Labor Economics
Problem Set

Due Date: Monday, July 29 at 12:30pm

- You may work together, but each student must submit their own solutions.
- Several questions will reference a paper. You do not need to fully read it, but you may find it helpful to read the introduction and look at the referenced part of the paper.
- Plagiarism will not be tolerated. This includes, but is not limited to, copying directly from another student, a paper, or textbooks. Your phrasing must be your own and should demonstrate that you understand the answer. Plagiarism will result in receiving a grade of zero for this assignment.
- You may quote a paper or other reference, but you must: 1) cite your quotation, 2) keep your quotation limited to 1-2 sentences, and 3) add further explanation to the quotation *in your own words*.
- You will receive an extra 5 points if you submit typed up solutions.
- Questions marked with a ★ are bonus exercises. They are optional, but highly encouraged, and you will receive points for answering them (grades above 100% are possible).

Question 1 [20 points]

For this question, you will be running your own regressions. Download the Excel file **Data.xlsx** from Courseworks. I will explain how to do this using Excel. However, you are welcome to use any statistical software (e.g. Stata, R, Python) if you know how to use them. For those using Excel, you need to first enable the Analysis ToolPak add-in (steps for Windows and Mac). The link for the Windows steps also explains the basics of how to run a regression in Excel. No matter how you do this, you must show your regression outputs in a table.[1]

The dataset shows a simulated dataset. The column headers represent each variable: y, x_1, x_2, x_3, x_4, f_1, f_2, ε (in that order). In the sheet **data**, you will find the dataset that you will use to answer these questions. In the sheet **formula**, you will see how the data was generated.

1. How many observations are in the dataset? Which variables are dummy variables?
2. Look at the formula for y in cell A2 of the formula sheet. This captures the true way that y is generated.
   (a) Write out this formula in the usual regression equation set up but leave the coefficients as parameters, e.g. something like $y = \beta_0 + \beta_1 x$. Make sure to write the equation as a function of x_1, x_2, x_3, x_4, f_1, f_2, and ε.
   (b) What are the true values for each coefficient?
3. Run a regression of y on all the variables, except ε (we will treat this as the unobservable error term). How do the results compare to your answer in 2(b)? (look at both the estimate values and the p-values)

[1] To save time, you can just take a screenshot of the table from Excel/whatever you are using. Do **not** take a picture of your computer screen with your phone. For Windows, use the Snipping Tool. For Mac, use the keyboard shortcuts.

4. Repeat (3), but now run a regression only on the first 100 observations. Repeat this again but run the regression on only the first 10 observations. What do you see changing in the results each time? What lesson do you draw from this?

★ 5. f_1 and f_2 can be interpreted as fixed effects (e.g. f_1 represents group 1 and f_2 represents group 2). These effectively capture average differences in groups (see this fantastic animated plot from Nick Huntington-Klein to see how this works).

   (a) What is the conditional mean of y given $f_1 = 1$? What about conditional on $f_2 = 1$? Use Excel's AVERAGEIF() function to do this quickly.

   (b) Why do your answers to 5(a) differ from the corresponding coefficient values in 2(b)? (hint: is there a third group?)

6. For each of the following parts, suppose that the variable listed was not observed. Re-run the regression in (3), except not including the listed variable.[2] What happens to the estimates as compared to (3)? Can you explain why this happened? (hint: look at the formulas)

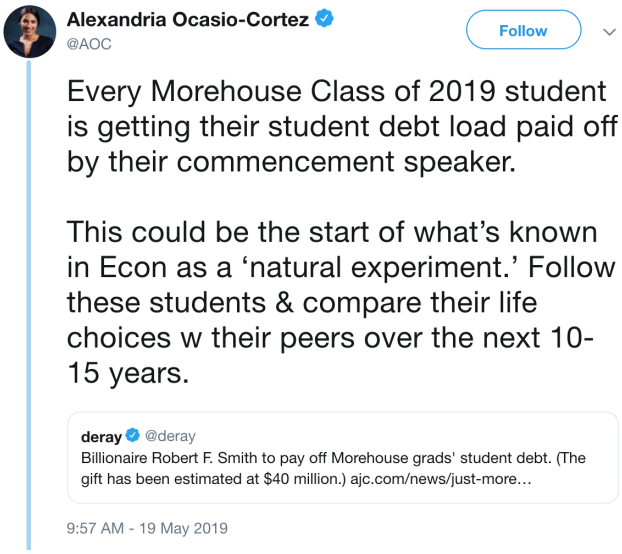
   (a) x_4

   (b) x_3

   (c) x_2

Question 2 [5 points]

