Final Class Project

Wages are a necessary evil for all economists to get comfortable with. They are easily measurable, but fraught with subtleties and unobserved characteristics that even the most accomplished econometrician cannot correctly avoid. Regardless, this analysis will attempt to do just that – analyze the wages of men and women from Country X in order to document and explore any potential gender wage gap. The **first** avenue of exploration is basic characteristics of the data set, and how it varies for men and women.

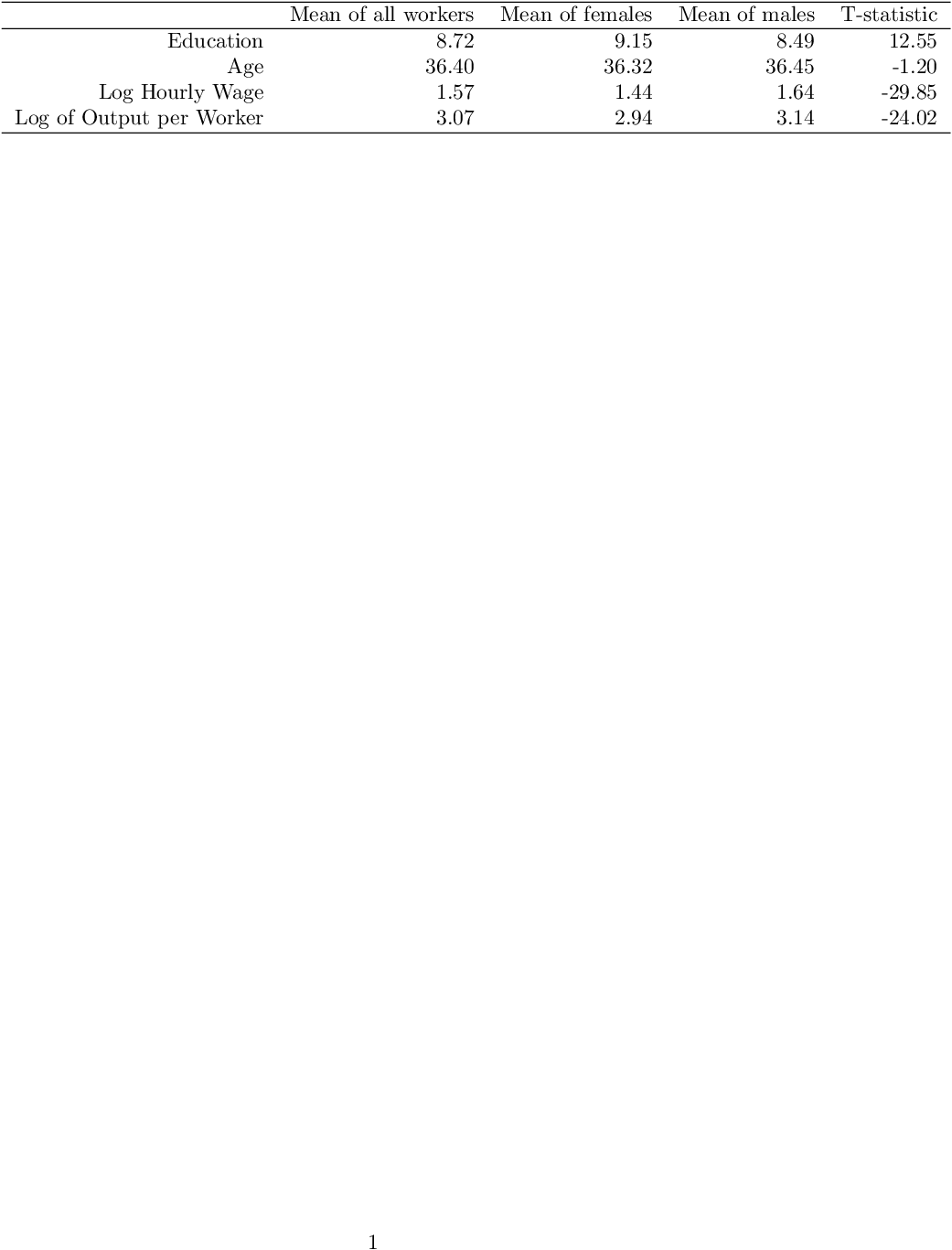


Table 1 – Mean Characteristics of Workers

This table documents the means of Education, Age, Wage and Productivity for everyone, males and females. We can see that on average, females have a higher education level than males yet women still earn a lower wage (on average). The average ages are similar in our dataset, while it is interesting to note that the average productivity of females is lower than that of males.

