Applied Econometrics and Public Policy Final Project

Pedram Doroudchi

5/18/2020

We begin with a data set with 16,969 observations, 10,575 of which are men and 6,394 women. Each individual is observed for six consecutive years, the first three years on one job and the next three years on a new job. We have information regarding each individual's age, years of education, gender, years of experience, and log hourly wage as of the first year of the new job. Log hourly wage in all other years was recorded as well. Lastly, mean log wage of other workers at the first and second jobs was recorded for each individual. Let's examine some important summary statistics below:

Table 1: Mean characteristics and log wage quartile percentages of various groups of workers on the first year of a new job

	sample	male	female	t
education	10.4841	10.2721	10.8348	-9.9029^*
age	33.5589	33.5737	33.5344	0.4358
log real hourly wage	1.7879	1.8664	1.6580	20.8163^*
log wage of coworkers	1.6921	1.7249	1.6378	11.6761^*
Q1	0.2503	0.1178	0.1325	
Q2	0.2499	0.1622	0.0876	
Q3	0.2498	0.1708	0.0790	
Q4	0.2500	0.1724	0.0776	

Note: t-statistic between male and female subgroups; $*$ indicates a statistically significant p-value $< 2.2e-16$

Analyzing Table 1, we do not observe a statistically significant disparity between average male and female age while we do with respect to their respective log hourly wages on the first year of a new job: men earn log hourly wages 0.2084 higher on average, even though women hold about a half year more education, a statistically significant difference. This is a surprising finding as on average one would expect that greater education undoubtedly leads to a higher salary. This result also holds with respect to the new-hires' same-sex coworkers: male coworkers earn log hourly wages 0.08 higher than their female counterparts on average. We also observe that male new-hires received a log hourly salary about 0.14 higher than that of their fellow male coworkers while female new-hires earned about 0.02 higher than their fellow female coworkers.

Looking at only the log real hourly wage quartiles for the sample, the sample is fairly equally divided among the quartiles. However, when looking at the fractions of males and females that compose these quartiles, the income inequality becomes quite stark. Starting with the first log wage quartile, men and women are quite similar in percentage, but there are about 1.47% more women. The upper three quartiles are dominated by men: men account for 7.46%, 9.18%, and 9.48% more of the total percentage in these quartiles respectively. It becomes clear in this sample that men are getting higher-paying jobs than women.

The gender wage gap apparent in the numbers becomes even clearer in Figure 1, found on the following page.

Figure 1: Density curves of log hourly wages for men and women

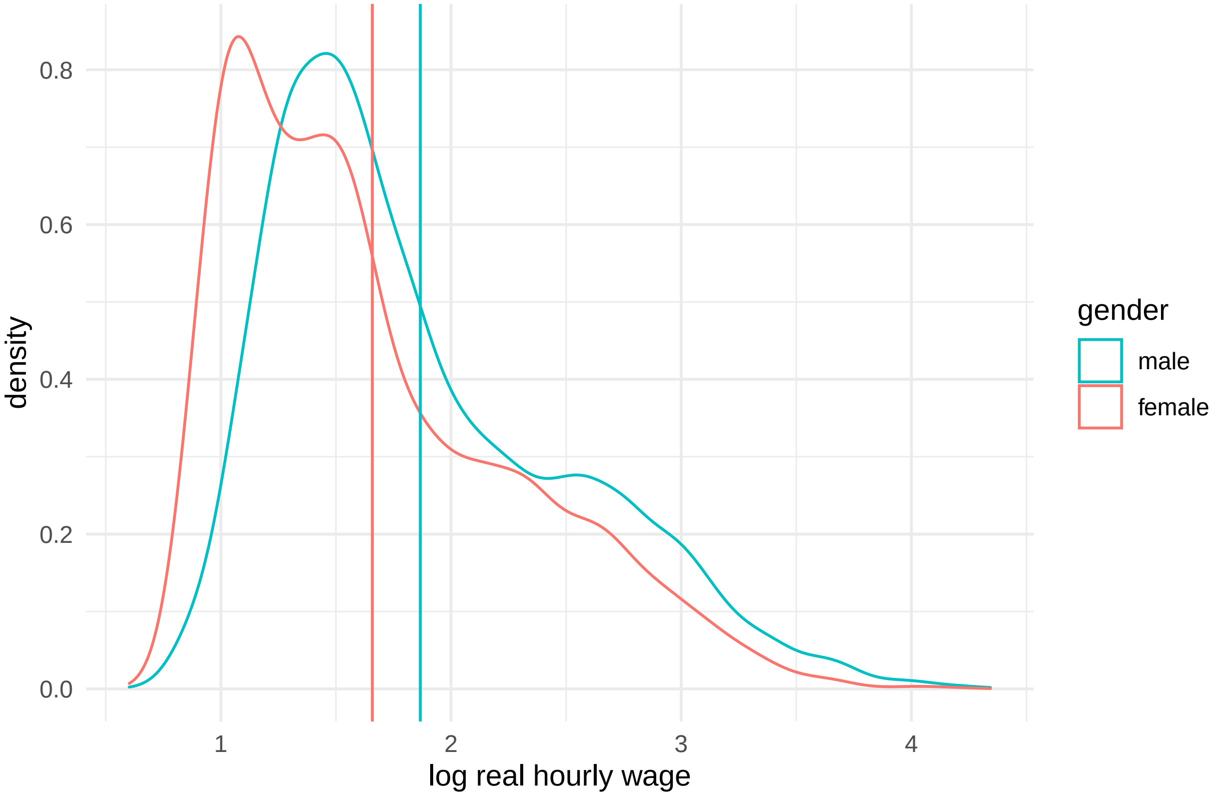


Figure 1 displays the prominence of the gender pay gap in the context of our sample as the smoothed histograms, or density functions, of male and female log hourly wages are pitted against one another. The male curve is clearly shifted to the right of the female curve, indicating that men are making higher salaries than women at all income quartiles, thereby earning higher log hourly wages on average. The male density curve presents a fatter right tail, making clear the fact that more men than women make up the higher income quartiles while more women than men make up the lower income quartiles.

The vertical lines in Figure 1 above indicate the respective mean log hourly wages of males and females, their difference indicating the raw wage gap of -0.2084035. What is causing the apparent wage gap discovered above? Is it due to an inherent male privilege that runs within our society? Is it that males inevitably receive the higher-paying jobs because it is men already holding these other positions, perpetuating the cycle of gender income inequality in the modern-day workplace? Or is it due to an innate cognitive ability immeasurable by any model? To truly understand the observed wage gap, we next attempt to construct some standard wage models.

Table 1, found on the following page, displays the results of four rather simple regressions. Model (1) regresses log wages on a female dummy, serving as an indicator for gender, and a constant. The coefficient on the female indicator should produce the raw wage gap we observe above in theory. Model (2) regresses log wage on a constant, education, a cubic in experience (potential experience, or the number of years since an individual has completed school), and a female dummy. This model seeks to gain explanatory power in regard to the wage gap. Finally, two additional gender-specific models are fit, regressing log wage on a constant, education, and a cubic in experience, in order to determine whether there is a difference in the effects of these regressors between genders.

