FINAL EXAM

May 11th, 2023

I - INSTRUCTIONS

- You will have from 7am to 7pm to complete this exam.
- By 7pm, save your exam and submit it and your R **script** to Canvas.
- The exam is out of 100 points. If you get bogged down on something you may want to move on to gather points from questions you understand more quickly.
- In open-answer questions please be concise.
- If you have a clarification question you may email the entire teaching team. We will do our best to respond.
- If you're unsure of how to answer a question, just use your best judgment and explain why you did what you did.

II – EXAM POLICIES

Although collaboration in the problem sets is permitted, exams are different. There is no collaboration allowed. This exam is open-book, open-notes, open-Internet, but you **may not speak to each other or anyone else, in person or on the Internet or in any form.**

The standard rules for plagiarizing apply – you must type your own answers and code.

Before you begin, please read this excerpt from the HKS Honor Code and write your initials in the box below:

Cheating on assignments or exams, plagiarizing or misrepresenting the ideas or language of someone else as one's own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs. Using someone else's words or concepts without attribution is a serious violation of the Academic Code. It is the student's responsibility to learn and use the proper forms of citation. If students submit work either not their own or without clear attribution to the original source, including but not limited to the Internet, they will be subject to discipline by the HKS Administrative Board, ranging from a warning to required withdrawal or expulsion from Harvard Kennedy School.

		Your Initials
	I certify that my work in this exam complies with the Harvard Kennedy School Academic Code and that I did not communicate with anyone during the exam.	SC

Introduction:

In this exam, you will be building on the results from Deshpande and Mueller-Smith (2022):

> Manasi Deshpande, Michael Mueller-Smith, Does Welfare Prevent Crime? The Criminal Justice Outcomes of Youth Removed from SSI, *The Quarterly Journal of Economics*, Volume 137, Issue 4, November 2022, Pages 2263-2307, https://doi.org/10.1093/qje/qjac017

As in the original paper, we want to explore the effect of losing Supplemental Security Income benefits on criminal justice and employment outcomes. We will examine four hypothetical scenarios in which we imagine that the United States Social Security Administration implemented various types of administrative procedures from 1997-2017. As such, you should expect discrepancies in institutional details of all four scenarios compared to the actual policy setting.

You are encouraged to refer to the original paper to help understand the context, though the data you will use is completely simulated, so the paper's results may or may not hold.

Part I [28 points + 4 bonus points] - Instrumental Variables

The file `api210_final_part01.dta` contains data on people born between 1976 and 1998 who received Supplemental Security Income as children.

In this new scenario, the Social Security Administration implemented an audit from 1994 to 2016, in which children who received Supplemental Security Income (SSI) were randomly selected in each state to complete an eligibility review when they turned 18. It can be assumed that within each state, all children who received SSI had an equal chance of being selected for review, but the probability of being selected varied between each state. Almost all child SSI beneficiaries not selected for review were allowed into the adult SSI program. In contrast, those selected for review had an increased chance of losing their benefits after they turned 18. The number of criminal charges on the criminal record of each person in the data between the ages of 18 and 38 is then observed. The variables in the dataset are:

variable	label
recipient_id	SSI Recipient ID
st_id	State in which the recipient resides at age 18
ssi_lost	Did the recipient lose SSI benefits after age 18?
charges	Total number of criminal charges between ages 18 and 38
reviewed	Was the recipient selected for review when they turned 18?
physical	Does the recipient have a physical condition?
age_receive	Age when the recipient first received SSI

1. Write down the regression equation(s) you would use to estimate the effect of losing SSI benefits in adulthood on the number of charges that a person has on their criminal record

from the ages of 18 to 38. Use all appropriate and valid controls. What is the coefficient of interest? **(4 points)**

First Stage: $ssi\_lost = \alpha_0 + \alpha_1 * reviewed + \alpha_2 * state + \alpha_3 * physical + \alpha_4 * age\_received + \mathcal{E}$
Second Stage: $charges = \beta_0 + \beta_1 * ssi\_lost + \beta_2 * state + \beta_3 * physical + \beta_4 * age\_received + \mathcal{E}$

The coefficient of interest is β_1 since it shows the influence of losing SSI on charges, holding all else (state, age, physical) constant.

