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

Econ 142 Final Project

Rowan Pan

5/18/2020

Load packages

```
library(stargazer)
library(matrixStats)
library(lemon)
knit_print.data.frame <- lemon_print
library(dplyr)
library(AER)
library(ggplot2)
```

Load the data

```
# dd dataset
dd <- read.csv("~/Desktop/Econ142/FinalProj/project2020_dd.csv")
head(dd)
```

y	age	educ	female	exp	yl1	yl2	yl3	yp1	yp2	owage2	owage1
2.570021	48	12	0	30	2.620926	2.616735	2.608906	2.784972	2.791555	1.191211	1.122809
2.298948	46	12	1	28	2.296236	2.284491	2.275459	2.309241	2.320770	2.397997	2.485690
2.401098	46	12	1	28	2.347023	2.354368	2.302642	2.409914	2.419748	2.397585	2.485626
1.230435	42	12	0	24	1.232543	1.256243	1.177653	1.198927	1.174227	1.103486	1.016370
2.298948	48	12	0	30	2.296236	2.284491	2.275459	2.309241	2.320770	2.397997	2.485690
2.298948	47	12	0	29	2.296236	2.284491	2.275459	2.309241	2.320770	2.397997	2.485690

Prepare Data

```
# males in dd dataset, 10575 obs
male_dd <- subset(dd, dd$female == 0)

# females in dd dataset, 6394 obs
female_dd <- subset(dd, dd$female == 1)

# calculate column means of these categories
table1_categories <- c("educ", "age", "y", "owage2")


male.means <- colMeans(male_dd[,table1_categories])
female.means <- colMeans(female_dd[,table1_categories])
all.means <- colMeans(dd[,table1_categories])


# get column SD for t-statistic
male_sds <- colSds(as.matrix(male_dd[,table1_categories]))
female_sds <- colSds(as.matrix(female_dd[,table1_categories]))
tstat <- (male.means-female.means)/(sqrt((male_sds^2/10575) + (female_sds^2/6394)))

# create new dataframe
table1 <- data.frame(all.means, female.means, male.means, tstat)
colnames(table1) <- c("All", "Female", "Male", "T statistic")
rownames(table1) <- c("Education", "Age", "Log Wage", "Mean Log Wage of Other Worker")
```

```
# find quartiles of the data
dd$quartiles <- ntile(dd$y, 4)

quartile1 <- subset(dd, dd$quartiles==1)
quartile2 <- subset(dd, dd$quartiles==2)
quartile3 <- subset(dd, dd$quartiles==3)
quartile4 <- subset(dd, dd$quartiles==4)

quartile1_male <- mean(quartile1$female==0)
quartile1_female <- mean(quartile1$female==1)
quartile2_male <- mean(quartile2$female==0)
quartile2_female <- mean(quartile2$female==1)
```

```
quartile3_male <- mean(quartile3$female==0)
quartile3_female <- mean(quartile3$female==1)
quartile4_male <- mean(quartile4$female==0)
quartile4_female <- mean(quartile4$female==1)


quartile1_frac <- c(quartile1_female, quartile1_male)
quartile2_frac <- c(quartile2_female, quartile2_male)
quartile3_frac <- c(quartile3_female, quartile3_male)
quartile4_frac <- c(quartile4_female, quartile4_male)

ttable <- t(data.frame(quartile1_frac, quartile2_frac, quartile3_frac, quartile4_frac))
```

```
# analysis of experience buckets
exp5 <- subset(dd, dd$exp==5)
exp6_13 <- subset(dd, dd$exp >= 6 & dd$exp <= 13)
exp14_20 <- subset(dd, dd$exp >= 14 & dd$exp <= 20)
exp21_30 <- subset(dd, dd$exp >= 21 & dd$exp <= 30)

exp5_male <- mean(exp5$female == 0)
exp5_female <- mean(exp5$female == 1)
exp6_13_male <- mean(exp6_13$female == 0)
exp6_13_female <- mean(exp6_13$female == 1)
exp14_20_male <- mean(exp14_20$female == 0)
exp14_20_female <- mean(exp14_20$female == 1)
exp21_30_male <- mean(exp21_30$female == 0)
exp21_30_female <- mean(exp21_30$female == 1)
```

Table 1

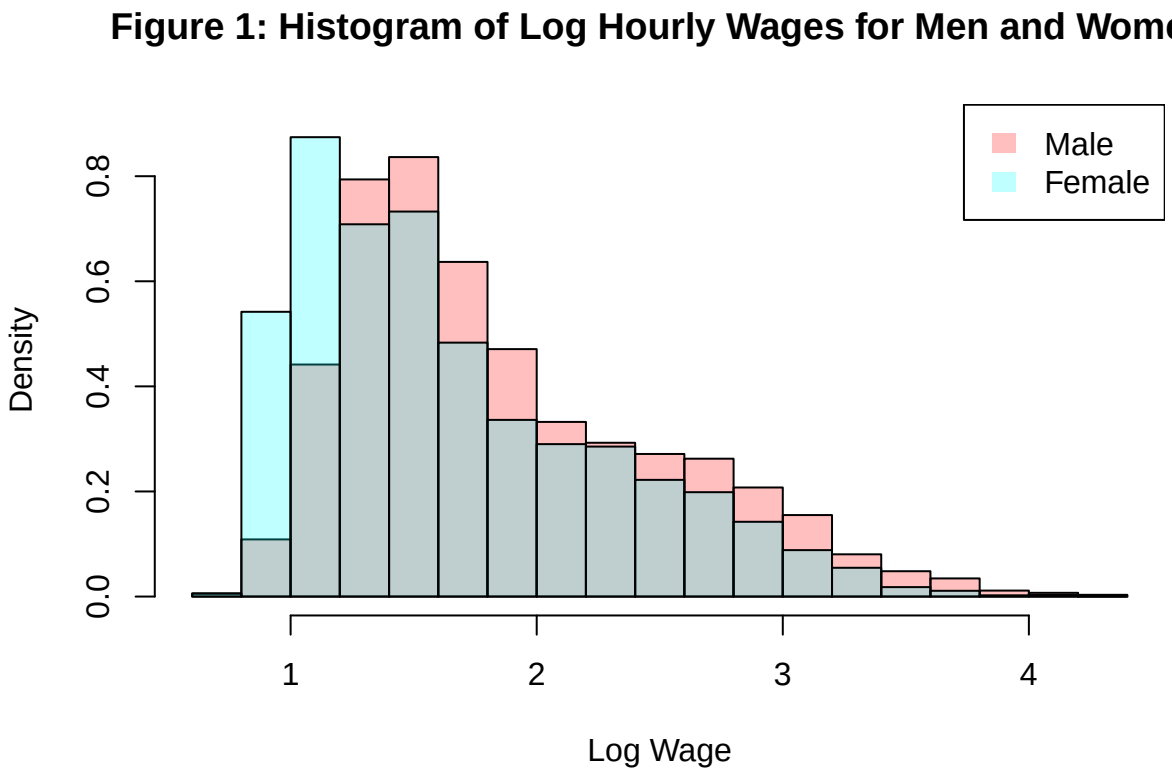
Table 1: Chacteristics of All, Female, and Male Workers with T-Test

	All	Female	Male	T-statistic
Observations (N)	16,969	10,575	6,394	
Education	10.484	10.835	10.272	−9.903
Age	33.559	33.534	33.574	0.436
Log Wage	1.788	1.658	1.866	20.816
Mean Log Wage of Co-Workers	1.692	1.638	1.725	11.676
Wage Quartile 1 Fraction		0.529	0.471	
Wage Quartile 2 Fraction		0.351	0.649	
Wage Quartile 3 Fraction		0.317	0.683	
Wage Quartile 4 Fraction		0.310	0.690	
Fraction with Experience = 5		0.521	0.479	
Fraction with Experience from 6 to 13		0.412	0.588	
Fraction with Experience between 14 and 20		0.348	0.652	
Fraction with Experience between 21 and 30		0.368	0.632	

Figure 1

```
m_hist <- hist(male_dd$y, col=rgb(1,0,0,1/4),
               main = "Figure 1: Histogram of Log Hourly Wages for Men and Women",
               xlab = "Log Wage", freq = F, ylim = c(0,0.9))
f_hist  <- hist(female_dd$y, col=rgb(0,1,1,1/4), freq = F, add = T)

legend('topright',c("Male", "Female"), fill = c(rgb(1,0,0,1/4), rgb(0,1,1,1/4)),
       border = NA)
```



Narrative:

Female workers tend to have a higher level of education (10.835 vs. 10.272) but a lower log wage (1.658 vs 1.866) than male workers, as we can see from Table 1. Furthermore, we can see that female wages are more right skewed and have more mass on the lower end of the wage spectrum, according to figure 1. Although the data contains a lot more observations of male workers than female workers, we can still see that there are progressively high fractions of male workers in higher wage quartiles. The same trend generally holds true for years of experience: the fraction of men increases in general as the years of experience increases. So, in general, men tend to be on average less educated, better paid, and the difference is even bigger for higher wage quartiles and years of experience buckets.