Table 2: Various models regressing log hourly wage on combinations of education, years of experience cubed, gender, and a constant

	Dependent variable:			
	log real hourly wage			
	(1)	(2)	(3a)	(3b)
educ		0.140^{***}	0.139^{***}	0.142^{***}
		(0.001)	(0.001)	(0.002)
exp^3		0.00002^{***}	0.00002^{***}	0.00002^{***}
		(0.00000)	(0.00000)	(0.00000)
female	-0.208^{***}	-0.280^{***}		
	(0.010)	(0.007)		
Constant	1.866^{***}	0.291^{***}	0.292^{***}	0.011
	(0.006)	(0.013)	(0.017)	(0.021)
Observations	16,969	16,969	10,575	6,394
R^2	0.024	0.541	0.507	0.570
Adjusted R^2	0.024	0.541	0.507	0.570
Residual Std. Error	0.637 (df = 16967)	0.437 (df = 16965)	0.453 (df = 10572)	0.409 (df = 6391)

Note: (1) and (2) pooled, (3a) male-only, (3b) female-only; *p<0.1; **p<0.05; ***p<0.01

We find that model (1), fitting only a constant, returns the raw wage gap as expected. Adding explanatory variables such as education and years of experience cubed in (2) suggests that the wage gap observed is understated as the effect of being female grows from -0.208 to -0.280. This is rather surprising as we observe a significant increase in the R^2 from 0.024 to 0.541, also the same values we get for the adjusted R^2, indicating that our variables have legitimate explanatory power. Models (3) and (4), respectively fit to males and females only, suggest that the effect of the education and experience explanatory variables are similar, but note that the coefficient on education for females is 0.003 higher, while the vast difference between the two models can be observed by comparing the models' sharply contrasting constant coefficients, 0.292 and 0.011 respectively, indicating that males tend to receive a dramatically larger average log wage, assuming zero years of education and experience. This also indicates a sign of severe overfitting with respect to model (3) due to the positive education effect yet negative gender effect on log wage.

The wage difference above may be dissected into two components: one that is due to differences in group characteristics and one that cannot be explained by such differences and may be attributed to outside or unobserved variables not accounted for in the model. Enter the Oaxaca decomposition! On the next page, we will attempt three flavors of the Oaxaca decomposition in order to better understand the individual effects of our explanatory variables in Table 1.

The Oaxaca decomposition, assuming the explanatory variables are the same in both samples, is defined as:

$$\bar{y}^b - \bar{y}^a = \sum_{j=2}^{K} \left(\bar{x}_j^b - \bar{x}_j^a\right) \hat{\beta}_j + \hat{\beta}_{K+1} \tag{1}$$

We know the raw wage gap to be the lefthand side of equation (1):

$$\bar{y}^f - \bar{y}^m = 1.658045 - 1.866448 = -0.2084035$$

This implies women earn log wages about 0.208 less than men on average. We can break down this result into individual regressor components:

$$\sum_{j=2}^{K} \left(\bar{x}_j^f - \bar{x}_j^m\right) \hat{\beta}_j + \hat{\beta}_{K+1} = \left(\bar{x}_{educ}^f - \bar{x}_{educ}^m\right) \hat{\beta}_{educ} + \left(\bar{x}_{exp^3}^f - \bar{x}_{exp^3}^m\right) \hat{\beta}_{exp^3} + \hat{\beta}_{female}$$

$$= (10.83485 - 10.27206)(0.140263) + (0.140263 - 7322.828)(1.841987e - 05) + -0.2797659 = -0.2084035$$

When the coefficients of the explanatory variables are not the same in both samples, the Oaxaca decomposition may alternatively be expressed in two additional yet equivalent ways:

$$\bar{y}^b - \bar{y}^a = (\bar{x}^b - \bar{x}^a)^T \hat{\beta}^a + (\bar{x}^b)^T (\hat{\beta}^b - \hat{\beta}^a) \tag{2}$$

$$= (\bar{x}^b - \bar{x}^a)^T \hat{\beta}^b + (\bar{x}^a)^T (\hat{\beta}^b - \hat{\beta}^a) \tag{3}$$

The first term in each of the above equations expresses the portion of the wage difference that is due to differences in group characteristics while the second term reflects a portion that cannot be explained by such differences and may be attributed to outside or unobserved variables not accounted for in the model, such as discrimination.

After manually performing the above decompositions in R, for the first terms we observe:

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^a = 0.07019461$$

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^b = 0.07310372$$

For the second terms we calculate:

$$(\bar{x}^b)^T (\hat{\beta}^b - \hat{\beta}^a) = -0.2785981$$

$$(\bar{x}^a)^T (\hat{\beta}^b - \hat{\beta}^a) = -0.2815072$$

Adding them up, indeed we find the observed gender wage gap:

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^a + (\bar{x}^b)^T(\hat{\beta}^b - \hat{\beta}^a) = 0.07019461 - 0.2785981 = -0.2084035$$

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^b + (\bar{x}^a)^T(\hat{\beta}^b - \hat{\beta}^a) = 0.07310372 - 0.2815072 = -0.2084035$$

Because females are better educated than males, these estimates do not explain any of the observed wage gap, in fact suggesting that the wage gap is understated. We must improve our model better to understand what exactly is causing the observed gender wage gap. Adding the mean log wage of an individual's coworkers as an explanatory variable, we observe the results below:

Table 3: Various models regressing log hourly wage on combinations of education, years of experience cubed, gender, mean log wage of other workers on the second job, and a constant

	Dependent variable:			
	log real hourly wage			
	(4)	(5)	(6a)	(6b)
educ		0.089***	0.090***	0.087***
		(0.001)	(0.001)	(0.002)
exp^3		0.00001***	0.00002***	0.00001***
		(0.00000)	(0.00000)	(0.00000)
female	−0.121***	−0.196***		
	(0.007)	(0.006)		
owage2	1.006***	0.651***	0.678***	0.616***
	(0.007)	(0.007)	(0.009)	(0.011)
Constant	0.131***	−0.280***	−0.355***	−0.361***
	(0.013)	(0.013)	(0.017)	(0.019)
Observations	16,969	16,969	10,575	6,394
R^2	0.549	0.688	0.667	0.708
Adjusted R^2	0.548	0.688	0.667	0.708
Residual Std. Error	0.434 (df = 16966)	0.360 (df = 16964)	0.372 (df = 10571)	0.337 (df = 6390)
Note:	(4) and (5) pooled, (6a) male-only, (6b) female-only; *p<0.1; **p<0.05; ***p<0.01			

With the results of the regressions above, let's again perform both equivalent forms of the Oaxaca decomposition that assume the coefficients of the explanatory variables between genders differ. By looking at models (6a) and (6b), male-only and female-only respectively, we definitely notice a sharp difference with respect to the *owage*2 coefficient, which represents the effect of an individual's coworkers' mean wages.

Again, the decomposition may be expressed as:

$$\bar{y}^b - \bar{y}^a = (\bar{x}^b - \bar{x}^a)^T \hat{\beta}^a + (\bar{x}^b)^T(\hat{\beta}^b - \hat{\beta}^a) \tag{2}$$

$$= (\bar{x}^b - \bar{x}^a)^T \hat{\beta}^b + (\bar{x}^a)^T(\hat{\beta}^b - \hat{\beta}^a) \tag{3}$$

Let's analyze the estimate of the first term, which expresses the portion of the wage difference that is due to differences in group characteristics:

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^a = -0.01520622$$

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^b = -0.009182264$$

We note here that the effect of adding coworkers' wages to the model was strong enough to actually produce explanatory power. Now, our regressors explain anywhere between 4.4% and 7.3% of the wage gap. Let's look at the explanatory power of *owage*2 when it isn't masked by the negating effect of education, which still suggests that the wage gap is understated, but not as much so:

$$(\bar{x}^f - \bar{x}^m)\hat{\beta}^m_{owage2} = -0.05902288$$

$$(\bar{x}^f - \bar{x}^m)\hat{\beta}^f_{owage2} = -0.05362588$$

These coefficient estimates suggest that the effect of coworkers' wages explains somewhere between 25.7% and 28.3% of the observed wage gap, producing a huge difference compared to the previous model.

Recall that the second term reflects a portion that cannot be explained by such differences and may be attributed to outside or unobserved variables not accounted for in the model. For the second terms we calculate:

$$(\bar{x}^b)^T (\hat{\beta}^b - \hat{\beta}^a) = -0.1931973$$

$$(\bar{x}^a)^T (\hat{\beta}^b - \hat{\beta}^a) = -0.1992212$$

We see that though the effect of coworker's wages makes quite a difference to the gender wage gap, over 90% may be attributed to variables outside of the model. In particular, let's observe the portion of the wage gap that cannot be explained by coworkers' wages:

$$(\bar{x}^f)(\hat{\beta}^f_{owage2} - \hat{\beta}^m_{owage2}) = -0.1014841$$

$$(\bar{x}^m)^T (\hat{\beta}^f_{owage2} - \hat{\beta}^m_{owage2}) = -0.1068811$$

Even though the effect of coworkers' wages explains about a quarter of the wage gap, differences in its gender-specific coefficients account for over half of the potential discrimination at play. Though the wage effect of working with highly paid coworkers is quite large, the effect is much smaller for women than for men.

