoehorn the data. But, as I mentioned previously it is still enough of an issue that it might show some problems with the sample.

Question 3: Randomization Validity / 15 Points

a. Does the randomization appear legitimate? (Hint: something went wrong)

No, as seen in the constant terms of the previously answer, the two groups seem to have different means. This can be seen in a simple ttest of the means of the two groups.

b. Provide evidence in the data that suggests flawed randomization. You can create a table,

So, as mentioned previously, we see there is an issue of the means of the two groups. It it is now important to check at which step did the randomization fail. Looking back at how I selected controls, I see classroom, Sex, Race, and School were expected to be randomized. Classroom has such small samples it is hard to run a chi-squared test (But we will uncover its bias soon enough). Sex and Race both pass. However, the school test fails significantly, and shows it isn't random.

Key
frequency *expected frequency*

school	treated: No	treated: Yes	Total
Hamilton	88	58	146
	74.1	71.9	146.0
Lincoln	101	91	192
	97.4	94.6	192.0
Washington	53	86	139
	70.5	68.5	139.0
Total	242	235	477
	242.0	235.0	477.0

Pearson chi2(2) = 14.4201 Pr = 0.001

This chart shows that Hamilton and Washington were not putting kids in the right groups at the right proportions. But there bias countered each other so it would not be immediately noticeable. To further uncover the culprits and the direction of the bias, we look at the means of each individual school and the groups they placed.

Means, Standard Deviations and Frequencies of lagged_exam_score_normalized

school	treated No	treated Yes	Total
Hamilton	.18861869	-.1487353	.05460135
	.81386825	.67722605	.77791936
	88	58	146
Lincoln	-.01228638	-.07073268	-.03998749
	.69636242	.72301576	.70785882
	101	91	192
Washington	.1116302	.03710018	.0655181
	.85856793	.72752196	.77787739
	53	86	139
Total	.08790877	-.05052215	.01970905
	.77936772	.71443082	.7505009
	242	235	477

Hamilton is the most extreme culprits, with a difference of .3373. Washington and Lincoln both did the same thing, albiet to a lesser degree. This means that the "Yes" group is basically the group that was worth off. What is even more strange to me is that the schools were put in unequal groups. It wasn't like they were just placing the worse half in one and the better half (generally) in the other, the results seemed much more intentional. It seemed like Hamilton and Washington both purposely tried to set up positive and negative numbers on both side, while Lincoln casually grouped its worse students into the yes category.

Question 4: Estimating the True Treatment Effect / 15 Points

a. Calculate the best estimate of the true treatment that corrects for the randomization error.

So, to correct for the sample problems, I figured we should resample the data. I tried bootstrapping the data (essentially replicating data points from underrepresented groups) and these are the results I got.

Table 5: With Controls

	(1) Exam Score Untreated	(2) Exam Score Treated
Lagged Exam Score (Normalized)	0.928	0.814
	(21.04)	(15.49)
Free Lunch	-0.123	0.0386
	(-1.61)	(0.43)
Two Parents	0.0295	0.0273
	(0.33)	(0.28)
Exam Missing	0.00571	-0.142
	(0.04)	(-1.20)
Lagged Missing	-0.0625	0.0571
	(-0.61)	(0.53)
Constant	-0.00359	0.0407
	(-0.05)	(0.48)
Observations	242	235

t statistics in parentheses

$R^s q$-adj (Untreated): .6582 R^2-adj (Treated): .5036

As you can tell, bootstrapping didn't really affect the table. This is most likely because the data we request simply doesn't exist in this sample so replicating it will have little effect.

Instead, to estimate the treatment effect, lets look at the difference in expected value. We know that the expected value of the lagged exam score is 0, 18% have two parents, so we can use .18, and .75 for Free Lunch.

Expected value untreated: -0.0954

Expected value treated: 0.0701

So, we can say on average, we expect the treatment to help the students. However there are still obvious weaknesses to this analysis. We know that the treatment certainly helps students below the curve, but there's is not strong reason to suspect it also helps those above the curve. Probably the most significant result was the change of coefficient in free lunch. Whatever method this treatment is using certainly helps students of lower socio-economic status. I don't think it is safe to say this treatment is better for everyone, but after running this analysis, it can certainly benefit quite a few students.