Table 2

(a) Pooled

```
# constant and female dummy
pooled_simple <- lm(y ~ female, data = dd)
# constant, education, cubic in experience, female dummy
pooled_lm1 <- lm(y ~female + educ + exp + I(exp^2) + I(exp^3), data = dd)
```

(b) Separate for Men and Women

```
# men
men_lm1 <- lm(y ~ educ + exp + I(exp^2) + I(exp^3), data = male_dd)

# women
women_lm1 <- lm(y ~ educ + exp + I(exp^2) + I(exp^3), data = female_dd)
```

```
stargazer(pooled_simple, pooled_lm1, men_lm1, women_lm1,
          column.labels = c("Pooled", "Pooled", "Men", "Women") ,
          omit.stat=c("f", "ser"),
          dep.var.labels = c("Log Wage in Period 0"),
          covariate.labels = c("Female", "Education", "Experience", "Experience Squared", "Experience Cu
          header = F,
          title = "Regression Models for Pooled and Male/Female Log Wages")
```

Table 2: Regression Models for Pooled and Male/Female Log Wages

	Dependent variable:			
	Log Wage in Period 0			
	Pooled	Pooled	Men	Women
	(1)	(2)	(3)	(4)
Female	-0.208^{***}	-0.273^{***}		
	(0.010)	(0.007)		
Education		0.149^{***}	0.149^{***}	0.149^{***}
		(0.001)	(0.001)	(0.002)
Experience		0.033^{***}	0.034^{**}	0.043^{***}
		(0.011)	(0.015)	(0.016)
Experience Squared		0.001	0.001	-0.0002
		(0.001)	(0.001)	(0.001)
Experience Cubed		-0.00004^{***}	-0.00005^{***}	-0.00002
		(0.00001)	(0.00002)	(0.00002)
Constant	1.866^{***}	-0.264^{***}	-0.315^{***}	-0.529^{***}
	(0.006)	(0.060)	(0.082)	(0.085)
Observations	16,969	16,969	10,575	6,394
R^2	0.024	0.562	0.532	0.588
Adjusted R^2	0.024	0.562	0.532	0.588
Note:				*p<0.1; **p<0.05; ***p<0.01

Oaxaca Decompositions (2 alternatives): Pooled vs. Separate

$\bar{y}^b - \bar{y}^a$ where b is males and a is females,

$= (\bar{x}^b)'\hat{\beta}^b - (\bar{x}^a)'\hat{\beta}^a$

$= (\bar{x}^b - \bar{x}^a)'\hat{\beta}^a + (\bar{x}^b)(\hat{\beta}^b - \hat{\beta}^a)$

$= (\bar{x}^b - \bar{x}^a)'\hat{\beta}^b + (\bar{x}^a)(\hat{\beta}^b - \hat{\beta}^a)$

First way of Oaxaca Decomposition: Using Pooled Regression

```
copy_male <- male_dd
copy_female <- female_dd

# add new column of cubed exp
copy_male["exp2"] <- copy_male['exp']^2
copy_female["exp2"] <- copy_female['exp']^2
copy_male["exp3"] <- copy_male['exp']^3
copy_female["exp3"] <- copy_female['exp']^3

# pooled betas
betas_pooled <- coef(pooled_lm1)

# x bar and y bar for males and females
xbar_m  <- c(1,0,mean(copy_male$educ, na.rm=TRUE), mean(copy_male$exp, na.rm = TRUE),
             mean(copy_male$exp2, na.rm = TRUE), mean(copy_male$exp3, na.rm = TRUE));
xbar_f  <- c(1,1,mean(copy_female$educ, na.rm=TRUE),  mean(copy_female$exp, na.rm = TRUE),
             mean(copy_female$exp2, na.rm = TRUE), mean(copy_female$exp3, na.rm = TRUE));
ybar_male <- xbar_m %*% betas_pooled
ybar_female <- xbar_f %*% betas_pooled

# log wage gap
oaxaca_pooled <-  ybar_male - ybar_female
```

Second Way of doing Oaxaca Decomposition: Using split male and female regressions

```
# get betas from male and female regression
betas_male <- coef(men_lm1)
betas_female <- coef(women_lm1)
new_cat <- c("educ", "exp", "exp2", "exp3")

# construct right hand side of the second Oaxaca method
rhs1 <- betas_male[1] - betas_female[1];
rhs2 <- (colMeans(copy_male[,new_cat], na.rm=TRUE)
         - colMeans(copy_female[,new_cat], na.rm=TRUE)) %*% betas_male[2:5];
rhs3 <- colMeans(copy_female[,new_cat], na.rm=TRUE) %*% (betas_male[2:5]-betas_female[2:5]);

# log wage gap
oaxaca_separate <- c(rhs3 + rhs2 + rhs1)
```

Oaxaca Decomposition by Method

	Pooled	Separate
Oaxaca Decomposition of Log Wage Gap	0.2084035	0.2084035

Note: The estimates of pooled and separate are roughly the same after rounding.

Narrative:

Both ways of the decomposition give a 0.2084 log point wage gap between men and women. Men and women have roughly the exact same coefficient on the education variable, suggesting that education plays an equal role in raising wages for both male and female workers.

However, women have a much lower constant term than men. The constant for men is statistically signifcant while it isn't for women. This means that there are some unobserved characteristics or phenomenon for men that are not captured by our sample. Examples of those could be connections, ambition, interest, and choice difference between men and women in society. This difference in the constant term may be the factor that understates the gender wage gap.

Table 3

```
pooled_t3_simple <-  lm(y ~ female + owage2, data = dd)
pooled_t3 <- lm(y ~ female + owage2 + educ + exp + I(exp^2) + I(exp^3), data = dd)

men_t3 <- lm(y ~ owage2 + educ+ exp + I(exp^2) + I(exp^3), data = male_dd)
women_t3 <- lm(y ~ owage2 + educ + exp + I(exp^2) + I(exp^3), data = female_dd)

stargazer(pooled_t3_simple, pooled_t3, men_t3, women_t3,
          column.labels = c("Pooled", "Pooled", "Men", "Women") ,
          omit.stat=c("f", "ser"),
          dep.var.labels = c("Log Wage in Period 0"),
          covariate.labels = c("Female", "Other Worker Wage","Education",
                               "Experience", "Experience Squared", "Experience Cubed"),
          header = F,
          title = "Table 3: Regression Models for Pooled and Male/Female Log Wages")
```

Table 3: Regression Models for Pooled and Male/Female Log Wages

	Dependent variable:			
	Log Wage in Period 0			
	Pooled	Pooled	Men	Women
	(1)	(2)	(3)	(4)
Female	-0.121^{***}	-0.192^{***}		
	(0.007)	(0.006)		
Other Worker Wage	1.006^{***}	0.638^{***}	0.662^{***}	0.604^{***}
	(0.007)	(0.007)	(0.009)	(0.011)
Education		0.098^{***}	0.099^{***}	0.094^{***}
		(0.001)	(0.001)	(0.002)
Experience		0.025^{***}	0.022^{*}	0.039^{***}
		(0.009)	(0.012)	(0.013)
Experience Squared		0.001^{*}	0.001^{*}	-0.0003
		(0.001)	(0.001)	(0.001)
Experience Cubed		-0.00003^{***}	-0.00004^{***}	-0.00001
		(0.00001)	(0.00001)	(0.00001)
Constant	0.131^{***}	-0.716^{***}	-0.809^{***}	-0.816^{***}
	(0.013)	(0.049)	(0.068)	(0.070)
Observations	16,969	16,969	10,575	6,394
R^2	0.549	0.703	0.685	0.720
Adjusted R^2	0.548	0.703	0.684	0.720
Note:				*p<0.1; **p<0.05; ***p<0.01

Oaxaca Decomposition, Pooled Way:

```
# pooled betas
betas_pooled_t3 <- coef(pooled_t3)

# x bar and y bar for males and females
xbar_m_t3  <- c(1,0, mean(copy_male$owage2, na.rm = TRUE), mean(copy_male$educ, na.rm=TRUE),
                mean(copy_male$exp, na.rm = TRUE), mean(copy_male$exp2, na.rm = TRUE),
                mean(copy_male$exp3, na.rm = TRUE));
xbar_f_t3  <- c(1,1, mean(copy_female$owage2, na.rm = TRUE), mean(copy_female$educ, na.rm=TRUE),
                mean(copy_female$exp, na.rm = TRUE), mean(copy_female$exp2, na.rm = TRUE),
                mean(copy_female$exp3, na.rm = TRUE));
ybar_male_t3 <- xbar_m_t3 %*% betas_pooled_t3
ybar_female_t3 <- xbar_f_t3 %*% betas_pooled_t3

# log wage gap
oaxaca_pooled_t3 <-  ybar_male_t3 - ybar_female_t3
```

Separate way:

```
# get betas from male and female regression
betas_male_t3 <- coef(men_t3)
betas_female_t3 <- coef(women_t3)
categ_t3 <- c("owage2", "educ", "exp", "exp2", "exp3")

# construct right hand side of the second Oaxaca method
rhs1_t3 <- betas_male_t3[1] - betas_female_t3[1]
rhs2_t3 <- (colMeans(copy_male[,categ_t3], na.rm=TRUE)
            - colMeans(copy_female[,categ_t3], na.rm=TRUE)) %*% betas_male_t3[2:6];
rhs3_t3 <- colMeans(copy_female[,categ_t3], na.rm=TRUE) %*%
            (betas_male_t3[2:6]-betas_female_t3[2:6])

# log wage gap
oaxaca_separate_t3 <- rhs3_t3 + rhs2_t3 + rhs1_t3
```

Comparison of RHS of Separate Oaxaca before and after adding owage2:

	Before owage2	After owage2
rhs1 (constant)	0.2138200	0.0064446
rhs2 (difference in x's)	-0.0618744	0.0199844
rhs3 (difference in betas)	0.0564579	0.1819744
sum (log wage gap)	0.2084035	0.2084035

Oaxaca Decomposition by Method after owage2

	Pooled	Separate
Oaxaca Decomposition Log Wage Gap	0.2084035	0.2084035

Narrative:

As we can see in the pooled log wage gap, log wage gap is still 0.2084 after accounting for owage2. However, in the second Oaxaca method, the components rhs1, rhs2, rhs3 are very different from before accounting for owage2 despite adding up to the same 0.2084.