Adding up individual components of the decompositions, we indeed find the observed gender wage gap:

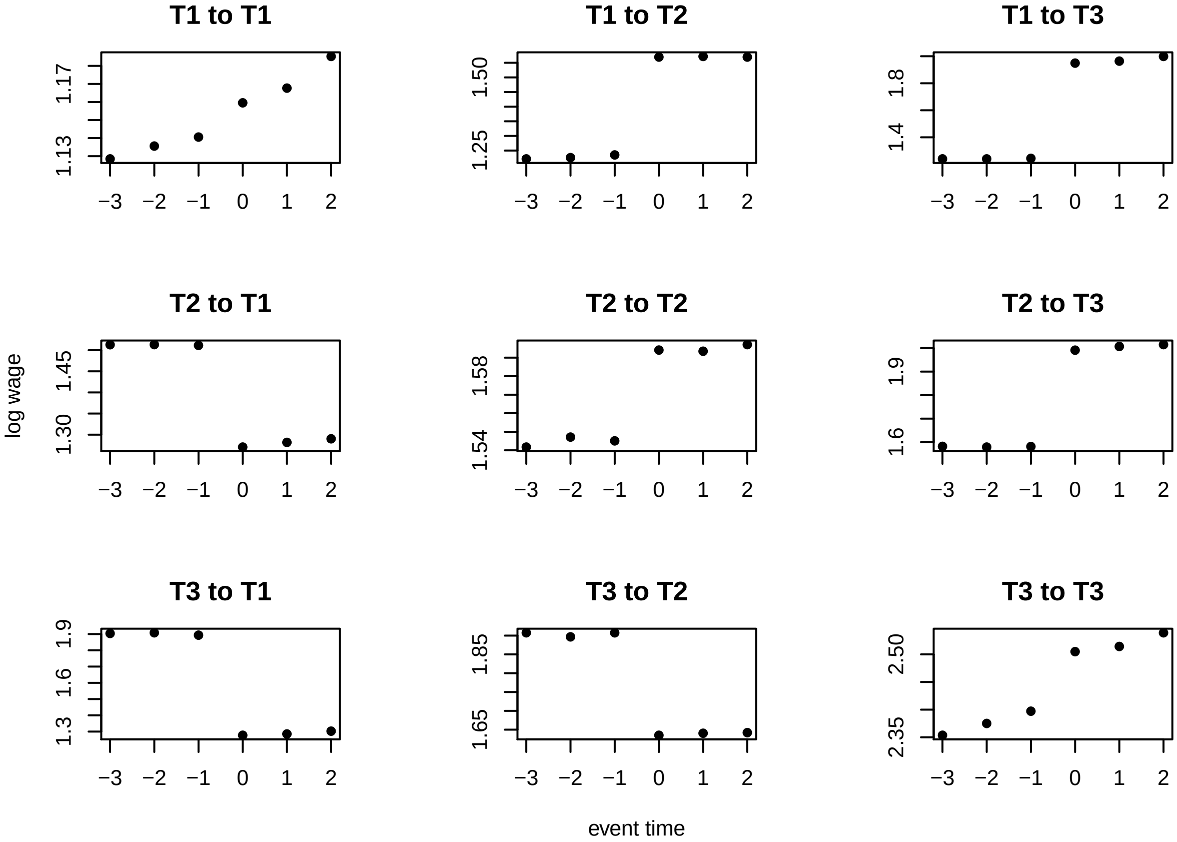
$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^a + (\bar{x}^b)^T (\hat{\beta}^b - \hat{\beta}^a) = -0.01520622 - 0.1931973 = -0.2084035$$

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^b + (\bar{x}^a)^T (\hat{\beta}^b - \hat{\beta}^a) = -0.009182264 - 0.1992212 = -0.2084035$$

The analysis above points to a fundamental reality of the gender wage gap. It is logical that if someone works with people who earn higher salaries, they must share similar job functions and skills to be able to collaborate efficiently. Thus, if they are performing similar jobs, they will be paid similarly. This is what the first term of the Oaxaca decomposition posits. Both women and men who work with higher-paid coworkers also get paid more. However, what the second term of the decomposition tells us is that getting a job with high-paid coworkers involves some form of good luck or solid connections, and men may have it better in this regard. Perhaps it is also that males are more aggressive in negotiating a higher salary from the start while women may feel that they do not have that luxury. The wage gap is self-perpetuating in nature because males have better connections set in higher places as it is more men that hold these positions, perpetuating the cycle of inequality and indicating that some level of gender bias may be at play.

Is it easier for those who work with higher-paid coworkers to find jobs with even higher-paid coworkers while it is more difficult for those at the bottom or in the middle? We must now shift our attention to the presence of job changers in the data in order to analyze wage changes as people move between jobs with higher and lower paid coworkers. This idea will help us better understand the difference in salary changes individuals experience in different terciles of coworkers' wages on the first job as they move between terciles of coworkers' wages on the second job. This event study is visualized on the following page.

Figure 2: Mean wages over time for people who start in each tercile of owage1 and move to each tercile of owage2



Looking at the event study graphs of Figure 2 above, we observe a general pattern of wage movements before a job change: wage stagnation or even decline. Workers expect to be rewarded for their hard work at the end of each year, and if not they will find a new job, just as the data shows. Exceptions to this rule are found for those in the first and third terciles who stay in their respective terciles after the job change, whose wages increase each year and continue to increase after the job change, but more so for those who remain in the third tercile. Our analysis has found a wage gap in favor of men who are generally paid more than women, which implies men make up more of the upper terciles than do women while there are more women than men in the lower terciles.

What's concerning is that those with the highest wages, the majority being men, continue to receieve the highest on-the-job raises and salary boosts when switching jobs. This reaffirms the previously discussed idea that the wage gap is self-perpetuating: males have an easier time attaining higher-paying positions because it is other men that hold these similar positions, with whom more men than women hold connections. It is interesting to note that those in the third tercile who remain in the third tercile received consistent raises over the three years before the job change (while those in the middle tercile received none), indicating that perhaps higher-paid individuals, consisting of more men than women, may search harder to find even higher-paying jobs with higher-paid coworkers, or may already have connections in place that provide them the additional opportunity.

In the models below, we seek to model the change in wages from the year before the job change to the year of the job change as a function of the change in the mean log wage of coworkers between job periods. These first-differenced models, as seen below, are assumed to control for all unobserved characteristics of individuals.

Table 4: Various models regressing first-differenced log hourly wage on combinations of years of experience as of period -1 squared, gender, change in mean log wage of coworkers, and a constant

	Dependent variable:			
	first-differenced log real hourly wage			
	(7)	(8)	(9a)	(9b)
$(exp-1)^2$		-0.0001^{***}	-0.0001^{***}	-0.0001^{***}
		(0.00001)	(0.00001)	(0.00001)
female	-0.019^{***}	-0.021^{***}		
	(0.004)	(0.004)		
Dwage	0.294^{***}	0.291^{***}	0.332^{***}	0.219^{***}
	(0.006)	(0.006)	(0.008)	(0.010)
Constant	0.045^{***}	0.082^{***}	0.080^{***}	0.065^{***}
	(0.003)	(0.004)	(0.005)	(0.005)
Observations	16,969	16,969	10,575	6,394
R^2	0.110	0.119	0.137	0.086
Adjusted R^2	0.109	0.118	0.137	0.086
Residual Std. Error	0.263 (df = 16966)	0.261 (df = 16965)	0.273 (df = 10572)	0.239 (df = 6391)

Note: (7) and (8) pooled, (9a) male-only, (9b) female-only; *p<0.1; **p<0.05; ***p<0.01