In May 2019, billionaire investor and philanthropist Robert F. Smith announced during the Morehouse College graduation ceremony that he would pay off all the student debt of the graduating class of 2019 (USA Today). In response, Representative Alexandria Ocasio-Cortez (D-NY) tweeted:



1. Do you agree that this case study could be a natural experiment? If so, describe how you would design the experiment and what outcomes you would be interested in studying. If not, argue what issues prevent this from being a natural experiment and why they cannot be overcome.

[2] For Excel, the input X-range must be contiguous (i.e. the columns must all be next to each other). This means you'll need to copy the sheet, delete the listed variable and run the regression on this (smaller) dataset.

2. Suppose that researchers did conduct this study. Putting aside issues of *internal* validity (which you addressed above), what issues of *external* validity might this study have? In other words, can we generalize the results?

3. In August 2018, NYU Medical School announced it will offer full scholarships to all current and future students in its program (NPR).

   (a) Suppose that a researcher wanted to compare the outcomes of pre-2018 NYU graduates to post-2018 graduates to see the effect of the scholarship program. Do you believe this will give us the causal effect of the program?

   ★ (b) How do the NYU and Morehouse cases differ? In particular, what about the timing of the announcements could generate different responses by the students? (putting aside that one was for college students and the other for medical students)

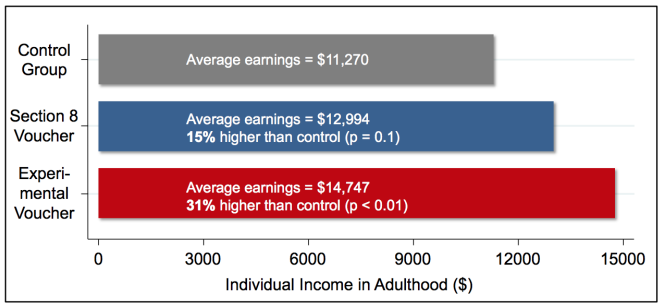
Question 3 [15 points]

The Moving to Opportunity (MTO) experiment was an experiment in the 90s to study the effects of giving low-income families living in public housing vouchers to move to new neighborhoods. There were 4,600 families in the study that were randomly assigned to three groups:

- Treatment 1: The Section 8 voucher group. They were offered the standard voucher.
- Treatment 2: The experimental voucher group. Their voucher restricted the neighborhoods the families could move to. They were restricted to moving to areas with a low poverty rate (<10%)
- Control: was not offered any vouchers

Chetty, Hendren and Katz (2016) study the long-term effects of this experiment. You can read a summary of their paper here.

1. The authors look at the incomes of the children in the 4,600 families once they grow up. They summarize their results in the following graph.



   Did both treatments have a causal effect in increasing incomes in adulthood? Were these statistically significant at the 5% level?

2. What are possible reasons that the experimental voucher has larger effects than the Section 8 voucher, even though it restricted the participants' choice?

3. While the MTO experiment *offered* vouchers to families, there was no requirement that they had to actually use them. Use the next questions to help you fill in the table below. We will be focusing on the children who were under 13 at the time of the experiment.

   (a) Calculate the **compliance** or **take-up rate** for each group. This is the fraction of people who used the voucher out of all the people who were offered a voucher:

   (b) The control group's mean income in adulthood was $11,270.3. Calculate the difference in mean income between each treatment group and the control group. This is the **intent to treat (ITT)**.

   (c) The ITT should suggest much smaller effects than the numbers reported in (1). The reason for this due to the imperfect compliance (not everyone actually used the voucher). Explain why this would create a make the effect of the voucher look smaller?