I agree more with model 2, the idea that getting a job with highly paid co-workers is likely a function of a person possessing high levels of cognitive skills or ambition, and likely varies by gender. This seems plausible because among people who possess high level skills and tremendous ambition, regardless of their gender, are likely to pursue jobs with high pay and enjoy being with other people who are also ambitious and are highly-skilled. The regression model shows that the effect of owage2 is higher for men than women, suggesting that men are more likely to be ambitious or possess traits that make them pursue prestigious jobs with high co-worker pay.

Table of Tercile Cutoffs for owage1 and owage2

	owage1	owage2
0%	0.7183212	0.7164346
33.33333%	1.3938337	1.4419053
66.66667%	1.7197770	1.7807178
100%	3.8084919	3.8042665

```
#create copy of dd
dd_copy <- dd

# find terciles for owage1 and owage2 and divide workers
dd_copy$owage1_tercile <- ntile(dd_copy$owage1, 3)
dd_copy$owage2_tercile <- ntile(dd_copy$owage2, 3)

# separate workers into nine groups based on terciles of owage1 and owage2
dd_copy$nine_groups <- ifelse((dd_copy$owage1_tercile==1 & dd_copy$owage2_tercile==1), 1,
                       ifelse(dd_copy$owage1_tercile==1 & dd_copy$owage2_tercile==2, 2,
                       ifelse(dd_copy$owage1_tercile==1 & dd_copy$owage2_tercile==3, 3,
                       ifelse(dd_copy$owage1_tercile==2 & dd_copy$owage2_tercile==1, 4,
                       ifelse(dd_copy$owage1_tercile==2 & dd_copy$owage2_tercile==2, 5,
                       ifelse(dd_copy$owage1_tercile==2 & dd_copy$owage2_tercile==3, 6,
                       ifelse(dd_copy$owage1_tercile==3 & dd_copy$owage2_tercile==1, 7,
                       ifelse(dd_copy$owage1_tercile==3 & dd_copy$owage2_tercile==2, 8,
                       ifelse(dd_copy$owage1_tercile==3 & dd_copy$owage2_tercile==3, 9, 0)))))))))
```

Table of Classification of 9 Groups of Workers

owage1 tercile	owage2 tercile	final 9 groups
1	1	1
1	2	2
1	3	3
2	1	4
2	2	5
2	3	6
3	1	7
3	2	8
3	3	9

```r
timeframe <- c("yl3", "yl2", "yl1", "y", "yp1", "yp2")

for (i in (1:9)){
  groupi <- subset(dd_copy, dd_copy$nine_groups == i)
  plot(seq(-3,2,1), colMeans(groupi[,timeframe]),
       main=paste("Figure 2: Plot of Mean Wages over Time of Group", seq(1,9,1)[i]),
       xlab = "Event Time", ylab = "Mean Log Wage in Period",  pch = 22, bg = "blue")
}
```