Now that we have the assumed true causal effects of coworker wages in the pooled, male-only, and female-only models, we can modify the Oaxaca decompositions performed on Table 3 by subtracting the estimated effect of coworkers' wages using the assumed causal coefficients and in turn re-estimating the other model coefficients by regressing the desired predictors on the difference, like so:

$$y_i - \beta_1 x_{1i} = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + e_i \tag{4}$$

We may then use the updated regressor coefficients in place of those used previously with respect to Table 3. Let's first analyze the estimate of the first term, which expresses the portion of the wage difference that is due to differences in group characteristics:

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^a = 0.02829239$$

$$(\bar{x}^b - \bar{x}^a)^T \hat{\beta}^b = 0.04379245$$

In particular, the contribution of differences in coworkers' wages is found to be:

$$(\bar{x}^f - \bar{x}^m)\hat{\beta}^m_{Dwage} = -0.02895978$$

$$(\bar{x}^f - \bar{x}^m)\hat{\beta}^f_{Dwage} = -0.01910219$$

After filtering for the true causal effect of coworker wages, the coefficient estimates directly above suggest that the effect of coworkers' wages now explains only between 9.2% and 13.9% of the observed wage gap, less than half of the previous model estimate between 25.7% and 28.3%. We find that most of the explanatory power of coworkers' wages was due to unobserved characteristics of the individuals that have now been left out. As in the case of our initial models that did not include coworker wages, we find that these controlled variables do not explain any portion of the wage difference and in fact suggest that the wage gap is understated due to the additional amount of mean education women possess.

Recall that the second term reflects a portion that cannot be explained by such differences and may be attributed to outside or unobserved variables not accounted for in the model. Since the first term is positive, we expect this term to compensate the positivity of the first term with a negative number larger than the raw wage gap. We find:

$$(\bar{x}^b)^T(\hat{\beta}^b - \hat{\beta}^a) = -0.2366959$$

$$(\bar{x}^a)^T(\hat{\beta}^b - \hat{\beta}^a) = -0.2521959$$

In particular, let's observe the portion of coworkers' wages attributable to outside factors such as gender discrimination:

$$(\bar{x}^f)(\hat{\beta}^f_{owage2} - \hat{\beta}^m_{owage2}) = -0.18536$$

$$(\bar{x}^m)^T(\hat{\beta}^f_{owage2} - \hat{\beta}^m_{owage2}) = -0.1952176$$

Summing individual components of the decompositions, we indeed find the observed gender wage gap:

$$(\bar{x}^b - \bar{x}^a)^T\hat{\beta}^a + (\bar{x}^b)^T(\hat{\beta}^b - \hat{\beta}^a) = 0.02829239 - 0.2366959 = -0.2084035$$

$$(\bar{x}^b - \bar{x}^a)^T\hat{\beta}^b + (\bar{x}^a)^T(\hat{\beta}^b - \hat{\beta}^a) = 0.04379245 - 0.2521959 = -0.2084035$$

Differences in the gender-specific coefficients of $owage2$ account for the vast majority of potential discrimination at play. We essentially find here that the wage effect of working with highly paid coworkers is not as large as previously thought, rather representing potential gender bias, discrimination, or some other variable not accounted for that is at the root of the observed wage gap.

Part II

We shift our attention to a student-level data set containing information on 112,008 students in Chile who finished high school and were eligible to enter college. Our topic of interest involves an exam called the "PSU," a standardized test which Chilean students must write at the end of high school. Students who score at least a 475 (exam scores between 300 and 700) and who have a family income below the 80th percentile are eligible for government loan support for college costs, while students who score below a 475 cannot receive the loan. The data set includes information on students' PSU scores, whether the student scored above a 475, whether the student entered college, high school GPA (between 0 and 70), whether the student attended a private high school, whether students' mothers and fathers have more than a high school education, gender, and family income quintiles, excluding the top quintile since students from the top quintile are not eligible for the government loan program.

We begin by formulating a proof in order to construct the mean characteristics of the compliers for a fuzzy regression discontinuity (RD) model using an extension of a two stage least squares (2SLS) model.

The compliers in this context are students with scores very close to 475 who will enter college if they get the loan, but will not enter college if they don't.

To begin the proof, note:

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

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

(i) Using expressions (1) and (2) above, we prove that:

$$E[w_i|C(0)] = \frac{E[w_i|AT(0) \; or \; C(0)]P(AT(0) \; or \; C(0)) - E[w_i|AT(0)]P(AT(0))}{P(C(0))} \tag{1}$$

$$E[w_i|C(0)]P(C(0)) = E[w_i|AT(0) \; or \; C(0)]P(AT(0) \; or \; C(0)) - E[w_i|AT(0)]P(AT(0)) \tag{2}$$

$$= \frac{E[w_i|AT(0)]P(AT(0)) + E[w_i|C(0)]P(C(0))}{P(AT(0) \; or \; C(0))} P(AT(0) \; or \; C(0)) - E[w_i|AT(0)]P(AT(0)) \tag{3}$$

$$= E[w_i|AT(0)]P(AT(0)) + E[w_i|C(0)]P(C(0)) - E[w_i|AT(0)]P(AT(0)) \tag{4}$$

$$\therefore E[w_i|C(0)]P(C(0)) = E[w_i|C(0)]P(C(0)) \tag{5}$$

(ii) Using the law of iterated expectations ($E[Y] = E[E[Y|X]]$), we show that:

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

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

$$\therefore E[w_i|AT(0) \; or \; C(0)]P(AT(0) \; or \; C(0)) = E[w_i|AT(0) \; or \; C(0)]P(AT(0) \; or \; C(0)) \tag{3}$$

(iii) Again, using the law of iterated expectations, it may be proven that:

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

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

$$\therefore E[w_i | AT(0)] P(AT(0)) = E[w_i | AT(0)] P(AT(0)) \tag{3}$$

(iv) Considering a “goofy” 2SLS RD model, we may show that the 2SLS estimate of β_1 is an estimate of $E[w_i|C(0)]$, as so:

$$E[w_i | C(0)] = \frac{E[w_i | AT(0) \text{ } or \text{ } C(0)] P(AT(0) \text{ } or \text{ } C(0)) - E[w_i | AT(0)] P(AT(0))}{P(C(0))} \tag{1}$$

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

$$= \frac{E[\delta_0 + \delta_1 z_i + \delta_2 x_i + \delta_3 x_i z_i + v_i | x_i \to 0, z_i = 1] - E[\delta_0 + \delta_1 z_i + \delta_2 x_i + \delta_3 x_i z_i + v_i | x_i \to 0, z_i = 0]}{P(C(0))} \tag{3}$$