   (d) To calculate the effect of actually using the voucher, we need to "scale up" the ITT. To do this, divide the ITT by the compliance rate. This gives us the **treatment on the treated (TOT)**. In general, how does the TOT change as the compliance rate changes (holding fixed ITT)?

   (e) Use the TOT to calculate the percentage change in income for each treatment group (relative to the control). Do these look similar to the results in (1)?

	# Used Voucher	# Offered Voucher	Mean Income	Compliance Rate	ITT	TOT	%Δ Effect
Section 8	1,296	1,969	12,379.6				
Experimental	680	1,427	12,894.3				

4. The ITT gives us the effect of *offering* someone a voucher. The TOT gives us the effect of *using* the voucher (for those people whose behavior was induced to change through the voucher offer). Suppose you were a policy maker looking to study the effect of housing vouchers. Which of these effects would you be more important for you? Explain your answer.

5. Call income in adulthood Y. Let D be a dummy for whether the voucher was used. Let Z be a dummy for whether the voucher was offered. For simplicity, let's just assume there is one treatment group. Consider the following regression equations:

$$Y = \beta_0 + \beta_1 Z + \varepsilon$$
$$Y = \alpha_0 + \alpha_1 D + \nu$$
$$D = \gamma_0 + \gamma_1 Z + \mu$$

   where ε, ν, μ are all error terms.

   (a) Which coefficient gives us the ITT? Which coefficient gives us the compliance rate?

   (b) We can use the voucher offer (Z) as an instrument for voucher usage (D). Explain why this would be a good instrument.

   (c) Show that if we do that then $\alpha_1 = \frac{\beta_1}{\gamma_1}$ (hint: plug in equations). How should we interpret α_1?

★ 6. Explore the Opportunity Atlas. Pick a city and zoom in to the census tract level. Play around with various outcomes and sub-groups and make sure you understand how to interpret the map. Give three interesting results you can see in the data. Make sure to include pictures of the maps (use the download as image option in the bottom left)

Question 4 [5 points]

Give short answers (2-3 sentences is enough) to the following questions

1. Imagine the following (fictional) scenario. School District A implemented a new policy for seniors in high school. If they have an average GPA above 3.75, they are allowed leave campus during their free period. A regression discontinuity (RD) analysis was done using a window of students who were within 0.1 points of the cutoff. It was found that this policy led to better outcomes for the students (higher test scores, better focus in class, less disciplinary action etc). School District B looks very similar to A on many observable characteristics. They have decided to implement the program in their schools, but given the results were so positive, they have decided to lower the minimum grade threshold to 3.5 in order to have more students participate. Do you agree with district B's logic? If we conducted an RD using the district B data, should we expect the same results as in A?

2. For each of the following, state what the extensive and intensive margin are:

   (a) Having children

   (b) Donating to charity

   (c) Discrimination

3. During the recent Supreme Court case over adding a citizenship question to the census, Justice Gorsuch made the following remark:

   > "Justice Neil Gorsuch weighed in. "There could be multiple reasons why individuals don't complete the form." He continued: "We don't have any evidence disaggregating the reasons why the forms are left uncompleted. What do you do with that? I mean, normally we would have a regression analysis that would disaggregate the potential cause and identify to a 95th percentile degree of certainty what the reason is that persons are not filling out this form and we could attribute it to this question. We don't have anything like that here. So what are we supposed to do about that?" (New York Times)

   Do you agree with Justice Gorsuch's claim? Explain your answer.

★ 4. In 2014, more than 90,000 people applied to be sanitation workers in NYC (only 500 were accepted!)

   > "One reason applicants may be lining up to become a sanitation worker is the pay. The starting salary is low, $33,746, but when you factor in overtime, it averages $47,371 in the first year. And after 5½ years, the salary jumps to an average of $88,616 dollars. That's not bad, considering the average annual pay for New York City transit workers is $77,991, New York City teachers is $68,151 dollars and New York City Parks Department employees is $50,042... And there are other perks too, such as 10 percent extra pay for night shifts, double pay for Sundays, 25 vacation days after six years of service and an unlimited number of sick days." (Al Jazeera America)

   Does the above fit the story of compensating wage differentials? Are there other potential explanations?