2. For the following questions, please answer using both the potential outcomes and variables notation, in plain English

    a. What is/are the key identifying assumption(s) in this context? (Describe all of them, but be very brief for each. One equation and one sentence for each explanation should do.) **(6 points)**

    The following key identifying assumptions should be met in this context:

    1. *Relevance: The instrument must be relevant to the treatment variable, which in this context means that the loss of SSI should be relevant to/influence the charges received.*
       On average, $E[D_i(1) - D_i(0)] \neq 0$ where $D_i(1)$ when SSI was lost and $D_i(0)$ when SSI was not lost.
    2. *Independence: The instrument is as good as randomly assigned, so in other words, everyone has an equal chance of getting dropped for SSI.*
       Mathematically, $D_i(1), D_i(0) \perp Z_i$.
    3. *Exclusion Restriction: The instrument only affects the outcome through the treatment, or in this case whether someone was selected for review should impact their charges only through whether they lost their SSI.*
       Mathematically, $Y_i(D_i(0), Z_i = 1) = Y_i(D_i(0), Z_i = 0)$
       $Y_i(D_i(1), Z_i = 1) = Y_i(D_i(1), Z_i = 0)$
    4. *Monotonicity: The instrument influences the treatment in the same direction for everyone.*
       Mathematically, $\pi_{1i} = (D_i(1) - D_i(0)) \geq 0$ for all i
       $E[D_i(1) - D_i(0)] = P(D_i(1) > D_i(0))$

    b. What type of causal effect do we recover if the assumptions hold? To what populations is this result relevant? **(2 points)**

    Assuming all the above-mentioned assumptions hold, we recover the Local Average Treatment Effect (LATE), which is the effect of treatment on a specific subgroup identifiable by their compliance status. The LATE in this study refers to compliers who would lose SSI only if reviewed, and not otherwise.

c. [BONUS] How does this population differ from the overall population in terms of having a physical condition? What proportion of recipients selected for eligibility review are compliers? **(4 points)**

This population (the compliers) might be different from the overall population in terms of having a physical condition in that people with physical conditions may tend to comply more with the review process than others.

3. Run the regression model you proposed above to estimate the effect of receiving adult SSI on the number of criminal charges in adulthood, using appropriate controls and standard error adjustments.

a. In words, justify your adjustment of the standard error **(2 points)**

Cluster-robust standard errors were necessary due to the clustering of the population in states, which could result in correlations in outcomes dependent on states (such as state-level characteristics that can influence the outcome in some direction). Using robust standard errors addresses potential violations of homoscedasticity and clustering in the data, leading to valid standard errors and more accurate inference.

b. Report the main coefficient, the standard error, the t-statistic, and the two-sided p-value for the null-hypothesis that the effect is null (**4 points**)
Abridged Table Below:

Table 1:

	Dependent variable:
	charges
ssi_lost	5.367***
	(0.147)
physical	−0.277**
	(0.121)
age_receive	0.443***
	(0.010)
factor(state_fips)2	0.923*
	(0.525)
factor(state_fips)4	1.068**
	(0.525)
factor(state_fips)5	1.122**
	(0.525)
factor(state_fips)6	0.707
	(0.525)

Coefficient of interest: 5.367
Standard Error: 0.147

t-statistic: 36.468
P Value: Approx. 2e-16 (significant even at the 1% level)

4. Interpret the estimate you provided above. You may assume that the identification assumptions were satisfied. **(2 points)**

   This regression shows that there is an increase of approximately 5.4 charges for an individual between the ages of 18 and 38 who loses his SSI benefits compared to an individual in the same age range who does not. This result is also statistically significant at the 1% level.