$$= \frac{\delta_0 + \delta_1 - \delta_0}{P(C(0))} \tag{4}$$

$$= \frac{\delta_1}{P(C(0))} \tag{5}$$

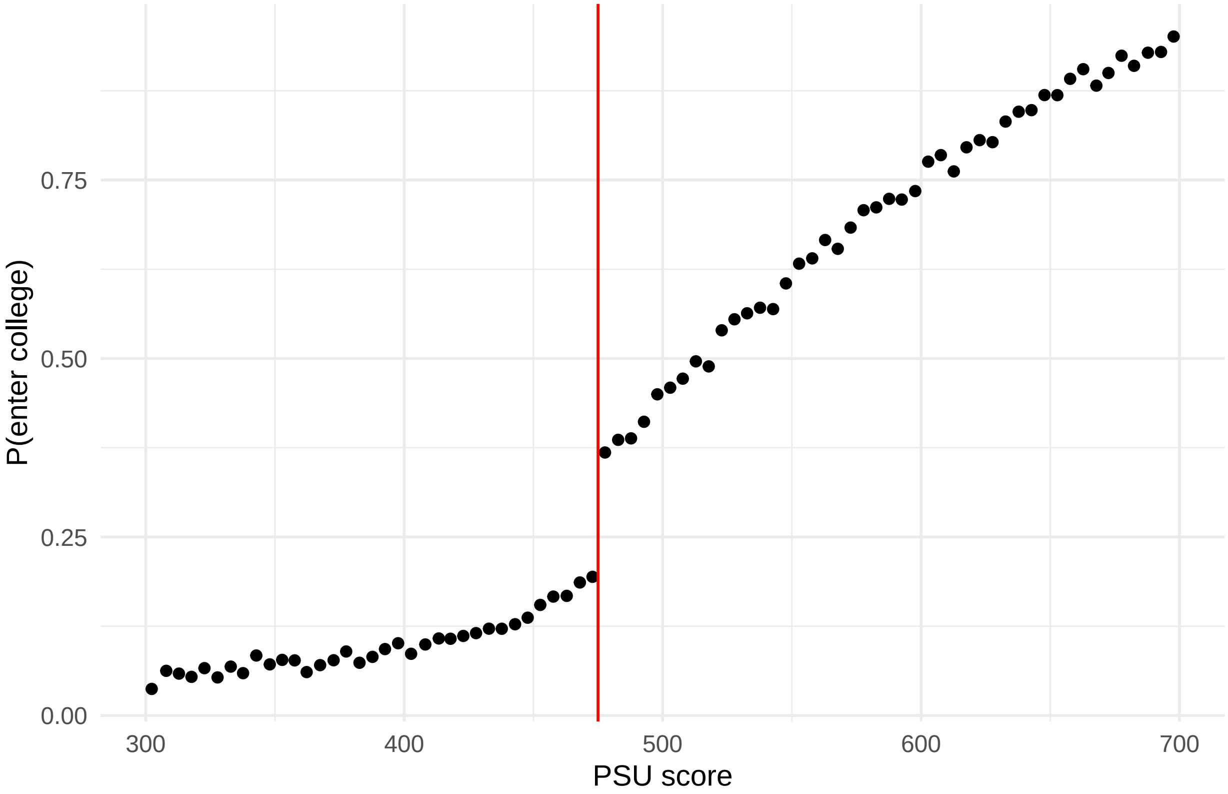
$$= \frac{\delta_1}{\pi_1} \tag{6}$$

$$= \frac{\beta_1 \pi_1}{\pi_1} \tag{7}$$

$$\therefore E[w_i | C(0)] = \beta_1 \tag{8}$$

We now seek to show the relationship between PSU score and the probability of entering college. This relationship is visualized in Figure 3 on the following page, grouping observations by 5-point bins with respect to PSU score.

Figure 3: Probability of entering college against mean PSU score



We observe a relatively flatter positive linear trend in the probability of entering college as mean PSU score increases below the 475-point cutoff. The discontinuity at the cutoff is apparent, where there is a sharp jump in the probability that a student enters college above that score. This jump represents the fraction of compliers in the sample. For scores above 475, the probability of entering college follows a much steeper positive linear trend as mean PSU score increases.

We next look to estimate local linear first stage models for the probability of attending college, all having the form:

$$D_i = \pi_0 + \pi_1 z_i + \pi_2 x_i + \pi_3 x_i z_i + \epsilon_i \tag{3}$$

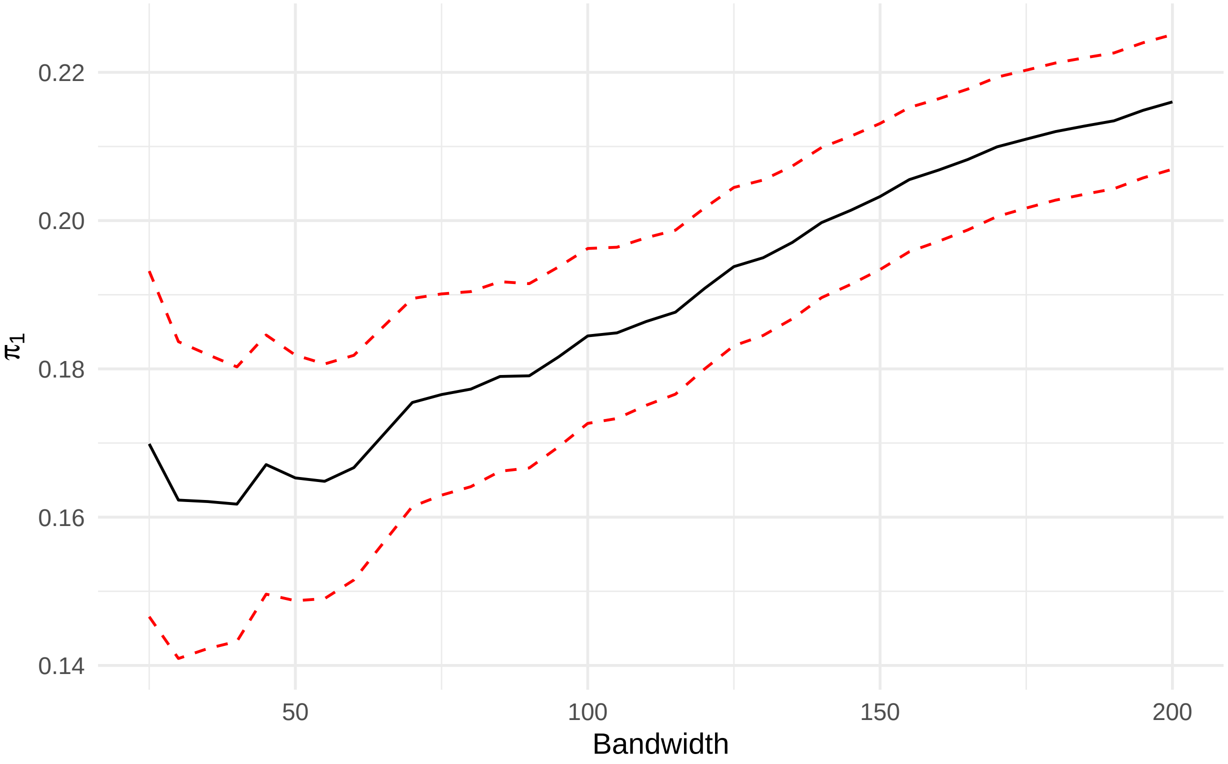
where $x_i = PSU_i - 475$, $z_i = 1[PSU_i \geq 475]$, and $D_i = entercollege$. We estimate the above model using bandwidths $B = \{25, 50, 75, 100\}$, as shown in Table 5 on the next page.

Below Table 5 lies Figure 4, which estimates model (4) using a sequence of bandwidths from 25 to 200 by 5-point intervals. 95% confidence intervals are shown as dashed red lines.

Table 5: Local linear first stage models for the probability of attending college using different "bandwidths" (B) around PSU score of 475 required for government loan eligibility

	Dependent variable:			
	entercollege			
	(1)	(2)	(3)	(4)
$1[PSU \geq 475]$	0.170***	0.165***	0.177***	0.184***
	(0.012)	(0.008)	(0.007)	(0.006)
$PSU - 475$	0.001	0.002***	0.001***	0.001***
	(0.001)	(0.0002)	(0.0001)	(0.0001)
interaction	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
Residual Std. Error	0.440 (df = 22842)	0.436 (df = 43848)	0.434 (df = 63066)	0.430 (df = 79519)
Note:	Note: (1) $B = 25$, (2) $B = 50$, (3) $B = 75$, (4) $B = 100$; *p<0.1; **p<0.05; ***p<0.01			

Figure 4: $1\left[\text{PSU} \geq 475\right]$ estimates against bandwidth $25 \leq B \leq 200$



Using Figure 4 and the estimates in Table 5, a bandwidth of 75 appears to be a reasonable choice. Analyzing Figure 4 above, it becomes clear that a bandwidth of 75 achieves a somewhat optimal bias-variance tradeoff since the confidence bands associated with the estimate at this bandwidth remain roughly constant as bandwidth increases, while the confidence intervals are not as large as smaller bandwidth estimates. The π_1 estimate associated with this bandwidth avoids the turbulent estimates associated with smaller bandwidths while avoiding the constant increasing linear bias of bandwidths greater than 75. Model (3) in Table 5, which is associated with a bandwidth of 75, appears to attain a solid compromise between the models associated with bandwidths of 25, 50, and 100 as the coefficient on $1[PSU \geq 475]$ is a rough average of the other models' respective estimates. Its coefficient standard error is 0.005 less than that of model (1) with $B = 25$, a significant difference, while it is only 0.001 greater than that of model (4) with $B = 100$.

