Economics 142
Final Project - Spring 2020
Due: 3 days after you start

You must submit your project electronically via GRADESCOPE within 3 days (72 hours) of the time you start the exam. Your answer should take the form of a **pdf document** with answers to each question that will include Tables, Figures and a narrative discussion. In an appendix ***you must also submit all code*** that generated the figures tables and calculations. You are allowed to use any auxilliary materials, including course notes and text books, to complete your assignment.

NOTE that part of your grade depends on the presentation of your results (clearly labelled, well-organized, professional-looking tables and figures), and part depends on your narrative discussion of your results.

You MUST report the tables and figures described below as actual tables (and figures). We will not accept an exam that only contains the code and raw output.

The project will use 2 different data sets: one for Part I, and one for Part II. The data sets are described in each section. Before you start the exam, make sure that you have correctly downloaded the data sets and that you can reproduce these means for the variables for each data set.

1 Part I

The data set for this part is called project2020_dd.csv. The data set (which is from Portugal) has 1 record per person for 16,969 observations – 10,575 for men and 6,394 for women. Each person is observed in 6 consecutive years. The timing convention for the measurement of all data is that **period 0** is the 4rd year of observation:

	$year\ l3$	$year\ l2$	$year\ l1$	$year\ 0$	$year\ p1$	$year\ p2$
$time:$	-3	-2	-1	0	$+1$	$+2$
$log\ wage:$	$yl3$	$yl2$	$yl1$	y	$yp1$	$yp2$

So we observe 3 years before year 0 (years l1, l2, and l3, "l" for lagged) and 2 years after year 0 (year p1 and p2, "p" for plus). The data are in "event study" format so year 0 is the first year of each person on a new job. In years -3, -2 and -1 they worked on "job 1" and in years 0, 1 and 2 they worked on "job 2".

The following variables are available and hold information about period 0:
- age (as of year 0)
- educ (=years of education, does not change over time)
- female

- exp = measure of # years since the person completed school, sometimes called "potential experience"
- y = log hourly wage, first year on the second job

The following additional wage information is available for other periods:
- yp1 = wage in year 1
- yp2 = wage in year 2
- yl1 = wage in year -1
- yl2 = wage in year -2
- yl3 = wage in year -3

In addition we observe the mean wages of **other workers** on the first job and the second job:

owage1 = mean log wage of other workers at the first job (held in period -3,-2,-1)

owage2 = mean log wage of other workers at the second job (held in period 0, 1,2)

Summary statistics for this sample are shown at the end of the exam. Make sure you can reproduce these!!

1.1 Table 1 and Figure 1

In this table you will compare the characteristics and wages of male and female workers, focusing on period 0. A suggested format for the table is 4 columns:

column 1 = characteristics for all workers
column 2 = characteristics for female workers
column 3 = characteristics for male workers
column 4 = test statistic comparing females and males (eg t-test)

The main characteristics of interest are:
- education: Note that education takes on only 4 values: 6, 9, 12 and 16

corresponding to elementary schooling; some high school, completed high school and university).
- age: note that the sample excludes people who are over 52 years old
- log of real hourly wage (y)
- mean log real hour wage for co-workers as of period 0 (i.e. owage2)

In addition to comparing means you could distinguish fractions in various intervals. For example, if you find the quartiles of log wages for all workers, you could compare the fractions of men and women in each quartile.

Figure 1

Plot the smoothed histograms of log hourly wages for men and women on one figure (suggestion: the "hist" function in R).

Narrative: Briefly discuss the main differences between men and women, using the table and figure to make your main points.

1.2 Table 2

In this table you will fit a series of standard wage models for wages in period 0, and construct Oaxaca decompositions of the wage gap between men and women.

a) fit 2 models using the pooled data for men and women

- including only a constant and a female dummy
- including a constant, education, a cubic in experience, and a female dummy

b) fit separate models for men and women that include a constant, education, and a cubic in experience. Use the models to construct standard Oaxaca decompositions as in Lecture 7 (recall there are 2 alternatives - construct BOTH - they won't give exactly the same answers).