Figure 2: Plot of Mean Wages over Time of Group 1

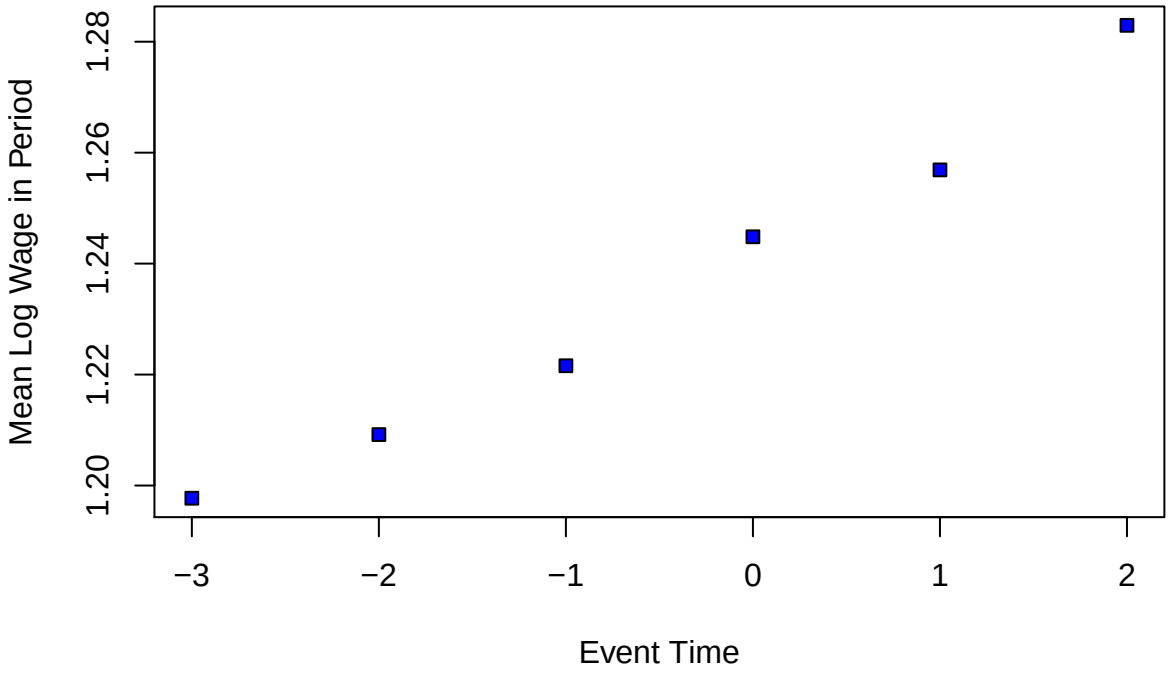


Figure 2: Plot of Mean Wages over Time of Group 2

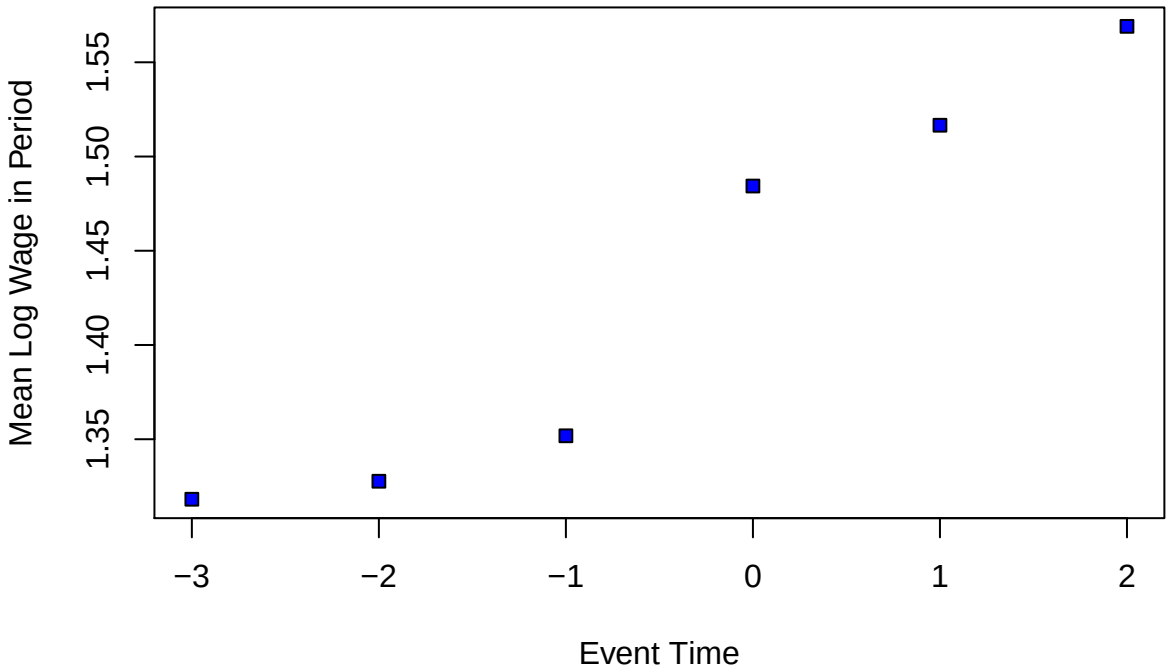


Figure 2: Plot of Mean Wages over Time of Group 3

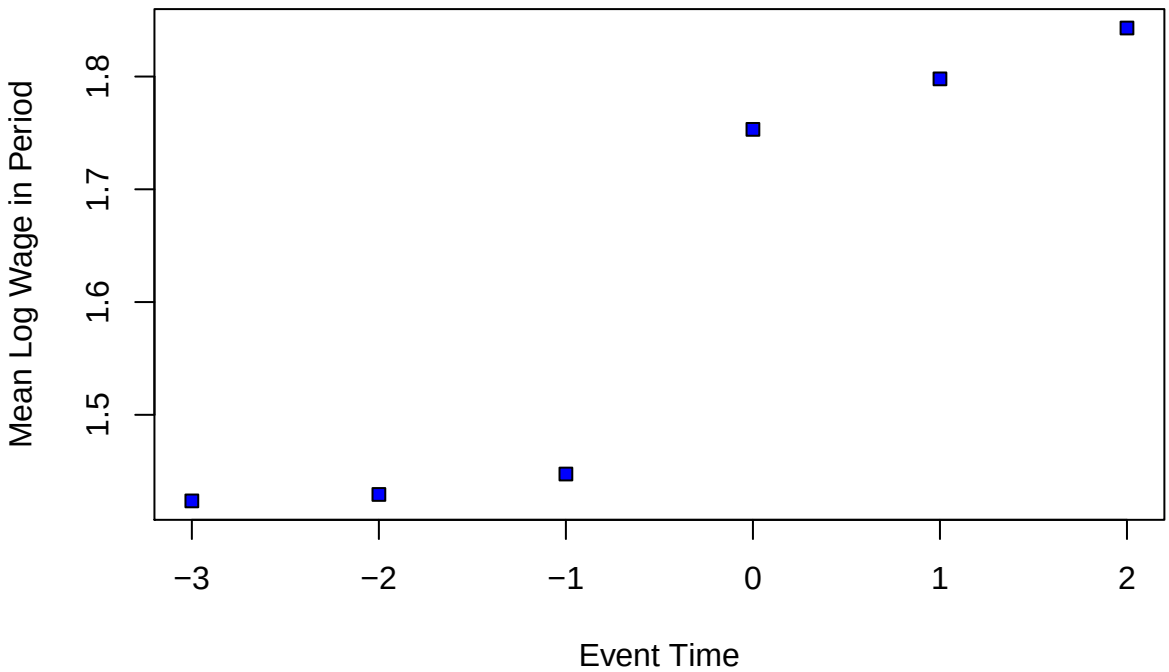


Figure 2: Plot of Mean Wages over Time of Group 4

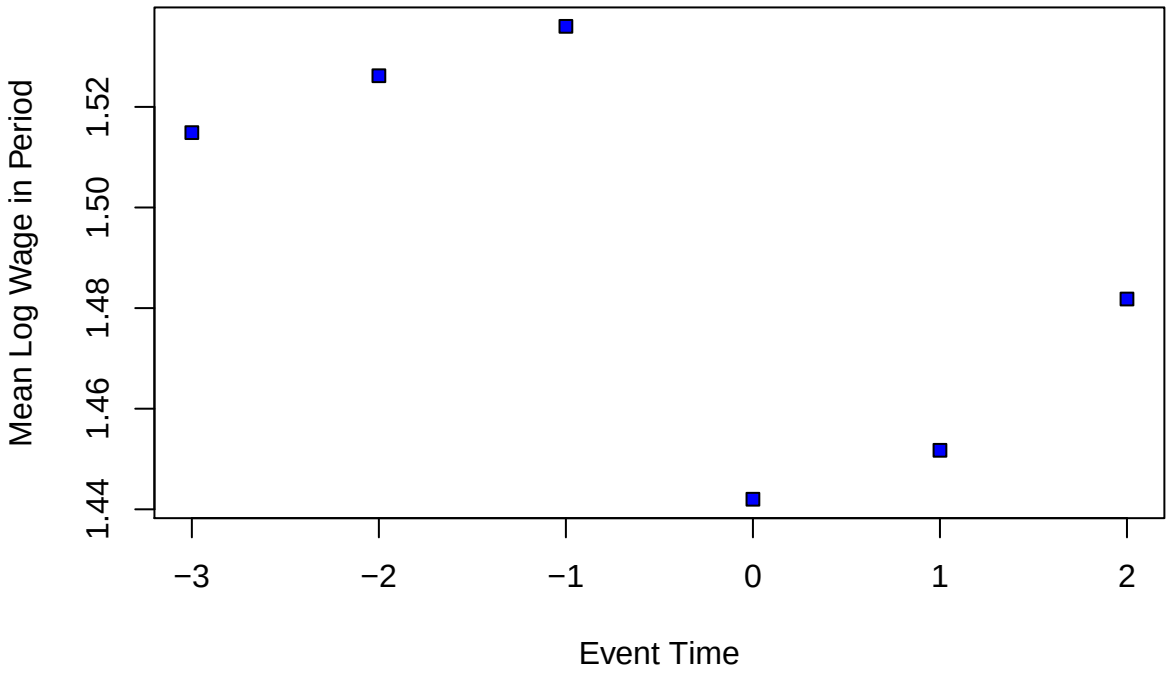


Figure 2: Plot of Mean Wages over Time of Group 5

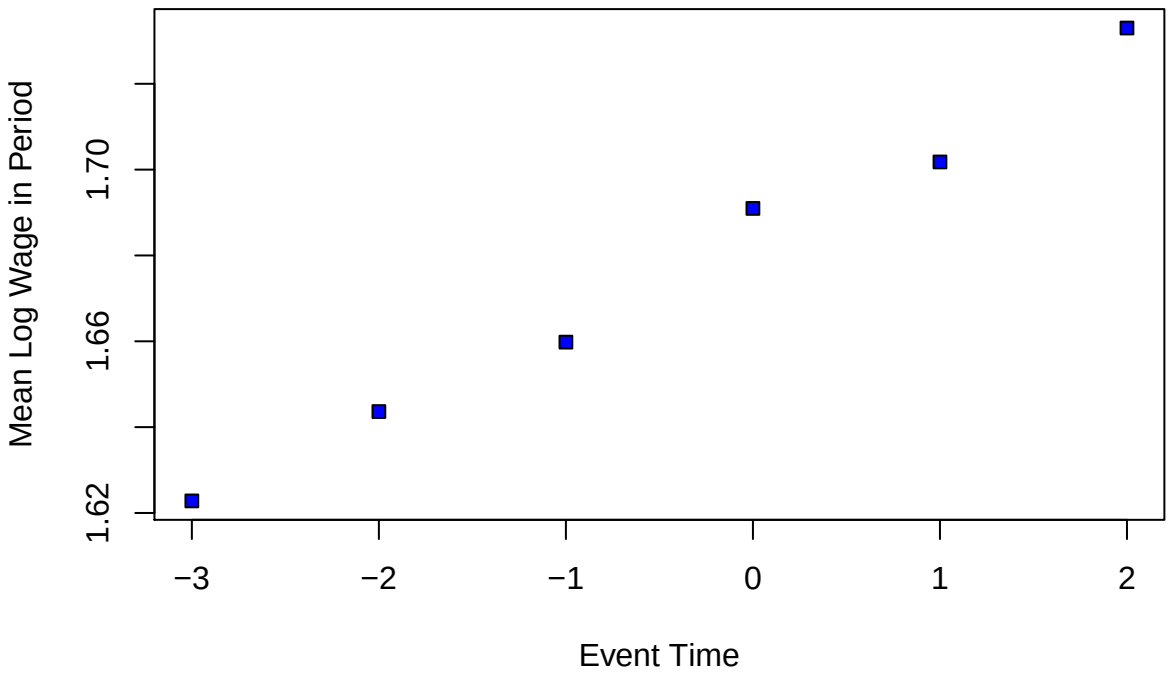


Figure 2: Plot of Mean Wages over Time of Group 6

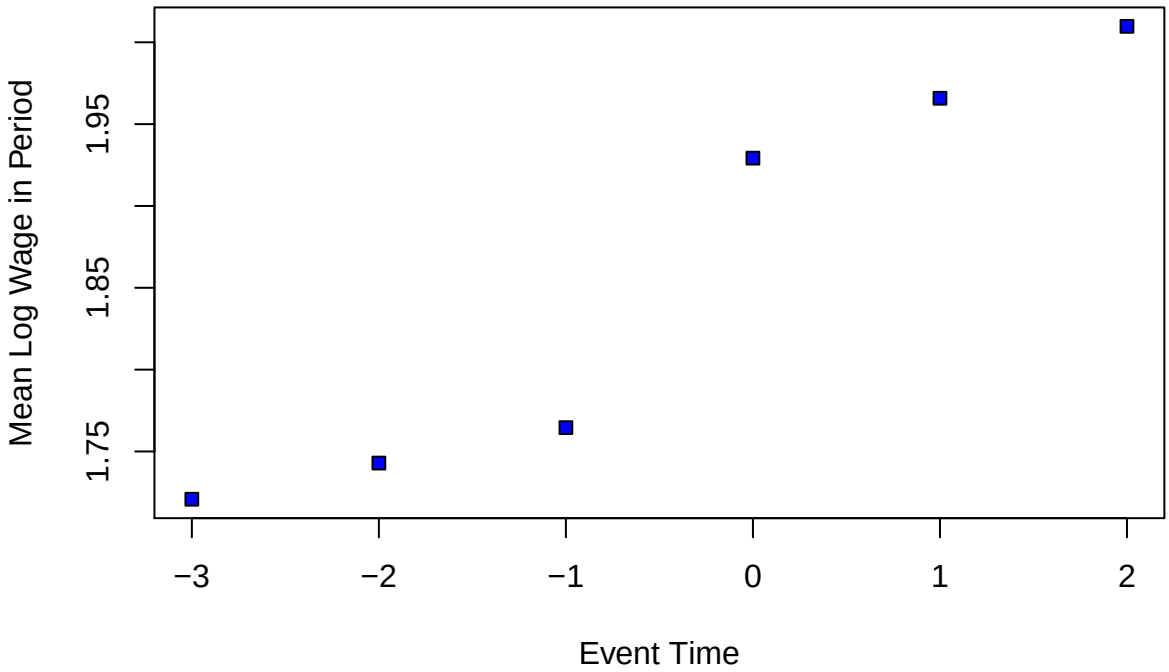


Figure 2: Plot of Mean Wages over Time of Group 7

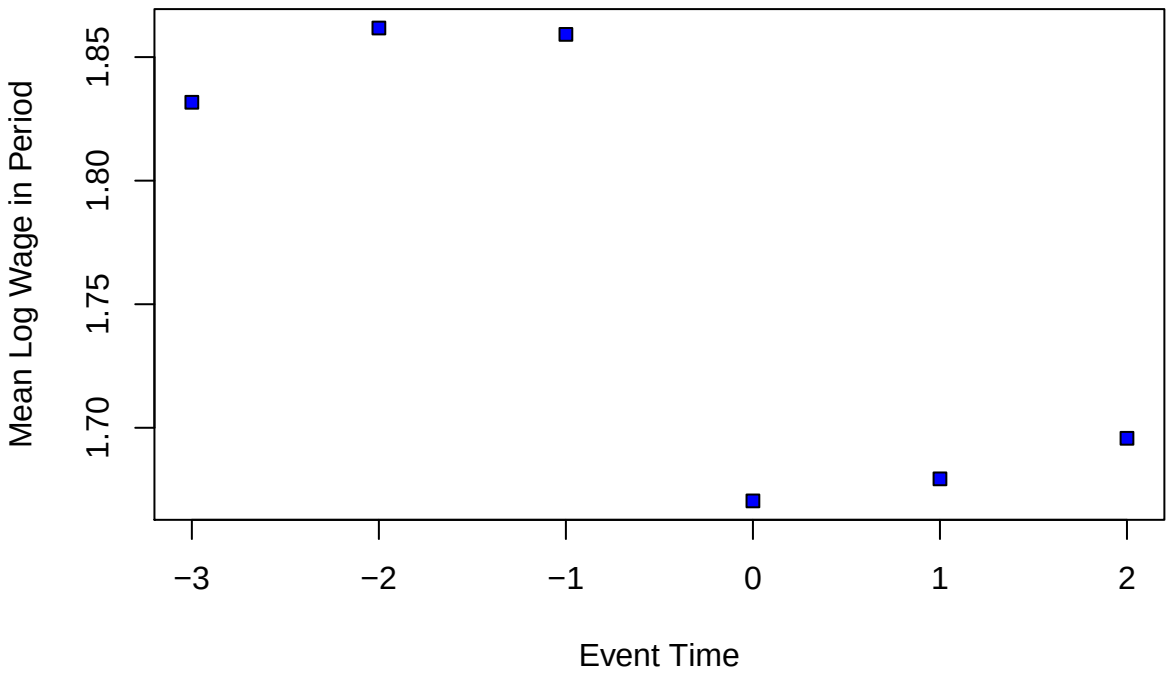


Figure 2: Plot of Mean Wages over Time of Group 8

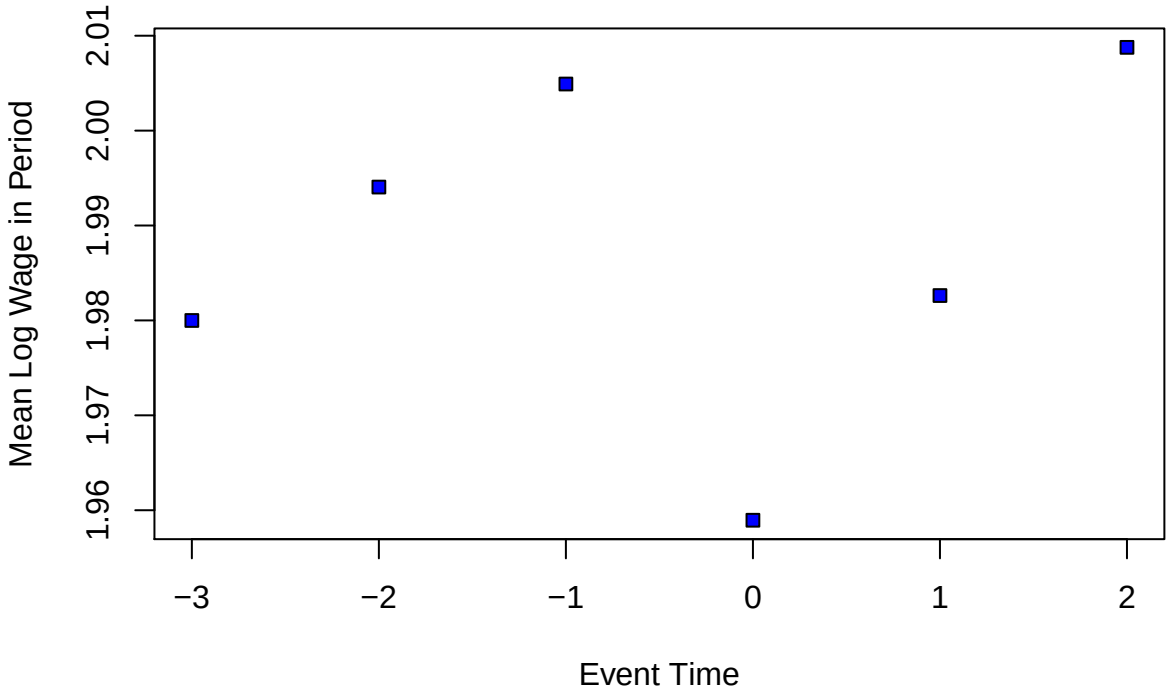
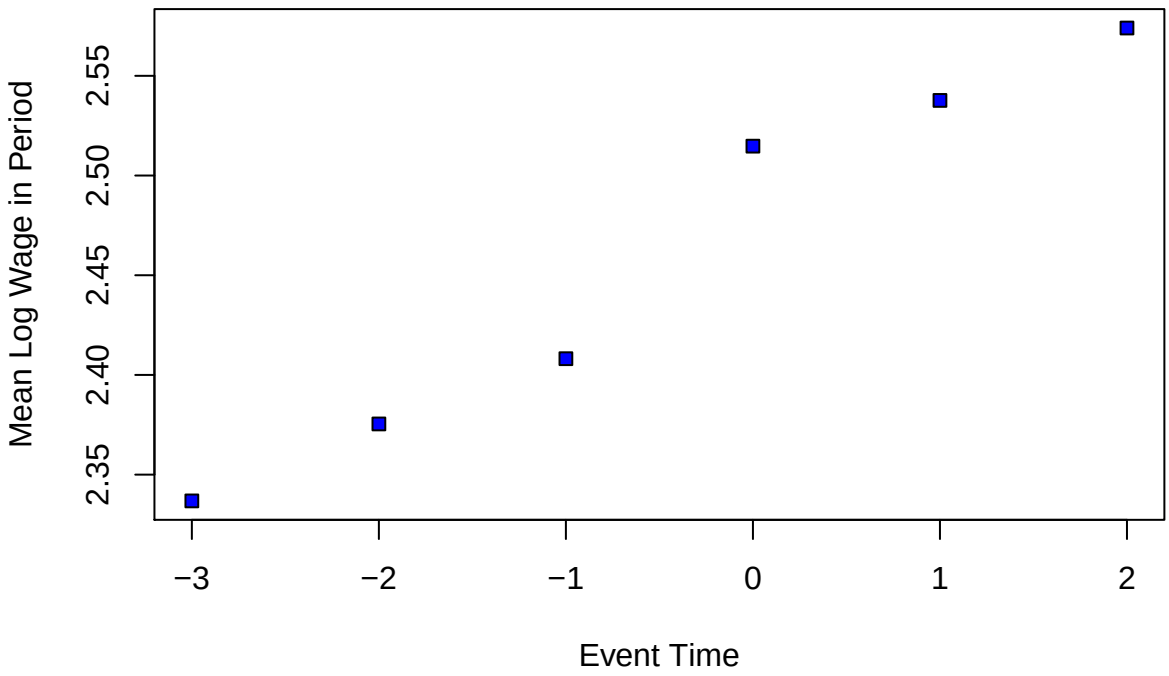


Figure 2: Plot of Mean Wages over Time of Group 9



Narrative:

These graphs provide more support for model 2. In the plots, we see that workers who move from a higher owage2 tercile to a lower owage2 tercile when they switch jobs see their own mean wages diminish along with their new co-workers. Clearly, this is not a consequence of "bad luck" and losing connections even if we assume model 1 to be true. This is likely a result of external factors, some of which could be diminishing ambition (perhaps due to age) or other external factors that cause them to move to a lower paying job (perhaps frictional and structural unemployment related causes). The thing that is concerning in the trends is what prompted these workers to choose a less "prestigious" job with lower owage2, which translates to probably a lower paying job properly. What is concerning to me is why a worker with at least three years of experience in a higher paying job would choose a lower paying job?