Finally, using the proof developed earlier and the bandwidth choice made above, let's provide estimates of the means of several characteristics of the compliers to the loan program:

Table 6: Various estimated characteristic means for entire sample, bandwidth of 75 points, compliers to the loan program, and complier-to-sample ratio

	sample	B = 75	compliers	ratio
family income Q1	0.4726	0.5127	0.4488	0.9497
family income Q2	0.2164	0.2167	0.2252	1.0409
family income Q3	0.1591	0.1471	0.1726	1.0844
family income Q4	0.1519	0.1235	0.1534	1.0097
female	0.5685	0.5885	0.6007	1.0567
GPA high	0.3189	0.2350	0.2931	0.9193
GPA mid	0.6266	0.7075	0.6604	1.0539
GPA low	0.0545	0.0575	0.0465	0.8528
mother > HS	0.1483	0.1170	0.1479	0.9971
father > HS	0.1556	0.1222	0.1506	0.9680

Using Table 6 above, we see that student compliers in the lowest quintile of family income make up the highest portion of compliers (0.4488) but are still underrepresented since the complier-sample ratio is 0.9497. On the other hand, compliers in the upper three quintiles are overrepresented, especially those in the third quintile with a ratio of 1.0844. This points to the fact that though the loan program does extend college access to more economically disadvantaged students, it should be doing much more for the majority of students in the lowest quintile of family income since students in the second, third, and fourth quintiles are overrepresented with respect to the sample. We should be seeing a far smaller percentage of students in the fourth quintile who are compliers if the loan program truly seeks to extend college access to the most economically disadvantaged students, who make up almost half of the sample.

We see that 60.07% of compliers are female, overrepresenting the sample with a complier-sample ratio of 1.0567. We also see that the majority of compliers, 66.04% precisely, have GPAs in the middle of the pack and overrepresent the sample by a factor of about 1.05. Compliers with low GPAs are quite underrepresented with a ratio of about 0.85 and represent less than 5% of the compliers. Those with high GPAs make up almost 30% of the compliers and are underrepresented compared to the sample, with a ratio of 0.92. Compliers with mothers and fathers who possess more than a high school education are underrepresented compared to the sample, but not by much, and more so for fathers than mothers. These ratios in particular should be even lower since students whose parents have less education are more likely to make less, and thus make up more of the lower family income quintiles.

After analyzing these statistics, it indeed becomes clear that the Chilean loan program is not doing everything in its power to extend college access to its most economically disadvantaged students.

Appendix

Part I

Table 1

```r
# import dataset for both parts
dd <- read.csv('project2020_dd.csv')
rd <- read.csv('project2020_rd.csv')

# find summary statistics for all workers
sample <- as.numeric(dd %>% summarize(mean(educ), mean(age), mean(y), mean(owage2)))

# find summary statistics for male and female workers
ddm <- subset(dd, female == 0)
ddf <- subset(dd, female == 1)
male <- apply(ddm[ ,c(3,2,1,11)], 2, 'mean')
female <- apply(ddf[ ,c(3,2,1,11)], 2, 'mean')

# find t-statistics for mean differences between genders
t <- c()
i <- 1
for (c in c(3,2,1,11)) {
  t[i] <- t.test(ddm[ ,c], ddf[ ,c])$statistic
  i <- i+1
}

# find quartiles of log wages for all workers
dd$quartile <- cut(dd$y, quantile(dd$y, seq(0,1,0.25)), labels = FALSE)

# find percentages in each quartile
perc_f <- c()
perc_m <- c()
perc_sample <- c()
for (i in 1:4) {
  perc_f[i] <- mean(dd$female == 1 & dd$quartile == i)
  perc_m[i] <- mean(dd$female == 0 & dd$quartile == i, na.rm = TRUE)
  perc_sample[i] <- perc_f[i] + perc_m[i]
}

# create Table 1
sample <- c(sample, perc_sample)
male <- c(male, perc_m)
female <- c(female, perc_f)
t <- c(t, rep(NA, 4))
t1 <- cbind(sample, male, female, t)
rownames(t1) <- c('education', 'age', 'log real hourly wage', 'log wage of coworkers',
                  'Q1', 'Q2', 'Q3', 'Q4'
                  )
colnames(t1) <- c('sample', 'male', 'female', 't')
stargazer(t1, digits = 4)
```

Figure 1

```
# plot density curves
ggplot(dd, aes(x = y, color = factor(female))) +
  geom_density() +
  geom_vline(xintercept = mean(ddm$y), color = '#00BFC4') +
  geom_vline(xintercept = mean(ddf$y), color = '#F8766D') +
  labs(title = 'Figure 1: Density curves of log hourly wages for men and women',
       x = 'log real hourly wage'
       ) +
  scale_color_manual(name = 'gender', values = c('#00BFC4', '#F8766D'), labels = c('male', 'female')) +
  theme_minimal()
```

Table 2

```
# fit two models using pooled male and female data
m1 <- lm(y ~ female, data = dd)
m2 <- lm(y ~ educ + I(exp^3) + female, data = dd)

# fit separate models for men and women
m3a <- lm(y ~ educ + I(exp^3), data = ddm)
m3b <- lm(y ~ educ + I(exp^3), data = ddf)

# create table
stargazer(m1, m2, m3a, m3b)
```

Table 2: standard Oaxaca decomposition

```
# construct Oaxaca decomposition

# unexplained wage gap
as.numeric(m2$coefficients['female']) +
# effect of education
(mean(ddf$educ) - mean(ddm$educ)) * as.numeric(m2$coefficients['educ']) +
# effect of experience cubed
(mean(ddf$exp^3) - mean(ddm$exp^3)) * as.numeric(m2$coefficients['I(exp^3)'])
```

```
## [1] -0.2084035
```

```
# compare with difference of mean log hourly wages between genders
mean(ddf$y) - mean(ddm$y)
```

```
## [1] -0.2084035
```

```
# lastly compare with coefficent on female dummy in first model with constant
as.numeric(m1$coefficients['female'])
```

```
## [1] -0.2084035
```

Table 2: alternative decompositions

```
# first alternative method

# contribution of difference in education
(mean(ddf$educ) - mean(ddm$educ)) * as.numeric(m3a$coefficients['educ']) +
# contribution of difference in experience cubed
(mean(ddf$exp^3) - mean(ddm$exp^3)) * as.numeric(m3a$coefficients['I(exp^3)']) +
# unexplained difference in education
(mean(ddf$educ)) * as.numeric((m3b$coefficients['educ'] - m3a$coefficients['educ'])) +
# unexplained difference in experience cubed
(mean(ddf$exp^3)) * as.numeric((m3b$coefficients['I(exp^3)'] - m3a$coefficients['I(exp^3)'])) +
# unexplained difference in constant coefficient
as.numeric((m3b$coefficients['(Intercept)'] - m3a$coefficients['(Intercept)']))
```

```
## [1] -0.2084035
```

```
# second alternative method

# contribution of difference in educaiton
(mean(ddf$educ) - mean(ddm$educ)) * as.numeric(m3b$coefficients['educ']) +
# contribution of difference in experience cubed
(mean(ddf$exp^3) - mean(ddm$exp^3)) * as.numeric(m3b$coefficients['I(exp^3)']) +
# unexplained difference in education
(mean(ddm$educ)) * as.numeric((m3b$coefficients['educ'] - m3a$coefficients['educ'])) +
# unexplained difference in experience cubed
(mean(ddm$exp^3)) * as.numeric((m3b$coefficients['I(exp^3)'] - m3a$coefficients['I(exp^3)'])) +
# unexplained difference in constant coefficient
as.numeric((m3b$coefficients['(Intercept)'] - m3a$coefficients['(Intercept)']))
```

```
## [1] -0.2084035
```