5. Do you think all the identifying assumptions are reasonable or justifiable in this context? Please explain each one using theory and/or empirical tests (i.e., using R, where appropriate), and show the results of any tests you carry out. **(8 points)**

   Relevance: *This assumption can be verified using the IV regression mentioned above:*

```
> first_stage <- lm(ssi_lost ~ reviewed + physical + age_receive + factor(state_fips), data=part_1)
> summary(first_stage)

Call:
lm(formula = ssi_lost ~ reviewed + physical + age_receive + factor(state_fips),
    data = part_1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4336 -0.2570  0.1032  0.2216  0.7273

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.1532050  0.0197151   7.771 8.54e-15 ***
reviewed             0.4915775  0.0056004  87.776  < 2e-16 ***
physical            -0.4135042  0.0052972 -78.061  < 2e-16 ***
age_receive          0.0091362  0.0005119  17.847  < 2e-16 ***
factor(state_fips)2  0.0184040  0.0267039   0.689   0.4907
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2669 on 10146 degrees of freedom
Multiple R-squared:  0.5939,    Adjusted R-squared:  0.5918
F-statistic: 279.9 on 53 and 10146 DF,  p-value: < 2.2e-16
```

   I think that the relevance assumption holds in this case since the p-values are very low. Intuitively, only the people who get reviewed can lose their benefits so this holds.

   Independence: *This assumption can be tested using t-tests on whether the controls are related to being reviewed are not.*

```
> t.test(physical ~ reviewed, data=part_1)

        Welch Two Sample t-test

data:  physical by reviewed
t = -0.028068, df = 8743.2, p-value = 0.9776
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.02009313  0.01952584
sample estimates:
mean in group 0 mean in group 1
      0.5054748       0.5057584
```

```
> t.test(age_receive ~ reviewed, data=part_1)

        Welch Two Sample t-test

data:  age_receive by reviewed
t = -0.28542, df = 8809.3, p-value = 0.7753
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.2345652  0.1749392
sample estimates:
mean in group 0 mean in group 1
       8.599935        8.629748
```

There is no systemic difference as can be seen from the high p-values and hence I think the independence assumption holds. Intuitively, we do not have a reason to believe that people were systemically picked for review and this can be verified from the t-tests above.

Exclusionary Restriction: *If the only impact of being selected for the lottery is the loss of SSI benefits, then this assumption is valid. Since being reviewed only alters the likelihood of losing SSI and not an individual's behavior regarding criminal charges, it is probable that this assumption is true.*

Monotonicity: *Monotonicity holds if the instrument influences treatment status in the same direction for everyone in the data (ie either increase or decrease but not both).*

Part II [34 points] – Regression Discontinuity

As before, the file `api210_final_part02.dta` contains data on people born between 1976 and 1998 who received SSI as children.

In this scenario, the review of adult SSI eligibility is NOT assigned randomly in each state, as in Part II. The US Congress enacted the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) in 1996, which required all child recipients of SSI to be reviewed for eligibility ***when they turned 18***. The act applied to children with an 18th birthday after August 22, 1996, while all children who had an 18th birthday before that date were automatically enrolled in adult SSI. Accordingly, we observe the number of criminal charges on the criminal record of each person in the data between the ages of 18 and 38.

variable	label
recipient_id	SSI Recipient ID
st_id	State in which the recipient resides at age 18
lost_ssi	Did the recipient lose SSI benefits after age 18?
charges	Total number of criminal charges between ages 18 and 38
bday18_rel	Date recipient turned 18 - August 22, 1996 (in days)
physical	Does the recipient have a physical condition?
age_receive	Age when the recipient first received SSI

1. Write down the regression model you would use to estimate the effect of losing SSI benefits on the number of criminal charges in adulthood. Use a linear functional form. Use all appropriate and valid controls. What is the coefficient of interest? **(4 points)**

   A Regression Discontinuity Design (RDD) will be relevant with the following model:
   *charges = $\beta_0 + \beta_1$*lost_ssi + β_2*bday18_rel + β_3*lost_ssi*bday18_rel + β_4*physical + β_5*age_receive + β_6*state + $\mathcal{E}$*
   Here, β_1 is our coefficient of interest since it shows the added impact on charges after SSI benefits are lost.