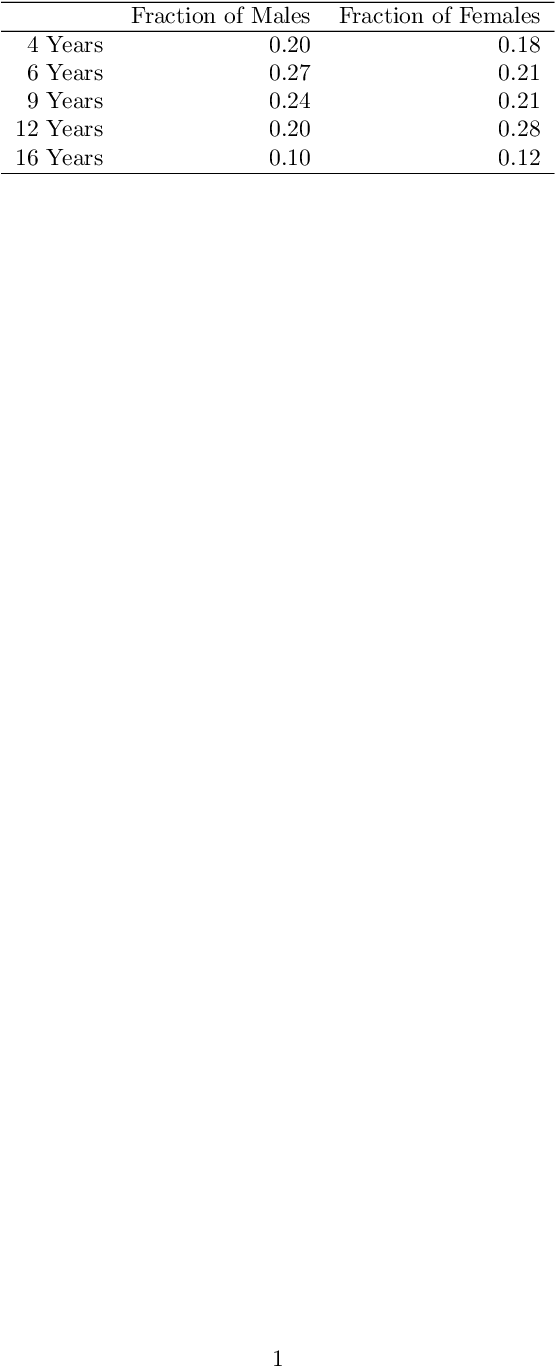


Table 1a – Fraction of Workers by Education Level

To go along with female’s highest average education level, we see a higher proportion of women have 12 and 16 Years of education, while men have a higher proportion of 4, 6 and 9 Years of education.