Narrative: Briefly discuss the decomposition. HINT: you will see that females are better educated than males so the models do not "explain" the gender gap. In fact, they suggest the observed raw wage difference between men and women understates the gender gap. What do you think of this?

1.3 Table 3

Now we are going to examine the effect of a new control variable, which is the mean log wage of a person's co-workers (the variable owage2)

Table 3 will report 4 models:

a) Fit 2 models using the pooled data for men and women

- including only a constant, a female dummy, and owage2
- including a constant, education, a cubic in experience, a female dummy, and owage2

b) fit separate models for men and women that include a constant, education, and a cubic in experience and owage2. Use these models to conduct a **pair of new decompositions** that accounts for the effect of higher-wage coworkers (as above - construct BOTH decompositions)

It will turn out that controlling for owage2 makes quite a difference to the gender gap. You will see that the wage effect of working with highly paid co-workers is quite large, and in the models that allow different returns by gender, the effect is smaller for women than men.

Narrative: In your narrative you will discuss alternative intepretations of the effect of working with highly paid co-workers. Before starting to write your narrative you will want to think carefully about two possible explanations for why people who work with higher-paid co-workers earn more:

- Model 1: getting a job with high-paid co-workers is largely a matter of good luck or connections, and men have better connections, or search harder to find higher paid coworker jobs.

- Model 2: getting a job with highly paid co-workers is only possible for workers who have high levels of cognitive skills or ambition, which is not measured in our data but potentially varies by gender. Think about the implications of these 2 models for the interpretation of the decompositions in Table 3.

1.4 Table 4 and Figure 2

In this part you use the fact that we have job changers in the data to conduct some event studies, and do an analysis of wage changes as people move between jobs with higher and lower paid co-workers.

a) begin by finding the terciles of owage1 (the *terciles* are the bottom, middle and top ***thirds*** of the data, ranked by the variable of interest). Classify all the first jobs (held in periods –3, -2, and -1) into 3 groups based on the tercile of owage1. Then find the terciles of owage2 and classify all the second jobs (held in periods 0, 1 and 2) into 3 groups. (You will notice that on average the second jobs have slightly higher co-worker pay, so the cutoff points for the terciles are a little higher for owage2). Now classify workers into 9 groups based on tercile of owage1 $\times$ tercile of owage2.

Figure 2

Show 9 separate plots of mean wages over time for people who start in each tercile of owage1 and go to each tercile of owage2. The x-axis for each plot will be "event time", which ranges from -3 to +2.

Narrative: Think carefully about the alternative models (Model 1 and Model 2) of why co-worker wages matter. Then discuss the event study graphs. Do these graphs provide more support for Model 1 or Model 2? Also: do you see any pattern of wage movements before a job change that lead you to be concerned?

For Table 4, you will model the change in wages from -1 to 0 (y - yl1) as a function of the change in the mean log wage of co-workers (owage2 - owage1).

a) Fit a set of models for the change in wages using the pooled data for men and women

- including only a constant, a female dummy, and Dwage=owage2-owage1

- including a constant, quadratic in experience as of period -1, a female dummy, and Dwage

b) fit separate models for men and women that include a constant, quadratic in experience as of period -1, and Dowage

*Note: experience in period -1 is just experience in period 0 minus 1.

Narrative: The main issue in this part of the narrative is the comparison between the effect of coworker average wages in OLS models (Table 3) and first-differenced models that control for all unobserved characteristics of people (Table 4).

One way to summarize the two sets of results is to ask: what fraction of the OLS effect of co-worker wages do we see in the first-differenced models? If, for example, the OLS model for males gives a coefficients on coworker wages of 0.66, but the differenced model gives a coefficient of 0.33, then you might conclude that one half of the OLS effect is a causal effect and the other half reflects differences in the unobserved skills of people who tend to work at high-coworker wage jobs.

If the true causal effect of coworker wages for men (from the differenced model) is λ^m and the true causal effect of coworker wages for women (from the differenced model) is λ^f explain how you would modify the decompositions you developed from Table 3 to adjust for the **true** effects of co-worker wages. Carry out this alternative decomposition: what does it imply?