Table 3

```
# fit two models using pooled male and female data
m4 <- lm(y ~ female + owage2, data = dd)
m5 <- lm(y ~ educ + I(exp^3) + female + owage2, data = dd)

# fit separate models for men and women
m6a <- lm(y ~ educ + I(exp^3) + owage2, data = ddm)
m6b <- lm(y ~ educ + I(exp^3) + owage2, data = ddf)

# create table
stargazer(m4, m5, m6a, m6b)
```

Table 3: alternative Oaxaca decompositions

```
# first alternative method

# contribution of difference in education
(mean(ddf$educ) - mean(ddm$educ)) * as.numeric(m6a$coefficients['educ']) +
# contribution of difference in experience cubed
(mean(ddf$exp^3) - mean(ddm$exp^3)) * as.numeric(m6a$coefficients['I(exp^3)']) +
# contribution of difference in coworkers' wages
```

```
(mean(ddf$owage2) - mean(ddm$owage2)) * as.numeric(m6a$coefficients['owage2']) +
# unexplained difference in education
(mean(ddf$educ)) * as.numeric((m6b$coefficients['educ'] - m6a$coefficients['educ'])) +
# unexplained difference in experience cubed
(mean(ddf$exp^3)) * as.numeric((m6b$coefficients['I(exp^3)'] - m6a$coefficients['I(exp^3)'])) +
# unexplained difference in coworkers' wages
(mean(ddf$owage2)) * as.numeric((m6b$coefficients['owage2'] - m6a$coefficients['owage2'])) +
# unexplained difference in constant coefficient
as.numeric((m6b$coefficients['(Intercept)'] - m6a$coefficients['(Intercept)']))
```

```
## [1] -0.2084035
```

```
# second alternative method

# contribution of difference in education
(mean(ddf$educ) - mean(ddm$educ)) * as.numeric(m6b$coefficients['educ']) +
# contribution of difference in experience cubed
(mean(ddf$exp^3) - mean(ddm$exp^3)) * as.numeric(m6b$coefficients['I(exp^3)']) +
# contribution of difference in coworkers' wages
(mean(ddf$owage2) - mean(ddm$owage2)) * as.numeric(m6b$coefficients['owage2']) +
# unexplained difference in education
(mean(ddm$educ)) * as.numeric((m6b$coefficients['educ'] - m6a$coefficients['educ'])) +
# unexplained difference in experience cubed
(mean(ddm$exp^3)) * as.numeric((m6b$coefficients['I(exp^3)'] - m6a$coefficients['I(exp^3)'])) +
# unexplained difference in coworkers' wages
(mean(ddm$owage2)) * as.numeric((m6b$coefficients['owage2'] - m6a$coefficients['owage2'])) +
# unexplained difference in constant coefficient
as.numeric((m6b$coefficients['(Intercept)'] - m6a$coefficients['(Intercept)']))
```

```
## [1] -0.2084035
```

Figure 2

```
# find terciles of owage1 and owage 2 and their product
terc1 <- quantile(dd$owage1, probs = seq(0,1,1/3))
terc2 <- quantile(dd$owage2, probs = seq(0,1,1/3))

# classify first jobs based on terciles for owage1
yl1 <- as.numeric(cut(dd$yl1, breaks = terc1, labels = FALSE))
yl2 <- as.numeric(cut(dd$yl2, breaks = terc1, labels = FALSE))
yl3 <- as.numeric(cut(dd$yl3, breaks = terc1, labels = FALSE))

# classify second jobs based on terciles for owage2
yp0 <- as.numeric(cut(dd$y, breaks = terc2, labels = FALSE))
yp0[yp0 == 1] <- 5
yp0[yp0 == 2] <- 6
yp0[yp0 == 3] <- 7
yp1 <- as.numeric(cut(dd$yp1, breaks = terc2, labels = 4:6))
yp1[yp1 == 1] <- 5
yp1[yp1 == 2] <- 6
yp1[yp1 == 3] <- 7
yp2 <- as.numeric(cut(dd$yp2, breaks = terc2, labels = 4:6))
yp2[yp2 == 1] <- 5
yp2[yp2 == 2] <- 6
```

```
yp2[yp2 == 3] <- 7

# classify into nine groups based on product of terciles
yl1_yp0 <- factor(yl1 * yp0)
yl1_yp1 <- factor(yl1 * yp1)
yl1_yp2 <- factor(yl1 * yp2)
yl2_yp0 <- factor(yl2 * yp0)
yl2_yp1 <- factor(yl2 * yp1)
yl2_yp2 <- factor(yl2 * yp2)
yl3_yp0 <- factor(yl3 * yp0)
yl3_yp1 <- factor(yl3 * yp1)
yl3_yp2 <- factor(yl3 * yp2)

# aggregate based on groups and find mean wage at each time point
yl3 <- rowMeans(cbind(aggregate(dd$yl3, by = list(yl3_yp0), FUN = mean)[ ,2],
                      aggregate(dd$yl3, by = list(yl3_yp1), FUN = mean)[ ,2],
                      aggregate(dd$yl3, by = list(yl3_yp2), FUN = mean)[ ,2]
                      )
                )
yl2 <- rowMeans(cbind(aggregate(dd$yl2, by = list(yl2_yp0), FUN = mean)[ ,2],
                      aggregate(dd$yl2, by = list(yl2_yp1), FUN = mean)[ ,2],
                      aggregate(dd$yl2, by = list(yl2_yp2), FUN = mean)[ ,2]
                      )
                )
yl1 <- rowMeans(cbind(aggregate(dd$yl1, by = list(yl1_yp0), FUN = mean)[ ,2],
                      aggregate(dd$yl1, by = list(yl1_yp1), FUN = mean)[ ,2],
                      aggregate(dd$yl1, by = list(yl1_yp2), FUN = mean)[ ,2]
                      )
                )
yp0 <- rowMeans(cbind(aggregate(dd$y, by = list(yl3_yp0), FUN = mean)[ ,2],
                      aggregate(dd$y, by = list(yl2_yp0), FUN = mean)[ ,2],
                      aggregate(dd$y, by = list(yl1_yp0), FUN = mean)[ ,2]
                      )
                )
yp1 <- rowMeans(cbind(aggregate(dd$yp1, by = list(yl3_yp1), FUN = mean)[ ,2],
                      aggregate(dd$yp1, by = list(yl2_yp1), FUN = mean)[ ,2],
                      aggregate(dd$yp1, by = list(yl1_yp1), FUN = mean)[ ,2]
                      )
                )
yp2 <- rowMeans(cbind(aggregate(dd$yp2, by = list(yl3_yp2), FUN = mean)[ ,2],
                      aggregate(dd$yp2, by = list(yl2_yp2), FUN = mean)[ ,2],
                      aggregate(dd$yp2, by = list(yl1_yp2), FUN = mean)[ ,2]
                      )
                )

# bind all calculated mean vectors into singular table
y_bar <- t(cbind(yl3,yl2,yl1,yp0,yp1,yp2))

# set column names
colnames(y_bar) <- c(5,6,7,10,12,14,15,18,21)

# plot all mean wages over time
mains <- c('T1 to T1',
           'T1 to T2',
           'T1 to T3',
           'T2 to T1',
           'T2 to T2',
           'T2 to T3',
           'T3 to T1',
           'T3 to T2',
           'T3 to T3')
par(mfrow = c(3,3), mar=c(4,4,3,3))
for (i in 1:9) {
  if (i == 4) {
    plot(x = c(-3,-2,-1,0,1,2),
     y = y_bar[,i],
     xlab = '',
     ylab = 'log wage',
     main = mains[i],
     pch = 16
     )
  }
  if (i == 8) {
    plot(x = c(-3,-2,-1,0,1,2),
     y = y_bar[,i],
     xlab = 'event time',
     ylab = '',
     main = mains[i],
     pch = 16
     )
  }
  if (i != 4 & i != 8) {
    plot(x = c(-3,-2,-1,0,1,2),
     y = y_bar[,i],
     xlab = '',
     ylab = '',
     main = mains[i],
     pch = 16
     )
  }
}
```