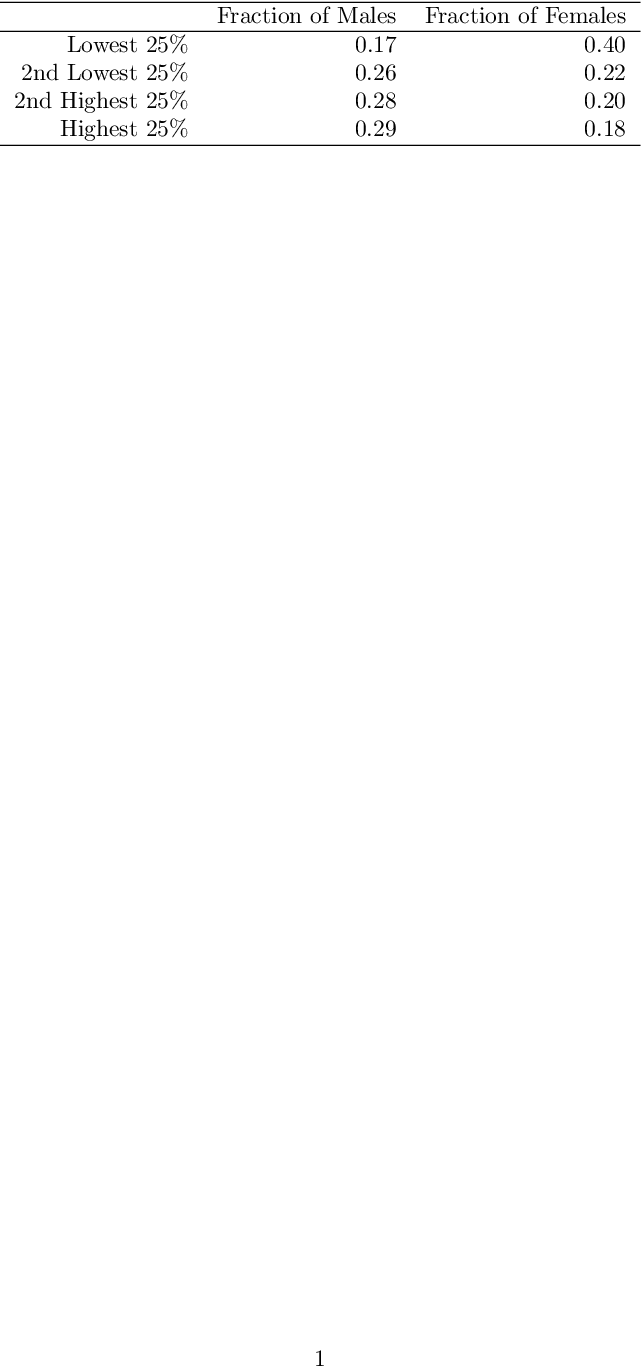


Table 1b – Fraction of Workers by Wage Quartile

Here, we see an interesting trend that helps explain women’s lower average wage – there are 40% of women who are in the overall lowest quartile of earnings. If we spread those 40% out over the other 3 quartiles, we would get a more even distribution of women’s wages that would push up their average earnings to be on par with men.

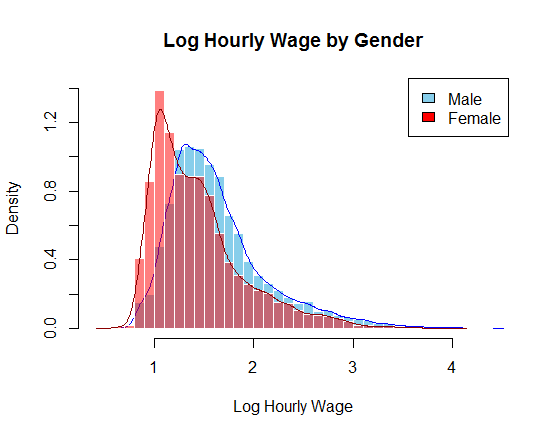


Figure 1a – Wage Distribution by Gender

This histogram helps exemplify how women’s wages are often lower than men’s. The mean log wage is 1.57. We see that roughly below 1.57, a women has a higher probability of earning a specific wage than a man (corroborated by the wage quantile fractions), while above 1.57 mean have a consistently higher probability of earning a specific wage than women.

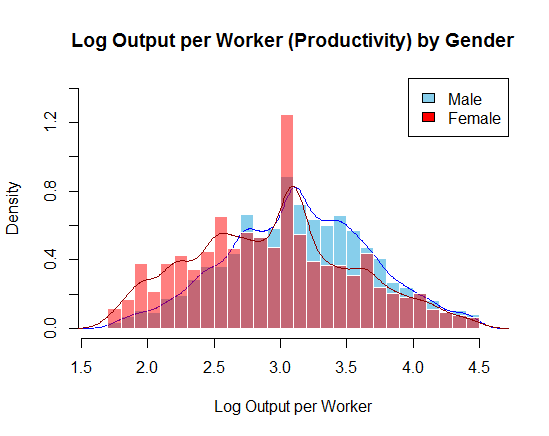


Figure 1b – Productivity Distribution by Gender

This histogram indicates that females tend to work at less productive firms than men. We see females have a higher probability of output levels <2.75, while males have a higher probability of output levels >2.75.