HINT: If you have a regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} ... + e_i$ and you know the true value of β_1 then you can get the correct estimates of the other coefficients by estimating the model:

$$y_i - \beta_1 x_{1i} = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} ... + e_i \tag{1}$$

Based on the estimated effects of co-worker wages from the differenced model, plus the coefficients of the other variables from model (1), you can do a decomposition of the effect of all the various characteristics because

$$\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}_1 + \widehat{\beta}_2 \overline{x}_2 + ..$$

2 Part II

In this part you will use an extended version of the data set from problem set 9. Recall that this data set contains student level data for 112,008 students in Chile who finished high school and were eligible to enter college. In Chile students write a standardized test at the end of high school, called the "PSU" test. Students who score at least 475 points on the PSU test (with family incomes below the 80th percentile) are eligible for a loan from the government for college costs, while students who score less than 475 points cannot receive the loan. The data set is called project2020_rd.csv.

The variables on the data set are:

- psu = PSU test score (ranges from 300 to 700; the scores are numbers like 300.0, 300.5, 301.0, 301.5, 302.0....)
- over475 = 1 if PSU score is 475 or higher
- entercollege = 1 if student entered college
- gpa = high school GPA (scored from 0 to 70, 70 is "perfect")

- privatehs = 1 if student went to a private high school
- hidad = 1 if father has more than a high school education
- himom = 1 if mother has more than a high school education
- female = 1 if female student
- quintile = family income quintile (this has values 1,2,3,4 - families from the top quintile are not eligible for the loan program and are excluded from the data). Note that the shares of the student population in these quintiles is not equal, since the quintile cutoffs are based on all families including (richer) families with no kids.

Summary statistics for this sample are shown at the end of the exam. Make sure you can reproduce these!! (Note that female and quintile were not included in the data for PS9 so you have to download the new version).

2.1 Compliers in an RD Model

In preparation for this part of the exam, review lecture 15, where we discussed compliers, always takers, and never takers in the context of an experiment with incomplete compliance. Also, review Lecture 17 where we discussed a "fuzzy RD" model.

a) The goal of this part of the exam is to develop a proof that we can construct the mean characteristics of the **compliers** for a fuzzy RD design using an extension of the "goofy 2sls" model presented in Lecture 15. Here is our notation.

x_i = running variable (in our case PSU score *re-centered.* So this is PSU-475

D_i = assignment status = 1 if attend college

z_i =1$[x_i \geq 0]$ (indicator for having $x_i \geq 0$)

w_i= vector of characteristics of student i

NOTE: because we are using x as the running variable we have to use w for the characteristics of each person. w plays the role of x in Lecture 15: we are interested in getting the means of w for students who "comply" with the RD - these are the students with scores very close to 475 who will enter college if the get the loan, but will not enter college if they don't.

We will assume we have a local linear f*uzzy RD.* So the assignment probability model (using $Pr()$ to denote the probability) is:

$$\begin{aligned} Pr(D_i = 1|x_i) = E[D_i|x_i] &= a_0 + a_1 x_i,\ x_i < 0 \\ &= b_0 + b_1 x_i,\ x_i \geq 0 \end{aligned}$$

As in Lecture 17, define:

$$\begin{aligned}\pi_1 &= lim_{x\to 0^+} E[D_i|x_i] - lim_{x\to 0^-} E[D_i|x_i] \\ &= b_0 - a_0\end{aligned}$$

This is the jump in college entry we would observe at 475 points (if we had an infinite sample). It represents the fraction of **compliers** who switch from $D_i = 0$ to $D_i = 1$ as $x_i$goes from just below 0 to just above 0. We will call this group $C(0)$ to denote the "compliers when $x \approx 0$". Thus:

$$\pi_1 = Pr(C(0))$$

When $x_i$is just below 0 (i.e., when $x_i \to 0$ and $z_i = 0$) only the **always takers** have $D_i = 1$. We will call this group $AT(0)$ to denote the "always takers when $x \approx 0$". Thus the mean of w_i (the vector oc characteristics) of $AT(0)$ is