Table 4

```
pooled_fig2_simple <- lm(I(y-yl1) ~ female + I(owage2-owage1), data = dd)
pooled_fig2 <- lm(I(y-yl1) ~ female + I(owage2-owage1) + I(exp-1) + I((exp-1)^2), data = dd)

men_fig2 <- lm(I(y-yl1) ~  I(owage2-owage1) + I(exp-1) + I((exp-1)^2), data = male_dd)
women_fig2 <- lm(I(y-yl1) ~  I(owage2-owage1) + I(exp-1) + I((exp-1)^2), data = female_dd)

stargazer(pooled_fig2_simple, pooled_fig2, men_fig2, women_fig2,
          column.labels = c("Pooled", "Pooled", "Men", "Women") ,
          omit.stat=c("f", "ser"),
          dep.var.labels = c("Change in Log Wage from Period -1 to Period 0"),
          header = F,
          covariate.labels = c("Female", "Dwage",
                               "Experience in Period -1", "Experience Squared in Period -1"),
          title = "Regression Models for Pooled and Male/Female Change in Log Wages")
```

Table 4: Regression Models for Pooled and Male/Female Change in Log Wages

	Dependent variable:			
	Change in Log Wage from Period -1 to Period 0			
	Pooled	Pooled	Men	Women
	(1)	(2)	(3)	(4)
Female	-0.019^{***}	-0.022^{***}		
	(0.004)	(0.004)		
Dwage	0.294^{***}	0.290^{***}	0.332^{***}	0.218^{***}
	(0.006)	(0.006)	(0.008)	(0.010)
Experience in Period -1		-0.009^{***}	-0.011^{***}	-0.008^{***}
		(0.002)	(0.002)	(0.002)
Experience Squared in Period -1		0.0002^{***}	0.0002^{***}	0.0001
		(0.00005)	(0.0001)	(0.0001)
Constant	0.045^{***}	0.153^{***}	0.163^{***}	0.123^{***}
	(0.003)	(0.013)	(0.018)	(0.018)
Observations	16,969	16,969	10,575	6,394
R^2	0.110	0.120	0.139	0.088
Adjusted R^2	0.109	0.120	0.139	0.087
Note:				*p<0.1; **p<0.05; ***p<0.01

Alternative Decomposition

```
# alternative decomposition regressions
adjusted_model_male <- lm(I(y-0.332*(owage2)) ~ educ  + owage2 +  exp + I(exp2) + I(exp3), data = copy_m
adjusted_model_female <- lm(I(y-0.218*(owage2))~ educ+ owage2 + exp + I(exp2) + I(exp3), data = copy_fem

stargazer(adjusted_model_male, adjusted_model_female, summary = F, header = F,
          title = "Alternative Decomposition of Causal Effect of Coworker Wages",
          covariate.labels = c("Education", "owage2", "Experience", "Experience Squared", "Experience Cu
          column.labels = c("Male", "Female"),
          dep.var.labels = c("y-0.332*dwage", "y- 0.218*dwage"))
```

```
stargazer(adjusted_model_male, adjusted_model_female, header = F, summary = F)
```

Table 8

	Dependent variable:	
	I(y - 0.332 ∗(owage2))	I(y - 0.218 ∗(owage2))
	(1)	(2)
educ	0.099***	0.094***
	(0.001)	(0.002)
owage2	0.330***	0.386***
	(0.009)	(0.011)
exp	0.022*	0.039***
	(0.012)	(0.013)
I(exp2)	0.001*	−0.0003
	(0.001)	(0.001)
I(exp3)	−0.00004***	−0.00001
	(0.00001)	(0.00001)
Constant	−0.809***	−0.816***
	(0.068)	(0.070)
Observations	10,575	6,394
R^2	0.562	0.638
Adjusted R^2	0.561	0.638
Residual Std. Error	0.363 (df = 10569)	0.330 (df = 6388)
F Statistic	2,708.813*** (df = 5; 10569)	2,250.413*** (df = 5; 6388)
Note:		*p<0.1; **p<0.05; ***p<0.01

Narrative:

Looking at the coefficients on owage2 and owage2 - owage 1 in table 3 and 4, respectively,

Table 3's owage2

- Pooled: 0.638, men = 0.662, women = 0.604

Table 4's owage-owage1

- Pooled 0.290, men = 0.332, women = 0.218

Observations

- For the pooled model, the OLS model represents 0.290/0.638 = 45% of the causal effect and the 55% represents unobserved factors.
- For men, the OLS model represents 0.332/0.662 = 50% of the causal effect, and the other half represents unobserved factors.
- For women, the OLS model represents 0.218/0.614 = 35% of the causal effect, and the other 65% represents unobserved factors.

The difference between men and women in the causal effect of the OLS estimate indicates that owage2 (effect of moving to a new job) has a stronger effect on increasing log wage for men than for women. Likw mentioned in model 2 earlier, this could be due to multiple factors like ambition, interests, connections, and also sometimes luck.

To get a better estimate of the true causal effect of coworker wages for men and women, we can develop a decomposition where we used the coefficient on dwage (owage2-owage1) from table 4 and regard it as the true value of beta 1 (using the hint in model 1). We use perform a decomposition as follows: regress the exogenous variable $\beta_{dwage} \times (y - owage2)$ on the regressors from table 3 (education, experience, experience squared, experience cubed).

The model will look like this:

$\beta_{dwage}(y - owage2) = \beta_0 + \beta_2 educ + \beta_3 exp + \beta_4 exp^2 + \beta_5 exp^3 + e_i$

Differencing out using the coefficient dwage will remove the differences in the unobserved skills of people who tend to work at high-coworker wage jobs and only keep the causal effect of owage2.And the difference between the coefficients of owage2 between men and women lessened, meaning the impact of co-worker pay and its impact on wages is reduced between men and women. This, in short, means that men and women are more similar in terms of owage2 than we had predicted earlier in table3's OLS.

And since we know $\bar{y} = \beta_0 + \beta_{dwage} \bar{owage2} + \beta_2 \bar{educ} + \beta_3 \bar{exp} + \beta_4 \bar{exp^2} + \beta_5 \bar{exp^3}$, we can get the correct estimates for the other variables and perform a decomposition for men and women.

Part II (Chile PSU Data)

2.1 Compliers in an RD Model

(i)

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

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

We know that $E[w_i|AT(0) \text{ or } C(0)]$ is a weighted average of always-takers and compliers given in this expression:

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

Rearranging,

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

(ii)

Law of Iterated Expections: $E[Y] = E[E[Y|X]]$

$E[w_i D_i|x_i \to 0, z_i = 1]$

Since $w_i D_i$ has two outcomes:

$$w_i D_i = \begin{cases} w_i & \text{if } D_i = 1 \\ 0 & \text{if } D_i = 0 \end{cases}$$

Therefore, $E[w_i D_i|x_i \to 0, z_i = 1] = E[E[w_i D_i|x_i \to 0, z_i = 1] \mid D_i]$

$= E[w_i|D_i = 1, z_i = 1] \times Pr(D_i = 1|z_i = 1)$

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

(iii)

$E[w_i D_i|x_i \to 0, z_i = 0] =$

By similar reasoning as (ii):

$= E[E[w_i D_i|x_i \to 0, z_i = 0] \mid D_i]$

$= E[w_i \mid D_i = 1, z_i = 0] \times Pr(D_i = 1|z_i = 0)$

$= E[w_i|AT(0)] \times Pr(AT(0))$

(iv)

We know that $\hat{\beta_1} = \frac{\hat{\delta_1}}{\hat{\pi_1}} = \frac{E[w_i D_i | x_i \to 0 \ , \ z_i = 1] - E[w_i D_i | x_i \to 0 \ , \ z_i = 0]}{E[D_i | x_i \to 0 \ , \ z_i = 1] - E[D_i | x_i \to 0 \ , \ z_i = 0]}$

The numerator is the difference between what we proved in (iii) and (ii):

Numerator $\hat{\delta_1}$ is: $E[w_i | AT(0) \ or \ C(0)] \times Pr(AT(0) \ or \ C(0)) - E[w_i | AT(0)] \times Pr(AT(0))$

which is equal to the numerator of our proof in part (i) for $E[w_i | C(0)]$

Denominator $\hat{\pi_1}$ is: $E[D_i | x_i \to 0 \ , \ z_i = 1] - E[D_i | x_i \to 0 \ , \ z_i = 0]$

$= Pr(AT(0) \ or \ C(0)) - Pr(AT(0)) = Pr(C(0))$ (from lecture 15)

Putting the numerator and denominator together, we have proved that the $\hat{\beta_1}$ in the "goofy regression" is equivalent to $E[w_i | C(0)]$, the mean characteristic of the compliers.

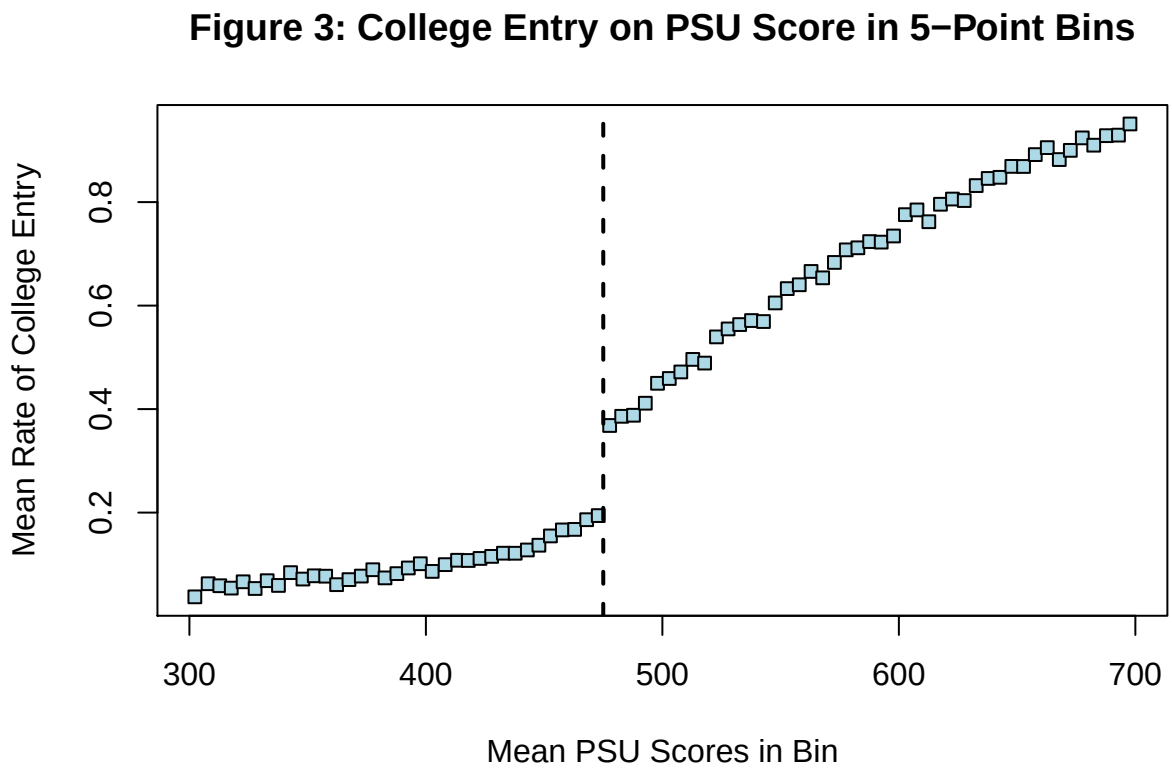
```r
# rd dataset
rd <- read.csv("~/Desktop/Econ142/FinalProj/project2020_rd.csv")
head(rd)
```