2. For the following questions, please answer using both the appropriate notation (hint: potential outcomes) and plain English.

   a. In addition to the usual identifying assumptions for instrumental variables, what is the key identifying assumption in this context? **(4 points)**

      The additional key identifying assumption in this context is that the potential outcome is continuous at the threshold, or in other words, if not for the treatment, the outcome of interest would have had a continuous pattern (no breaks) at the threshold.

b. What type of causal effect do we recover if the assumptions hold? To what populations is this result relevant? **(3 points)**
Similar to the previous question, we recover the LATE if these assumptions hold. The results are relevant to the people born after August 22, 1996 (the threshold at which treatment is applied)

3. Estimate the effect of a child SSI recipient losing their SSI benefits in adulthood on the number of criminal charges they accrue in adulthood by an RD design with the following specifications:
   - Estimate the causal effect of losing SSI in adulthood, rather than the effect of being above or below the birthdate cutoff for SSI eligibility review.
   - Use a local linear regression, with separate slopes for each cutoff side. Do not estimate quadratic or other higher-order terms.
   - No fixed effects are necessary, but include other appropriate and valid controls.
   - Use a bandwidth of 50 days below and above the cutoff.

   Use a triangular kernel. Recall that a triangular kernel with a given bandwidth b and a running variable centered (normalized) at the cutoff $\widetilde{R}_i$ is:

$$w_i = 1(\widetilde{R}_i \in [-b, b]) \cdot \left(1 - \left|\frac{\widetilde{R}_i}{b}\right|\right)$$

   Make sure that you calculate the appropriate standard errors. If your regression includes multiple coefficients, make sure it is clear which one is your final estimate. **(10 points)**

Table 1:

	Dependent variable:
	charges
ssi_lost	14.999***
	(0.373)
bday18_rel	0.013**
	(0.005)
physical	−0.914***
	(0.225)
age_receive	0.994***
	(0.017)
ssi_lost:bday18_rel	0.015
	(0.016)
Constant	6.522***
	(0.229)
Observations	5,143
R^2	0.673
Adjusted R^2	0.673
Residual Std. Error	4.346 (df = 5019)
F Statistic	2,068.473*** (df = 5; 5019)
Note:	*p<0.1; **p<0.05; ***p<0.01

The final estimate is ssi_lost which is equal to 14.999 and statistically significant at the 1% level.

4. Interpret the estimate you provided above. You may assume the identifying assumptions were satisfied. **(3 points)**

This can be interpreted as: Assuming all else constant, individuals that lose benefits at age 18 have 15 additional criminal charges compared to those who do not (between the ages 18 - 38).

5. Do you think the identifying assumption(s) is/are reasonable in this context? Please explain using both theory and/or the empirical tests you would need to carry out if studying this yourself. (You do not need to implement the tests in R) **(10 points)**

In order to check whether the identifying assumptions hold in this context, I would perform the following tests:

- *Density Tests: We can check whether the share of observations across birthdays follows a reasonable distribution (ie there are no systemic biases). Ideally, since date of birth is exogenous, it is safe to assume that this test would pass, and therefore the independence and exclusionary restriction should hold.*
- *Continuity of observable variables / Balance Tests: We can check whether all other observable variables are similar before and after the threshold ie in the treatment and control groups.*
- *Robustness Checks: Finally, we can perform robustness checks by estimating the RD design using different bandwidths or functional forms (such as quadratic or cubic models) to see if the estimated treatment effects are consistent across different specifications.*

Part III [38 points + 3 bonus points] – Difference in Differences

In the final scenario, it is assumed that starting from 2005, the Social Security Administration (SSA) discontinued benefits for 40% of SSI recipients due to budget constraints, after reviewing the medical conditions of all SSI recipients who turned 18 years old that year. The available criminal activity data is limited to the years 2000 to 2009 for beneficiaries who turned 18 years old at the beginning of 2005 and were assessed for eligibility.