Question 5 [15 points]

Consider a reform that decreased the generosity of welfare benefits in New Jersey. At the same time that the reform occurred, there was no reform to the generosity of welfare in neighboring Pennsylvania. Suppose we were able to collect the following data on average hours of work and average welfare benefit amounts for recipients (per month) in the two states:

	New Jersey		Pennsylvania	
	Average hours of work	Average benefit	Average hours of work	Average benefit
Pre-Reform	60	$1,000	55	$1,000
Post-Reform	80	$600	70	$800

1. Given the available data, explain why a difference-in-difference (DID) would be an appropriate strategy here if the goal is to estimate the effects of the reform on labor supply? What assumption do you need to make for this to be a valid strategy?
2. What is the estimate of the effect of the reform on hours of work of welfare recipients?
3. What is the estimate of the effect of the reform on the average welfare benefit amount?
★ 4. What is the implied elasticity of labor supply to the welfare benefit amount? How should we interpret this?
    - An elasticity is always a ratio of percentage changes. Your answers to (2) and (3) will give you the *level* change, but to make it a *percentage* change you will need a baseline level. For this, use the average of the pre- and post-periods (this is called the midpoint method)

The above is a standard example of DID. DIDs are often presented as a comparison over time with a pre- and post-period. But, in fact, it doesn't have to be over time. Madrian (1994) studies a phenomenon called **job-lock**. This is when people are reluctant to change jobs because they are afraid of losing the health insurance they receive from their employer.

5. To estimate the job-lock effect, why is it a bad idea to just compare the job mobility rates (how often people change jobs) for people who get health insurance through their employer versus the mobility rates of those who get no health insurance from their employer?
6. Consider the following two groupings of people:
    (a) People with high medical expenses vs people with low medical expenses
    (b) People with only health insurance from their employer vs people who also have health insurance from another source (e.g. their spouse's insurance)

    Within each grouping, who do you think is more likely to experience job-lock? (i.e. is less likely to move jobs because of their health insurance status). As in (5), why is it a bad idea to compare the job mobility rates between the two groups to determine the effect of job-lock?
7. The standard DID can be written as a 2×2 table (rows are treatment/control and the columns are pre/post). We can setup two DIDs using the groupings in (5) and (6). Each column represents people without employer health insurance (No HI) and people with insurance (Has HI). This represents the groups suggested in (5). The rows represent the groups suggested in (6). Each cell shows the probability that someone in that group will change their job.

	No HI	Has HI
Low Expenses	0.253	0.092
High Expenses	0.224	0.051

	No HI	Has HI
No other HI	0.256	0.085
Other HI	0.244	0.115

In the left table, think of three possible characteristics as dummy variables:

- JL = 1 if the person experiences (significant) job lock, 0 if not
- EX = 1 if the person has high medical expenses, 0 if not
- HI = 1 if the person has a job with health insurance, 0 if not

Then, we want to run this regression, where Y is the probability of job change:

$$Y = \beta_0 + \beta_1 \cdot JL + \beta_2 \cdot EX + \beta_3 \cdot HI + \varepsilon$$

We are interested in estimating β_1. Make an argument for why $JL = HI \times EX$. Set up a DID using the above equation (hint: what does the equation look like for each cell?). Using the information in the table, what is the estimate for β_1?

★ 8. Repeat the steps for (7) for the table on the right (make sure to show your equation and explain each of your variables). Do you get similar estimates? Do both of these show evidence of job lock? Explain your answer.

Another extension to the standard DID is to look at multiple groups where treatment was applied at different times. For example, suppose we wanted to study the effect of policy D. There are three states ($S1, S2, S3$). State 1 implemented policy D in 2015, state 2 implemented it in 2017, and state 3 never implemented it.

9. Draw a dataset that summarizes this information. The columns should be: state, year, D (a dummy variable for whether the state implemented the policy), and Y (the outcome variable). The dataset should go from 2014 to 2018. You can leave the Y column blank.

10. The diff-in-diff regression we want to run is:

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \phi_i + \delta_t + \varepsilon_{it}$$

What is the unit of observation? What do ϕ_i and δ_t capture?