$$E[w_i|AT(0)] = E[w_i|D_i = 1, x_i \to 0, z_i = 0] \tag{2}$$

When $x_i$is just above 0 (i.e., when $x_i \to 0$ and $z_i = 1$) both the always takers (group AT) and the compliers have $D_i = 1$. Thus the mean of w_i for the combined group of $AT(0)$ and $C(0)$ is:

$$E[w_i|AT(0) \text{ or } C(0)] = E[w_i|D_i = 1, x_i \to 0, z_i = 1] \tag{3}$$

(i) Use expressions (2) and (3) to prove that:

$$E[w_i|C(0)] = \frac{E[w_i|AT(0) \text{ or } C(0)] \times Pr(AT(0) \text{ or } C(0)) - E[w_i|AT(0)] \times Pr(AT(0)}{Pr(C(0))}$$

Hint: refer to Lecture 15.

(ii) Prove that

$$E[w_i D_i|x_i \to 0, z_i = 1] = E[w_i|AT(0) \text{ or } C(0)] \times Pr(AT(0) \text{ or } C(0))$$

Hint: law of interated expections. Refer to Lecture 15.

(iii) Prove that

$$E[w_i D_i|x_i \to 0, z_i = 0] = E[w_i|AT(0)] \times Pr(AT(0))$$

Hint: law of interated expections. Refer to Lecture 15.

(iv) Consider the "goofy 2sls RD model" with a first stage assignment model and a second stage (structural) model for the the outcome $w_{1i}D_i$ (where w_{1i} is one element of w_i):

$$\begin{aligned}D_i &= \pi_0 + \pi_1 z_i + \pi_2 x_i + \pi_3 x_i z_i + \epsilon_i \; 1st \; stage \\ w_{1i}D_i &= \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 x_i z_i + \nu_i \; structural \; model \\ w_{1i}D_i &= \delta_0 + \delta_1 z_i + \delta_2 x_i + \delta_3 x_i z_i + \nu_i \; reduced \; form\end{aligned}$$

Show that the 2sls estimate of $\beta_1$is an estimate of $E[w_i|C(0)]$.

Hint: first note that $\widehat{\beta}_1 = \widehat{\delta}_1/\widehat{\pi}_1$. Now use the reduced form to take expectations $E[w_i D_i|x_i \to 0, z_i = 1]$ and $E[w_i D_i|x_i \to 0, z_i = 0]$.

2.2 Estimating Characteristics of Compliers

2.2.1 Table 5 and Figures 3,4

In Figure 3 you will show the relationship between PSU and the probability of entering college. Using the data set, select a "bin size" for PSU and graph the mean rate of college entry for all observations in each bin against the mean PSU of scores in the bin. Try using a binsize of 5 points.

In Table 5 you will estimate local linear first stage models for the probability of attending college. These models all have the form

$$D_i = \pi_0 + \pi_1 z_i + \pi_2 x_i + \pi_3 x_i z_i + \epsilon_i \tag{4}$$

where $x_i = PSU_i - 475$, $z_i = 1[PSU_i \geq 475]$, and $D_i = entercollege$. Estimate model (4) using observations with $(475 - B \leq PSU \leq 475 + B)$ for "bandwidths" $B = \{25, 50, 75, 100\}$.

For Figure 4, estimate model (4) using a range of bandwidths. Plot the estimates of π_1, and the $\pm$ 2 standard error confidence bands for each estimate, against the bandwidth choice. (You could try bandwidths of 25, 50, ... 200)

Narrative: Using Figures 3 and 4 and the estimates in Table 5, discuss what you think is a reasonable bandwidth choice. Discuss how a bigger bandwidth may give a more biased but more precise estimate of π_1.