In addition, information is available on the medical conditions of the individuals in 1999, as well as data on the socioeconomic status of the individual's household for each year. This is the data that is available for analysis: `api210_final_part03.dta`.

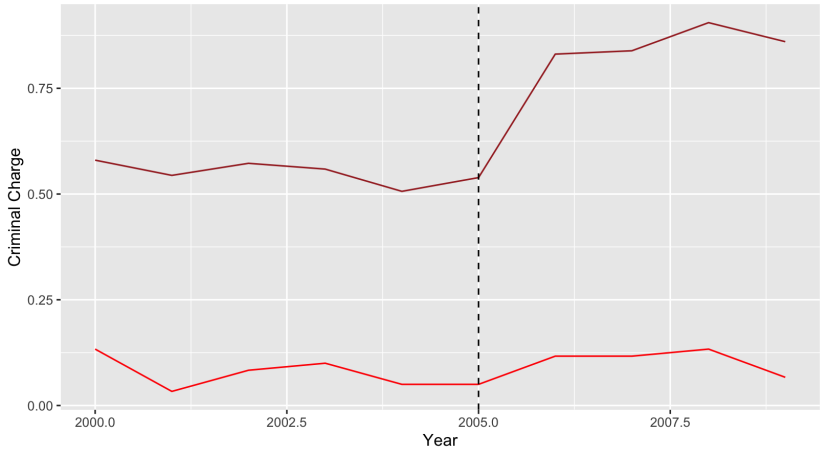
variable	label
recipient_id	SSI Recipient ID
st_id	State in which the recipient resides at age 18
ssi_lost	Did the recipient lose SSI benefits after age 18?
physical	Does the recipient have a physical condition (1999)?
year	Year
hh_index	Household socioeconomic index in each year
crime	Any criminal charges activity in each year

1. Write down the regression model you would use to estimate the <u>effect of losing benefits from the SSI program on the probability of having any criminal charge.</u> Use all appropriate and valid controls, and remember to include them correctly. What is the coefficient of interest? **(4 points)**

   $crime = \beta_0 + \beta_1 * ssi\_lost + \beta_2 * post2004 + \beta_3 * ssi\_lost * post2005 + \beta_4 * year + \beta_5 * physical + \beta_6 * hh\_index + \beta_7 * state + \mathcal{E}$
   Here, β_3 is our coefficient of interest since it shows the impact of losing benefits on the criminal charges likelihood upon turning 18 in 2005.

2. What is the key identifying assumption in this context? Draw and label a simple graph of the raw trends to illustrate the identifying assumption and explain it in plain English. **Note**: you can also use event study estimates. **(4 points)**

   The key identifying assumption in this DiD context is the parallel trends assumption: In the absence of the treatment, the treatment and control groups would have the same trend over time. Simply put, if some people did not lose their SSI benefits, they would have a similar trend as those who did. Therefore, any difference between the treatment and control groups is attributed to the treatment alone.



Here, the red line denotes the trend for the control group and the brown line denotes the trend for the treatment group. The trends are similar till the threshold (2005) beyond which the treatment trend jumps up.