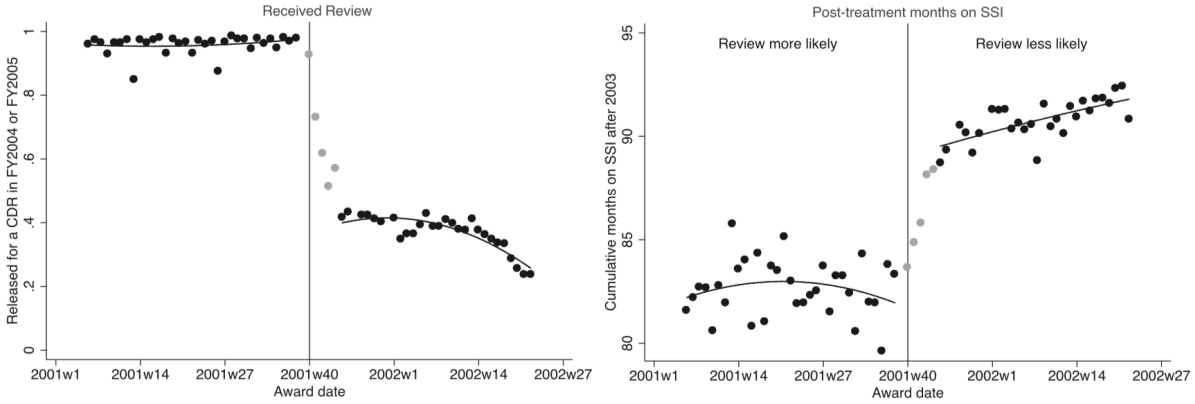
★ 11. Explain how using this setting we can set up four standard DIDs (by standard, I mean a two-state DID comparing the pre- and post-periods where one state implemented the policy and the other did not). The coefficient of interest β_1 in the above regression is just a weighted average of the four DIDs (hint: look at Figure 2 from this paper)

Question 6 [15 points]

Deshpande (2016) studies how parents respond when their child is removed from SSI.[3] There was a large cut in the budget for medical reviews between fiscal year 2004 (2004) and 2006.[4] Children due for a review in 2005 and 2006 were much less likely to actually receive it. Since children receive medical reviews three years after they enter SSI, looking at those who entered in 2001 (and therefore were scheduled for a review in 2004) versus those who entered in 2002 (and therefore were scheduled for a review in 2005) should provide a discontinuity to analyze. This analysis is done using a regression discontinuity (RD) analysis.

[3] Note that this is a different paper to the one we covered in class
[4] For simplicity, all years here will refer to fiscal years (FY)

1. The following graphs show the discontinuity being analyzed. On the left, we see a clear discontinuity in probability of being reviewed. What is the running variable? Why do we see a discontinuity in months on SSI as well (graph on the right)?



Source: Deshpande (2016), Figure 1

2. The author runs the following regression for a person i with a child on SSI:

$$Y_i = \alpha + \beta \cdot FY2001_i + \gamma AwardDate_i + \delta \left(AwardDate_i \times FY2001_i\right) + \kappa X_i + \varepsilon_i$$

where $FY2001_i$ is an indicator for whether the child first received SSI in 2001 (the award date). What does the equation look like for someone whose award date was in 2001 versus in 2002? What is the main coefficient of interest in this RD?

3. For an RD to be valid, we need the treatment and control group to look similar on observable characteristics before the quasi-experiment (i.e. balanced).

	Point Estimate	SE
Diagnosis		
Infectious	−0.0001	(0.0005)
Neoplasm	0.0005	(0.0019)
Endocrine	0.0011	(0.0013)
Blood	0.0014	(0.0019)
Mental	−0.0102	(0.0086)
Nervous	0.0063	(0.0043)
Sensory	−0.0044	(0.0033)
Circulatory	−0.0010	(0.0013)
Respiratory	0.0118***	(0.0039)
Digestive	0.0028*	(0.0017)
Demographics		
Male	0.0078	(0.0086)
Year of birth	0.0157	(0.0477)
Age at initial receipt	−0.0245	(0.0475)
Single mother	0.0101	(0.0092)
Young parent	0.0095	(0.0092)
Pretreatment outcomes		
Months on SSI	−0.0022	(0.0019)
Family disability applications	−0.0004	(0.0026)
Family disability receipt	50.07	(78.9)
Household earnings	319*	(189)
Total household income	329*	(191)