psu	female	quintile	entercollege	privatehs	hidad	himom	gpa	over475
396.0	1	1	0	0	0	0	60	0
402.5	1	1	0	0	0	0	65	0
485.0	1	3	0	0	0	0	55	1
461.5	0	2	0	0	0	0	0	0
394.0	1	1	0	0	0	0	62	0
409.0	1	1	0	0	0	0	57	0

Figure 3

```r
# bin data using aggregate function
binned_data <- aggregate(rd, list(cut(rd$psu, breaks= (max(rd$psu)-min(rd$psu))/5)), mean)

plot(binned_data$psu, binned_data$entercollege,
      xlab = "Mean PSU Scores in Bin",
      ylab = "Mean Rate of College Entry",
      main = "Figure 3: College Entry on PSU Score in 5-Point Bins",
      pch = 22, bg = "lightblue")
abline(v=475, lty = 2, lwd = 2)
```



```r
bandwidths <- c(25,50,75,100)

bandwidth25 <- rd[(rd$psu >= 475-bandwidths[1]) & (rd$psu <= 475+bandwidths[1]),]
entercollege_25 <- lm(entercollege ~ I(psu-475) + over475 + I((psu-475)*over475), data = bandwidth25)

bandwidth50 <- rd[(rd$psu >= 475-bandwidths[2]) & (rd$psu <= 475+bandwidths[2]),]
entercollege_50 <- lm(entercollege ~ I(psu-475) + over475 + I((psu-475)*over475), data = bandwidth50)

bandwidth75 <- rd[(rd$psu >= 475-bandwidths[3]) & (rd$psu <= 475+bandwidths[3]),]
entercollege_75 <- lm(entercollege ~ I(psu-475) + over475 + I((psu-475)*over475), data = bandwidth75)

bandwidth100 <- rd[(rd$psu >= 475-bandwidths[4]) & (rd$psu <= 475+bandwidths[4]),]
entercollege_100 <- lm(entercollege ~ I(psu-475) + over475 + I((psu-475)*over475), data = bandwidth100)

stargazer(entercollege_25, entercollege_50, entercollege_75, entercollege_100,
          header = F,
          column.labels = c('B = 25', 'B = 50', 'B = 75', 'B = 100') ,
          omit.stat=c("f", "ser"),
          dep.var.labels = c("Entercollege"),
          covariate.labels = c("PSU-475", "over475", "(PSU-475)*over475"),
          title = "Table 5: Regression Models for Probability of Attending College ")
```

Table 5

Table 5: Regression Models for Probability of Attending College

	Dependent variable:			
	Entercollege			
	B = 25	B = 50	B = 75	B = 100
	(1)	(2)	(3)	(4)
PSU-475	0.001	0.002***	0.001***	0.001***
	(0.001)	(0.0002)	(0.0001)	(0.0001)
over475	0.170***	0.165***	0.177***	0.184***
	(0.012)	(0.008)	(0.007)	(0.006)
(PSU-475)*over475	0.003***	0.002***	0.002***	0.002***
	(0.001)	(0.0003)	(0.0002)	(0.0001)
Constant	0.181***	0.188***	0.182***	0.176***
	(0.009)	(0.006)	(0.005)	(0.004)
Observations	22,846	43,852	63,070	79,523
R^2	0.066	0.110	0.152	0.194
Adjusted R^2	0.066	0.110	0.152	0.194
Note:				*p<0.1; **p<0.05; ***p<0.01

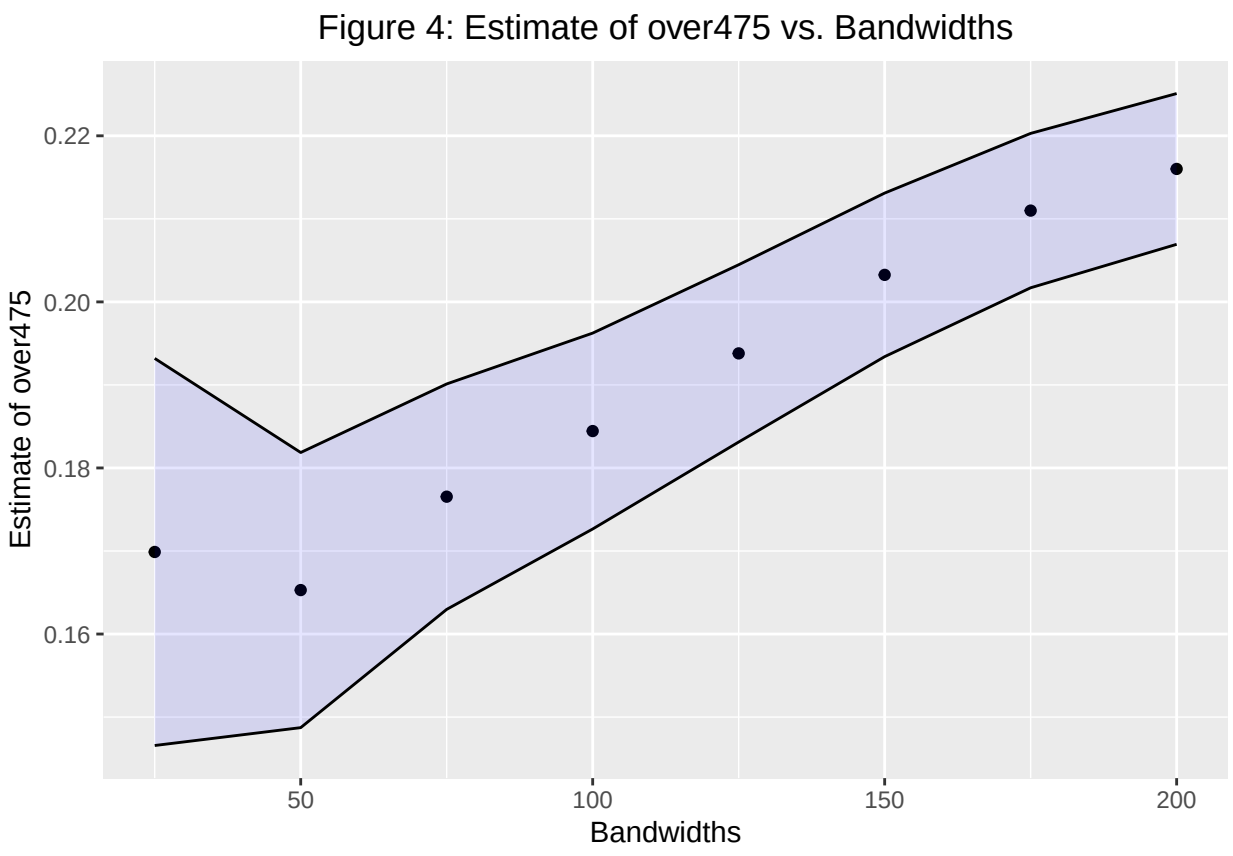
```r
# for figure 4: plot range of bandwidths from 25 to 200
bandwidth_fig4 <- seq(25,200,25)
estimates <- c()
h.ci <-  c()
l.ci <-  c()

for (b_size in bandwidth_fig4) {
  bandwidth_sample <- rd[(rd$psu >= 475 - b_size) & (rd$psu <= 475 + b_size),]
  model <- lm(entercollege ~ I(psu-475) + over475 + I((psu-475)*over475), data = bandwidth_sample)
  pi1 <-  model$coefficients[3]
  estimates <-  append(estimates, pi1)
  sd = summary(model)$coefficients[3,2]
  l.ci = c(l.ci, pi1 - 2*sd)
  h.ci = c(h.ci, pi1 + 2*sd)
}
```

Figure 4:

```r
band_and_est_jumps = data.frame(bandwidth_fig4, estimates)

ggplot(data = band_and_est_jumps, aes(x = bandwidth_fig4, y = estimates)) +
geom_point() + geom_ribbon(aes(ymax = h.ci, ymin = l.ci), fill = 'blue', alpha = 0.1) +
  labs(title = "Figure 4: Estimate of over475 vs. Bandwidths") +
  geom_line(aes(x = bandwidth_fig4, y = l.ci)) + geom_line(aes(x = bandwidth_fig4, y = h.ci)) +
    xlab("Bandwidths") + ylab("Estimate of over475")+
    theme(plot.title = element_text(hjust = 0.5))
```



Narrative:

I think a good choice of bandwidth is one that balances the competing priorities of accuracy (the existence of bias) and precision. A higher bandwidth is more precise because a larger sample size would tend the sample error to go down. We know that the standard error is inversely related to the square root of the sample size.