2.2.2 Table 6

Using the approach from Section 2.1, and the bandwidth choice you made in Section 2.2.1, provide estimates of the means of the following 10 characteristics of the compliers to the loan program:

- share with family income in quintile 1
- share with family income in quintile 2
- share with family income in quintile 3
- share with family income in quintile 4
- share female
- share with gpa in the interval $60 \leq gpa \leq 70$
- share with gpa in the interval $50 \leq gpa < 60$
- share with gpa in the interval $gpa < 50$
- share with mother education $>$ HS
- share with father education $>$ HS

In table 6, show

- means of the 10 characteristics for the entire sample
- means of the 10 characteristics for students within your selected bandwidth around the 475 point threshold (so, if your selected bandwidth is 65 points, show the means for students with 410-540 points)
- means of the 10 characteristics for the compliers

- ratio of the mean of each characteristic for the compliers versus the entire sample (e.g., if 50% of the overall sample are female and 60% of the compliers are female then the ratio is 1.20 - these numbers are not actual fractions of females, just examples)

Narrative: Table 6, discuss the claim that the loan program extends college access to more economically disadvantaged students. What else can you say about the compliers?

OVERALL MEANS

Variable	Obs	Mean	Std. Dev.	Min	Max
y	16,969	1.787921	.6452567	.6019443	4.343445
age	16,969	33.5589	5.693736	22	52
educ	16,969	10.48412	3.586129	6	16
female	16,969	.3768048	.4845996	0	1
exp	16,969	17.07478	6.446165	5	30
yl1	16,969	1.736934	.6102851	.6035661	4.237082
yl2	16,969	1.717822	.6004088	.5928036	4.30092
yl3	16,969	1.696991	.590419	.5906132	4.315834
yp1	16,969	1.806854	.6469906	.637028	4.322492
yp2	16,969	1.840162	.6508525	.6963268	4.399618
owage2	16,969	1.692074	.4662639	.7164346	3.804267
owage1	16,969	1.64842	.4720519	.7183212	3.808492

MEANS if female=0

Variable	Obs	Mean	Std. Dev.	Min	Max
y	10,575	1.866448	.6454139	.6019443	4.343445
age	10,575	33.57371	5.695776	22	52
educ	10,575	10.27206	3.558082	6	16
female	10,575	0	0	0	0
exp	10,575	17.30165	6.343187	5	30
yl1	10,575	1.810452	.6119204	.7177568	4.237082
yl2	10,575	1.790024	.6042797	.5928036	4.30092
yl3	10,575	1.768761	.5957848	.5906132	4.315834
yp1	10,575	1.887867	.6449437	.6404113	4.322492
yp2	10,575	1.922324	.6496422	.6963268	4.399618
owage2	10,575	1.724893	.4542091	.8453446	3.804267
owage1	10,575	1.68847	.4640381	.7183212	3.808492

MEANS if female=1

Variable	Obs	Mean	Std. Dev.	Min	Max
y	6,394	1.658045	.6237106	.6978359	4.244642
age	6,394	33.53441	5.690722	22	52
educ	6,394	10.83485	3.605044	6	16
female	6,394	1	0	1	1
exp	6,394	16.69956	6.596358	5	30
yl1	6,394	1.615344	.5877698	.6035661	4.190975

yl2	6,394	1.598407	.5744114	.6697909	4.253769
yl3	6,394	1.57829	.5617018	.7887472	4.055236
yp1	6,394	1.672866	.627873	.637028	4.247511
yp2	6,394	1.704275	.6297967	.7051765	4.268098
owage2	6,394	1.637794	.4806876	.7164346	3.418792
owage1	6,394	1.58218	.4777382	.8171813	3.589496

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
psu	college entry test score	112008	501.9008151	86.1494388	300.0000000	700.0000000
female		112008	0.5685040	0.4952872	0	1.0000000
quintile	family income quintile 1-5	112008	1.9904471	1.1132766	1.0000000	4.0000000
entercollege	1 if enter college	112008	0.4256839	0.4944485	0	1.0000000
privatehs	1 if private HS	112008	0.0271677	0.1625726	0	1.0000000
hidad	1 if dads education > HS	112008	0.1555871	0.3624651	0	1.0000000
himom	1 if moms education > HS	112008	0.1482841	0.3553829	0	1.0000000
gpa	hi school GPA	112008	56.2768552	8.3072761	0	70.0000000
over475	entry test 475	112008	0.6190183	0.4856303	0	1.0000000