Overall, we see that while women are often more educated than men, they still earn less and tend to work at firms with lower productivity.

Our **second** avenue of exploration will be running some standard wage models (S1 – S4) to construct Oaxaca Decompositions of the wage gap between genders.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  | *Dependent variable:* | | | |
|  |  | | | |
|  | y (Log Hourly Wage) | | | |
|  | S1 | S2 | S3 | S4 |
|  | | | | |
| Education |  | 0.083\*\*\* | 0.081\*\*\* | 0.086\*\*\* |
|  |  | (0.001) | (0.001) | (0.001) |
|  |  |  |  |  |
| Age |  | 0.095\*\*\* | 0.110\*\*\* | 0.064\*\*\* |
|  |  | (0.015) | (0.020) | (0.024) |
|  |  |  |  |  |
| Age^2 |  | -0.002\*\*\* | -0.002\*\*\* | -0.001 |
|  |  | (0.0004) | (0.001) | (0.001) |
|  |  |  |  |  |
| Age^3 |  | 0.00001\*\* | 0.00001\*\* | 0.00000 |
|  |  | (0.00000) | (0.00000) | (0.00001) |
|  |  |  |  |  |
| Female | -0.197\*\*\* | -0.251\*\*\* |  |  |
|  | (0.007) | (0.005) |  |  |
|  |  |  |  |  |
| Constant | 1.636\*\*\* | -0.859\*\*\* | -1.065\*\*\* | -0.675\*\* |
|  | (0.004) | (0.189) | (0.244) | (0.295) |
|  |  |  |  |  |
|  | | | | |
| Observations | 23,144 | 23,144 | 15,058 | 8,086 |
| R2 | 0.036 | 0.400 | 0.349 | 0.442 |
| Adjusted R2 | 0.036 | 0.400 | 0.349 | 0.441 |
| Residual Std. Error | 0.489 (df = 23142) | 0.386 (df = 23138) | 0.403 (df = 15053) | 0.351 (df = 8081) |
| F Statistic | 858.016\*\*\* (df = 1; 23142) | 3,089.985\*\*\* (df = 5; 23138) | 2,019.745\*\*\* (df = 4; 15053) | 1,597.209\*\*\* (df = 4; 8081) |
|  | | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | |

Table 2 – Standard Wage Regression Models (Including Age, Education)

We see that the female coefficient in S1 is -0.197, while the female coefficient in S2 is -0.251. This makes sense, as once we include the age coefficients into the regression, we remove a part of the omitted variable bias in S1 that underestimates the true effect of gender on wages.

The education coefficient in S3 is 0.081, while it is 0.086 in S4. There is not a great difference between the two, and is likely negligible. It is interesting to note that women are often more educated than men yet experience the same returns to their wage level. The age coefficients in S3 are 0.110, -0.002 and 0.00001 while the coefficients in S4 are 0.064, -0.001 and 0.00000. Clearly, women experience lower returns to age and work experience than men. One possible reason is that women often take breaks during their career for family commitments, and so they get older without advancing their career and improving their wage.

Using S3 and S4, one can construct an Oaxaca decomposition of the gender wage gap. We get:

The gender wage gap is 0.2. We see that Education contributes somewhere around to to the gender wage gap, while Age contributes somewhere around to to the gender wage gap. This indicates that Education actually closes the gender wage gap, while mean Age is not a large contributor to the wage gap. The Education story makes sense, as we discovered females have slightly higher returns to education *and* have a higher average level of education, so this should push their wages up relative to men. The Age story is a bit odd – it does not explain much of the gender wage gap, but we found wildly different Age coefficients for both genders. It is likely that since both males and females have a similar average Age, it does not contribute enough information by itself to explain the wage gap. The contribution of the Education coefficients is roughly, which shows women’s higher returns to education slightly close the wage gap. The contribution of the Age coefficients is roughly 1.67, which means that men’s higher returns to age contribute a *significant* amount to the gender wage gap.

An exploration of the age profiles of both genders would help glean some more insight into this phenomenon.