However, in our Regression discontinuity context (although not always), the bias will increase as a tradeoff of a large bandwidth and higher precision. Choosing a bandwidth is the choice of selecting how far to go to the left and the right when we are right near the boundary where the jump occurs. Observations immediately to the left and right of the boundary are assumed to be very similar to each other. However, the further you go in either direction – as the bandwidth gets larger – we encounter further observations where that have genuine differences from the closest ones to the jump. As a result, our assumption about the similarity of the bandwidth sample on either side no longer holds. As a result, the bias increases as we increase the bandwidth.

By figure 4, the choice of bandwidth I would choose is 100 because this bandwidth choice seems to be the point of inflection of the function before the function switches from convex to concave. In economics, this can be seen as diminishing marginal returns to the estimate of over475 (π_1) as we go over bandwidth of 100. To me, it is the most ideal choice because it strikes a good balance between precision and bias.

```r
# find characteristic means for entire sample
quintile1 <- subset(rd, rd$quintile == 1)
quintile2 <- subset(rd, rd$quintile == 2)
quintile3 <- subset(rd, rd$quintile == 3)
quintile4 <- subset(rd, rd$quintile == 4)

share_q1_all <- nrow(quintile1) / nrow(rd)
share_q2_all <- nrow(quintile2) / nrow(rd)
share_q3_all <- nrow(quintile3) / nrow(rd)
share_q4_all <- nrow(quintile4) / nrow(rd)

share_female_all <- nrow(subset(rd, rd$female == 1)) / nrow(rd)
share_gpa_60_70_all <- nrow(subset(rd, rd$gpa >= 60 & rd$gpa <= 70)) / nrow(rd)
share_gpa_50_60_all <- nrow(subset(rd, rd$gpa >= 50 & rd$gpa < 60)) / nrow(rd)
share_gpa_less50_all <- nrow(subset(rd, rd$gpa < 50)) / nrow(rd)

share_mother_overhs_all <- nrow(subset(rd, rd$himom == 1)) / nrow(rd)
share_father_overhs_all <- nrow(subset(rd, rd$hidad == 1)) / nrow(rd)

entire_sample <- c(share_q1_all, share_q2_all, share_q3_all, share_q4_all,
                   share_female_all, share_gpa_60_70_all, share_gpa_50_60_all, share_gpa_less50_all,
                   share_mother_overhs_all, share_father_overhs_all)



# find characteristic means for my chosen bandwidth
b <- 100

bdata <- rd[(rd$psu >= 475-b) & (rd$psu <= 475+b),]

share_q1_b <- nrow(subset(bdata, bdata$quintile == 1)) / nrow(bdata)
share_q2_b <- nrow(subset(bdata, bdata$quintile == 2)) / nrow(bdata)
share_q3_b <- nrow(subset(bdata, bdata$quintile == 3)) / nrow(bdata)
share_q4_b <- nrow(subset(bdata, bdata$quintile == 4)) / nrow(bdata)


share_female_b <- nrow(subset(bdata, bdata$female == 1)) / nrow(bdata)
share_gpa_60_70_b <- nrow(subset(bdata, bdata$gpa >= 60 & bdata$gpa <= 70)) / nrow(bdata)
share_gpa_50_60_b <- nrow(subset(bdata, bdata$gpa >= 50 & bdata$gpa < 60)) / nrow(bdata)
share_gpa_less50_b <- nrow(subset(bdata, bdata$gpa < 50)) / nrow(bdata)

share_mother_overhs_b <- nrow(subset(bdata, bdata$himom == 1)) / nrow(bdata)
share_father_overhs_b <- nrow(subset(bdata, bdata$hidad == 1)) / nrow(bdata)

b_sample <- c(share_q1_b, share_q2_b, share_q3_b, share_q4_b, share_female_b,
              share_gpa_60_70_b, share_gpa_50_60_b, share_gpa_less50_b,
              share_mother_overhs_b, share_father_overhs_b)
```

```
# find characteristic means for compliers

rd_copy <- rd
rd_copy$quintile1 <- ifelse(rd_copy$quintile==1, 1, 0)
rd_copy$quintile2 <- ifelse(rd_copy$quintile==2, 1, 0)
rd_copy$quintile3 <- ifelse(rd_copy$quintile==3, 1, 0)
rd_copy$quintile4 <- ifelse(rd_copy$quintile==4, 1, 0)

rd_copy$gpa_60_70 <- ifelse((rd_copy$gpa >= 60 & rd_copy$gpa <= 70), 1, 0)
rd_copy$gpa_50_60 <- ifelse((rd_copy$gpa >= 50 & rd_copy$gpa < 60), 1, 0)
rd_copy$gpa_less50 <- ifelse(rd_copy$gpa < 50, 1, 0)

share_q1_complier <- ivreg(I(quintile1*entercollege) ~ entercollege |
                           over475 + I(psu-475) + I(over475*(psu-475)), data = rd_copy)

share_q2_complier <- ivreg(I(quintile2*entercollege) ~ entercollege |
                           over475 + I(psu-475) + I(over475*(psu-475)), data = rd_copy)

share_q3_complier <- ivreg(I(quintile3*entercollege) ~ entercollege |
                           over475 + I(psu-475) + I(over475*(psu-475)), data = rd_copy)

share_q4_complier <- ivreg(I(quintile4*entercollege) ~ entercollege |
                           over475 + I(psu-475) + I(over475*(psu-475)), data = rd_copy)


share_female_complier <-  ivreg(I(female*entercollege) ~ entercollege |
                              over475 + I(psu-475) + I(over475*(psu-475)),data = rd_copy)

share_gpa_60_70_complier <- ivreg(I(gpa_60_70*entercollege) ~ entercollege |
                              over475 + I(psu-475) + I(over475*(psu-475)),data = rd_copy)

share_gpa_50_60_complier <- ivreg(I(gpa_50_60*entercollege) ~ entercollege |
                              over475 + I(psu-475) + I(over475*(psu-475)),data = rd_copy)

share_gpa_less50_complier <- ivreg(I(gpa_less50*entercollege) ~ entercollege |
                              over475 + I(psu-475) + I(over475*(psu-475)),data = rd_copy)

share_himom_complier <-  ivreg(I(himom*entercollege) ~ entercollege |
                              over475 + I(psu-475) + I(over475*(psu-475)),data = rd_copy)


share_hidad_complier <- ivreg(I(hidad*entercollege) ~ entercollege |
                              over475 + I(psu-475) + I(over475*(psu-475)),data = rd_copy)


complier_sample <- c(share_q1_complier$coefficients[2], share_q2_complier$coefficients[2],
                     share_q3_complier$coefficients[2], share_q4_complier$coefficients[2],
                     share_female_complier$coefficients[2], share_gpa_60_70_complier$coefficients[2],
                     share_gpa_50_60_complier$coefficients[2], share_gpa_less50_complier$coefficients[2]
                     share_himom_complier$coefficients[2], share_hidad_complier$coefficients[2])
```

```r
table6 <- data.frame(entire_sample, b_sample, complier_sample, complier_sample/entire_sample)
rownames(table6) <- c("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 60 <= GPA <= 70",
                      "Share with 50 <= GPA <  60", "Share GPA <  50",
                      "Share with Mother Education > HS", "Share with Father Education > HS")
colnames(table6) <- c("Entire Sample", "Bandwidth = 100", "Compliers", "Complier to Sample Ratio")

stargazer(table6, header = F, summary = F,
 title = "Table 6: Characteristics of Entire, Bandwidth = 100, Complier Samples")
```

Table 6

Table 6: Characteristics of Entire, Bandwidth = 100, and Complier Samples

	Entire Sample	Bandwidth = 100	Compliers	Complier to Sample Ratio
Share with Family Income in Quintile 1	0.473	0.503	0.302	0.639
Share with Family Income in Quintile 2	0.216	0.217	0.228	1.052
Share with Family Income in Quintile 3	0.159	0.150	0.212	1.334
Share with Family Income in Quintile 4	0.152	0.130	0.258	1.699
Share Female	0.569	0.586	0.473	0.831
Share with $60 \geq$ GPA ≤ 70	0.319	0.251	0.708	2.219
Share with $50 \geq$ GPA < 60	0.627	0.692	0.303	0.484
Share GPA < 50	0.055	0.057	-0.011	-0.197
Share with Mother Education $>$ HS	0.148	0.123	0.277	1.868
Share with Father Education $>$ HS	0.156	0.128	0.293	1.886

Narrative:

The claim that many advocates of the PSU loan program talk about is the idea that the loan program would provide opportunities to economically disadvantaged students that do well who otherwise would find it much harder to enroll in college. The jump at PSU score of 475 is the complier group who change from $D_i = 0$ *to* $D_i = 1$ at the boundary. The characteristics of the compliers seem to be drastically different from that of the entire sample.

Compliers in generally seem to be wealthier and more middle-class than the entire sample. The compliers seem to be more heavily male than the entire sample. The GPA of the compliers seem to be in general better than that of the entire sample with the majority scoring between 60 and 70 on GPA. Lastly, compliers seem to have highly educated parents which makes sense because well-educated parents tend to focus a lot more on the quality of education for their children and try harder to actively for the loan program.