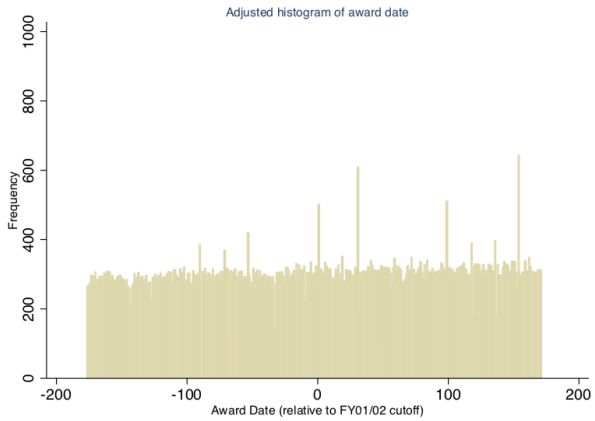
$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$. The table reports estimates for covariate balance tests using equation (2). Diagnoses constituting less than 1% of sample are not shown but are not significant. $N = 49{,}662$.

Source: Deshpande (2016), Table 2

The above table is constructed by running the regression shown in (2) and using different outcomes

for Y. The outcome used is in the first column.[5] The estimate of the coefficient of interest (your answer to (2)) is the second column. The standard error for this estimate is in the third column. Based on these results, does the sample look balanced? Are there any factors that you may be concerned about? Explain your answer.

★ 4. Another property an RD has to have is that the running variable cannot be manipulated. Explain how manipulation can occur in this situation and how it would bias the results. Using the figure below, do you think there is evidence of manipulation?



Discontinuity in density at cutoff

	Point Est.	*Std. Error*
FY01/02	0.058	(0.013)
FY00/01	0.030	(0.017)
FY99/00	0.046	(0.014)

Source: Deshpande (2016), Figure A5 (Appendix)

5. After estimating the regressions, the author finds the following:

> "As a result of the discontinuity in medical review receipt, the probability of receiving an unfavorable medical review decision drops by 5 percentage points across the cutoff... A child who enters the program on the last day of FY2001 spends approximately 4.5 fewer months on the program each year and receives $2,800 less in total SSI payment each year in the eight years following the medical review event than a child who enters the program on the first day of FY2002."

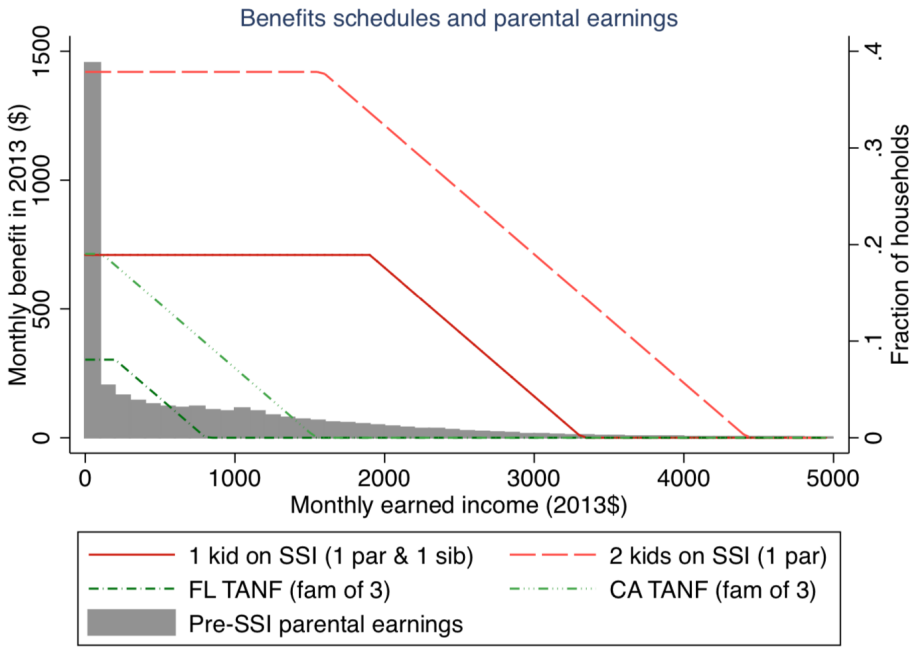
   (a) You can interpret the 5% as a compliance rate and –4.5 months and –2800 dollars as ITT. What would the TOT be? What is the difference in interpretation between the ITT and TOT in this setting? (make sure you've done Question 3 first to understand this)