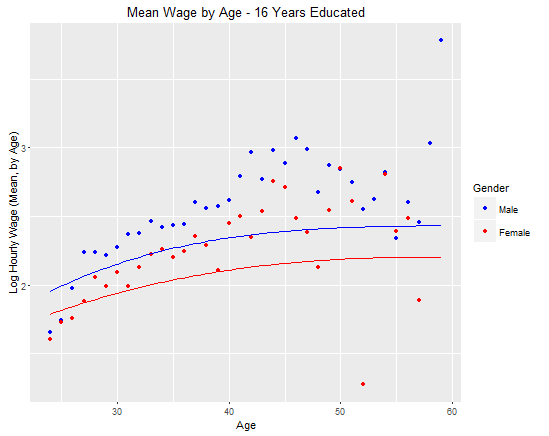
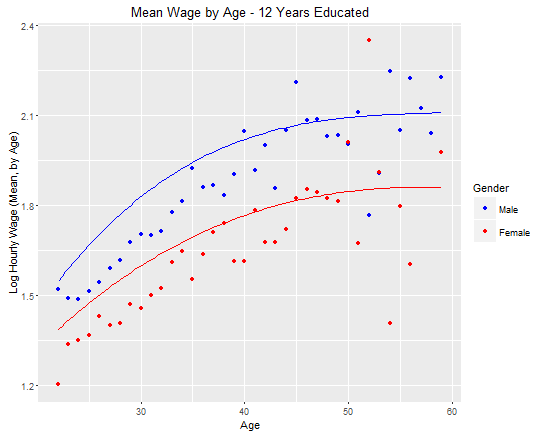
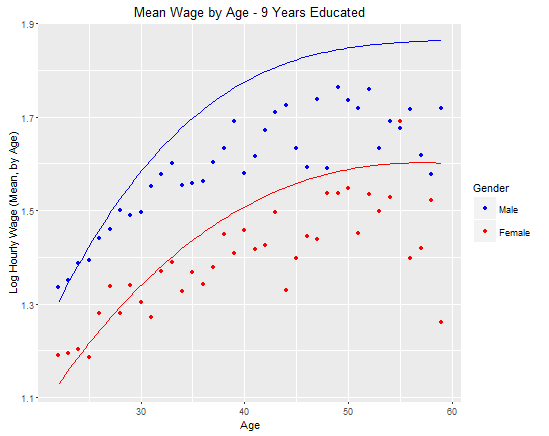
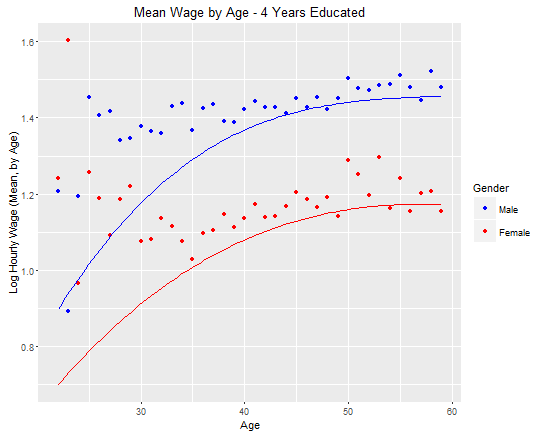


Figure 2 – Mean Wage vs Age by Education Level

All 5 panels have a similar shape: wage peaks quickly in your 20s, and then plateaus slowly as you get older. It is interesting to note that the average wage gap between genders for each age stays roughly the same – women never ‘catch up’ to men’s earnings as they get older.

A cubic function may not be the best way to describe the age profiles by gender. Another approach could be using cubic splines instead. Taking the data for workers with Education level of 12 Years, we estimate a cubic spline with knots K = 4,5,6…, 10 using a 2-sample cross-validation method. Unfortunately, the case for K = 3 returned an error, and was omitted from the analysis.