Table 4

```
# create change in wages response
dd$Dy <- dd$y - dd$yl1

# create change in others' wages predictor
dd$Dwage <- dd$owage2 - dd$owage1

# refresh male- and female-only data frames
ddm <- subset(dd, female == 0)
ddf <- subset(dd, female == 1)

# fit two models using pooled male and female data
m7 <- lm(Dy ~ female + Dwage, data = dd)
```

```
m8 <- lm(Dy ~ I((exp-1)^2) + female + Dwage, data = dd)

# fit separate models for men and women
m9a <- lm(Dy ~ I((exp-1)^2) + Dwage, data = ddm)
m9b <- lm(Dy ~ I((exp-1)^2) + Dwage, data = ddf)

# create table
stargazer(m7, m8, m9a, m9b)
```

Table 4: revised alternative Oaxaca decompositions

```
# fit revised models from Table 3
m6a_new <- lm(I(y - m9a$coefficients['Dwage']*owage2) ~ educ + I(exp^3), data = ddm)
m6b_new <- lm(I(y - m9b$coefficients['Dwage']*owage2) ~ educ + I(exp^3), data = ddf)

# first alternative method

# contribution of difference in education
(mean(ddf$educ) - mean(ddm$educ)) * as.numeric(m6a_new$coefficients['educ']) +
# contribution of difference in experience cubed
(mean(ddf$exp^3) - mean(ddm$exp^3)) * as.numeric(m6a_new$coefficients['I(exp^3)']) +
# contribution of difference in coworkers' wages
(mean(ddf$owage2) - mean(ddm$owage2)) * as.numeric(m9a$coefficients['Dwage']) +
# unexplained difference in education
(mean(ddf$educ)) * as.numeric((m6b_new$coefficients['educ'] - m6a_new$coefficients['educ'])) +
# unexplained difference in experience cubed
(mean(ddf$exp^3)) * as.numeric((m6b_new$coefficients['I(exp^3)'] - m6a_new$coefficients['I(exp^3)'])) +
# unexplained difference in coworkers' wages
(mean(ddf$owage2)) * as.numeric((m9b$coefficients['Dwage'] - m9a$coefficients['Dwage'])) +
# unexplained difference in constant coefficient
as.numeric((m6b_new$coefficients['(Intercept)'] - m6a_new$coefficients['(Intercept)']))
```

```
## [1] -0.2084035
```

```
# second alternative method

# contribution of difference in education
(mean(ddf$educ) - mean(ddm$educ)) * as.numeric(m6b_new$coefficients['educ']) +
# contribution of difference in experience cubed
(mean(ddf$exp^3) - mean(ddm$exp^3)) * as.numeric(m6b_new$coefficients['I(exp^3)']) +
# contribution of difference in coworkers' wages
(mean(ddf$owage2) - mean(ddm$owage2)) * as.numeric(m9b$coefficients['Dwage']) +
# unexplained difference in education
(mean(ddm$educ)) * as.numeric((m6b_new$coefficients['educ'] - m6a_new$coefficients['educ'])) +
# unexplained difference in experience cubed
(mean(ddm$exp^3)) * as.numeric((m6b_new$coefficients['I(exp^3)'] - m6a_new$coefficients['I(exp^3)'])) +
# unexplained difference in coworkers' wages
(mean(ddm$owage2)) * as.numeric((m9b$coefficients['Dwage'] - m9a$coefficients['Dwage'])) +
# unexplained difference in constant coefficient
as.numeric((m6b_new$coefficients['(Intercept)'] - m6a_new$coefficients['(Intercept)']))
```

```
## [1] -0.2084035
```

Part II

Figure 3

```r
# divide PSU into bins of size six
psu_bins <- cut(rd$psu,
                breaks = 80,
                ordered_result = TRUE
                )

# aggregate by bin, find mean scores and mean rate of college entry for each bin
bin_means <- aggregate(rd, by = list(psu_bins), FUN = mean)

# plot mean rate of college entry against mean PSU score for each bin
ggplot(bin_means, aes(x = psu, y = entercollege)) +
  geom_point() +
  theme_minimal() +
  geom_vline(aes(xintercept = 475), col = 'red')
```

Table 5

```r
# create respective variables and run regressions
rd$x <- rd$psu - 475
rd$z <- ifelse(rd$psu >= 475, 1, 0)
rd$D <- rd$entercollege
lst <- list(0,0,0,0)
i <- 1
for (B in seq(25,100,25)) {
  lst[[i]] <- lm(D ~ z*x, data = subset(rd, rd$psu >= 475 - B & rd$psu <= 475 + B))
  i <- i + 1
}

# create table
stargazer(lst)
```

Figure 4

```r
# estimate previous regression using a range of bandwidths
pi <- c()
se <- c()
li <- c()
ui <- c()
i <- 1
for (B in seq(25,200,5)) {
  model <- lm(D ~ z*x, data = subset(rd, rd$psu >= 475 - B & rd$psu <= 475 + B))
  pi[i] <- model$coefficients[2]
  se[i] <- coef(summary(model))[, 2][2]
  li[i] <- pi[i] - 2 * se[i]
  ui[i] <- pi[i] + 2 * se[i]
  i <- i + 1
}

# plot coefficient estimates against bandwidth values
```

```
B <- seq(25,200,5)
df1 <- data.frame(B,pi,li,ui)
ggplot(df1, aes(x = B)) +
  geom_line(aes(y = pi)) +
  geom_line(aes(y = li),
            color = 'red',
            linetype = 'dashed'
            ) +
  geom_line(aes(y = ui),
            color = 'red',
            linetype = 'dashed'
            ) +
  labs(title = TeX('$\\1\\left[PSU\\geq 475\\right]$ coefficient estimates against bandwidth $25\\leq B\
       x = 'Bandwidth',
       y = TeX('$\\pi_1$')) +
  theme_minimal()
```

Table 6

```
# family income in quintile 1
rd$q1 <- rd$quintile == 1

# family income in quintile 2
rd$q2 <- rd$quintile == 2

# family income in quintile 3
rd$q3 <- rd$quintile == 3

# family income in quintile 4
rd$q4 <- rd$quintile == 4

# GPA between 60 and 70
rd$gpa_high <- rd$gpa >= 60 & rd$gpa <= 70

# GPA between 50 and 60
rd$gpa_mid <- rd$gpa >= 50 & rd$gpa < 60

# GPA less than 50
rd$gpa_low <- rd$gpa < 50

# complier means
comp_means <- c()
i <- 1
for (j in c(13,14,15,16,2,17,18,19,7,6)) {
  comp_means[i] <- as.numeric(lm(I(subset(rd, rd$psu >= 400 & rd$psu <= 550)[ ,j]*D) ~ D + x + x:z,
                                 data = subset(rd, rd$psu >= 400 & rd$psu <= 550))$coefficients['D']
                  )
  i <- i + 1
}

# find means of characteristics for entire sample
sample_means <- apply(rd[ ,c(13,14,15,16,2,17,18,19,7,6)], 2, 'mean')
```

```
# find means of characteristics for sudents within bandwidth of 75
bandwidth_means <- apply(subset(rd, rd$psu >= 400 & rd$psu <= 550)[ ,c(13,14,15,16,2,17,18,19,7,6)],
                         2,
                         'mean'
                         )

# find ratio of complier means against sample means
ratios <- comp_means / sample_means

# combine characteristics into table
t6 <- cbind(sample_means,
             bandwidth_means,
             comp_means,
             ratios
             )
rownames(t6) <- c('Q1', 'Q2', 'Q3', 'Q4', 'female', 'GPA high', 'GPA mid', 'GPA low',
                  'mother > HS', 'father > HS'
                  )
colnames(t6) <- c('sample', 'B = 75', 'compliers', 'complier-sample ratio')

# create table
stargazer(t6, digits = 4)
```