   (b) Is a 5% compliance rate low? Should we have reasonably expected a 100% compliance rate?

6. The figure below shows a lot of information. Let's focus on the solid red line (1 kid on SSI), which shows the benefit schedule for SSI. The x-axis shows the parent's monthly income. The y-axis (on the left) shows the monthly SSI benefit amount.

   (a) Draw the parents' leisure-consumption budget constraint with and without SSI. Assume that their wage is $8 per hour. Plot consumption on the y-axis, with a price of 1. Assume they have 800 hours of leisure per month. To make it easier, let's simplify the schedule to be as follows: payment is $720 for income between 0 and 2000. For income between 2000 and 3200, the benefit amount is calculated as $B = 720 - 0.6(M - 2000)$, where M is monthly income. Beyond 3200, the benefit is zero.

[5] Note that sometimes the outcome refers to the parent and sometimes it refers to the child (the person who is actually on SSI). The diagnosis and first three demographic variables refer to the child, for example.



Source: Deshpande (2016), Figure A13 (Appendix)

(b) How does the benefit schedule differ from the EITC schedule? What impact do you think SSI might have on labor supply? (on the extensive and intensive margin)

(c) The histogram on the graph shows the distribution on incomes for the parent before their child entered SSI. Assuming that people's income would have stayed the same had their child not entered SSI, which people in the distribution experience an income effect and which people experience both an income and substitution effect? Explain your answer using indifference curves and the diagram you drew in (a).

★ (d) The author finds that parent's annual earnings increase by $470 for those in the treated group. The average annual loss in SSI payments in $330. Based on the information above, do you think this is more likely to be an income effect or a combined income and substitution effect? If it is a pure income effect, why is this a surprising result?

★ 7. The author also studies this policy change using a difference-in-difference (DID) setup. The people due for medical reviews in 2004 did receive them while the people due for reviews in 2006 (i.e. who were awarded SSI in 2003) effectively did not receive them (only 1.2% did). The regression is estimated for a person i in time t:

$$Y_{it} = \sum_{t=1992}^{2012} \beta_{1t}(Year_t \times FY2004_i) + \sum_{t=1992}^{2012} \beta_{2t} Year_t + \alpha X_{it} + \mu_i + \varepsilon_{it}$$

where Y_{it} is person i's outcome in time t, $Year_t$ is a dummy variable for year t and $FY2004_i$ is a dummy for whether person i's child was awarded SSI in 2004.

(a) This setup effectively checks the size of the difference between the treatment and control group in every single year. Show what the equation looks like for the treatment and for the control group. What set of coefficients capture the difference between them?

(b) Suppose that Y_{it} is person i's income in time t. Given that we want a DID, what should we expect of

the coefficients to satisfy the parallel trends assumption? What should the sign of the coefficients be so that we get the same result as the RD? (hint: in a DID, the post-period starts at a particular year - what is that particular year in this setting?)

(c) In the regression μ_i is a person fixed effect. Why can we include that now when it was not included in the RD estimate? (hint: think about the structure of the data)

References

Chetty, Raj, Nathaniel Hendren, and Lawrence F Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review*, 106(4): 855–902.

Deshpande, Manasi. 2016. "The Effect of Disability Payments on Household Earnings and Income: Evidence from the SSI Children's Program." *The Review of Economics and Statistics*, 98(4): 638–654.

Madrian, Brigitte C. 1994. "Employment-based health insurance and job mobility: Is there evidence of job-lock?" *The Quarterly Journal of Economics*, 109(1): 27–54.