3. What type of causal effect do we recover if the assumptions hold? Explain your response in plain English and use potential outcomes notation to show how this is identified. **(4 points)**

   If the abovementioned assumptions hold, we get the ATT estimate from this specification. The ATT represents the causal effect of losing SSI benefits on the criminal charges for those who got the treatment.
   Mathematically, ATT = $E[Y_i(1) - Y_i(0) \mid ssi\_los = 1]$

4. Estimate the regression model you proposed above and present the coefficient and standard error.

   a. In words, justify your use of fixed effects **(4 points)**

   I used fixed effects for year and state to address any potential biases in the regression. More specifically, year-fixed effects control for any unobservable factors that vary with time and can influence the criminal charge (such as economic policies, etc). Next, state specific fixed effects control for any unobservable factors that vary across states but remain constant over time (such as state-specific policies, etc).

   b. Report the coefficient and standard error. If your regression includes multiple coefficients, be sure your response is clear as to which one is the DID coefficient. **(8 points)**

Table 1:

	Dependent variable:
	crime
ssi_2005	0.701***
	(0.045)
physical	0.016
	(0.021)
hh_index	−0.00001
	(0.007)
Observations	1,000
R^2	0.400
Adjusted R^2	0.387
Residual Std. Error	0.328 (df = 978)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient if interest is ssi_2005 which is equal to 0.701 and is statistically significant at the 1% level in this case.

5. Interpret the estimate you provided above with proper units. You may assume the identifying assumptions were satisfied. **(2 points)**

This coefficient can be interpreted as Holding all else constant, individuals who lost their SSI benefits in 2005 are 0.701 units more likely to have some criminal charges compared to those individuals who did not.

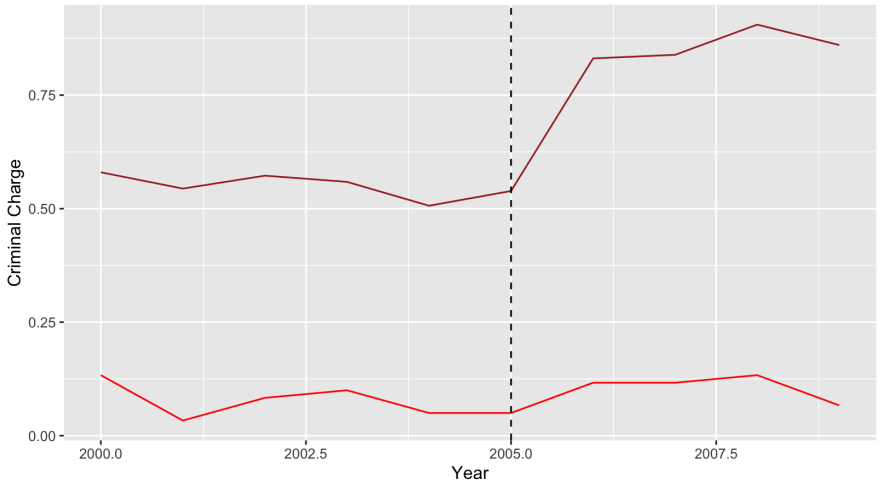
6. Do you think the identifying assumption is reasonable in this context? Please explain using both theory and empirical tests (i.e., in R, as necessary) and show the results of any tests you carry out. **(12 points)**

I think the identifying assumptions for this study hold reasonably, and propose the following two tests to confirm:

- *Balance Test: There is no statistically significant difference between the two groups.*

	estimate	statistic	p.value	parameter	conf.low	conf.high	var
1	0.47	c(t = 17.18)	0	c(df = 378.63)	0.42	0.53	criminal activity
2	-0.03	c(t = -0.57)	0.57	c(df = 421.96)	-0.11	0.06	physical
3	-0.06	c(t = -0.66)	0.51	c(df = 425.65)	-0.24	0.12	socioeconomic index

- *Trend Line: The trends between the treatment and control were roughly parallel before the 2005 threshold.*



7. **Bonus:** Explain why the negative weights problem recently identified in the *difference-in-differences* literature does NOT affect this context. Use plain English. (3 points)

The negative weights problem arises when the treatment and control groups are not fixed and the units switch the status between pre and post-treatment periods. However, in this context, it is not a major concern since the income threshold is unlikely to change during our observation period (that is, eligible individuals are likely to remain so throughout and vice versa).