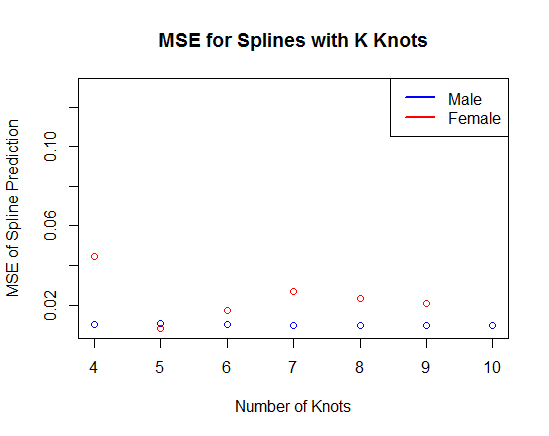


Figure 3 – MSE for Cubic Spline with K Knots, Both Genders

For reference, the MSE for males using the cubic function is 0.0132, while the MSE for females using the cubic polynomial is 0.0109. We can see that the optimal number of knots (yielding the lowest MSE) is K = 10 for males and K = 5 for females, which yield an MSE of 0.0097 and 0.0082 respectively. Note that the MSEs under the best case cubic spline is lower than the MSE under cubic polynomial fits. Refer to the graphs below, where purple is the cubic spline fit.

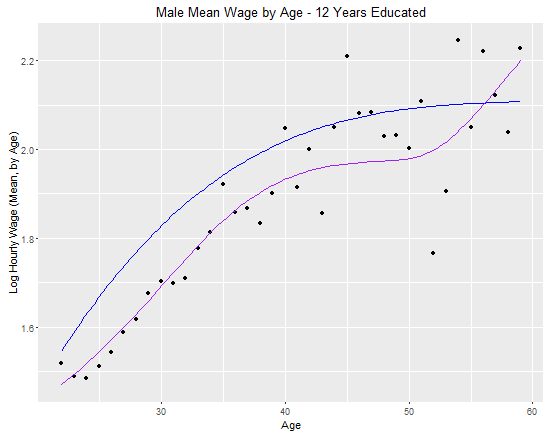
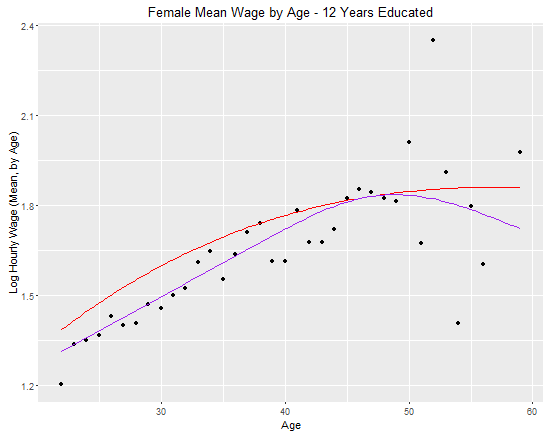
 

Figure 4 – Mean Wage by Age for 12 Years Educated – Polynomial vs Spline Fit

The **third** avenue of exploration will be regressions like S1-S4, but with productivity (va) added.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Dependent variable:* | | | |
|  |  | | | |
|  | y (Log Hourly Wage) | | | |
|  | M1 | M2 | M3 | M4 |
|  | | | | |
| Education | 0.083\*\*\* | 0.071\*\*\* | 0.071\*\*\* | 0.072\*\*\* |
|  | (0.001) | (0.001) | (0.001) | (0.001) |
|  |  |  |  |  |
| Age | 0.095\*\*\* | 0.093\*\*\* | 0.099\*\*\* | 0.076\*\*\* |
|  | (0.015) | (0.014) | (0.019) | (0.023) |
|  |  |  |  |  |
| Age^2 | -0.002\*\*\* | -0.002\*\*\* | -0.002\*\*\* | -0.001\*\* |
|  | (0.0004) | (0.0004) | (0.0005) | (0.001) |
|  |  |  |  |  |
| Age^3 | 0.00001\*\* | 0.00001\*\*\* | 0.00001\*\* | 0.00001 |
|  | (0.00000) | (0.00000) | (0.00000) | (0.00000) |
|  |  |  |  |  |
| Female | -0.251\*\*\* | -0.197\*\*\* |  |  |
|  | (0.005) | (0.005) |  |  |
|  |  |  |  |  |
| Employer |  | 0.233\*\*\* | 0.254\*\*\* | 0.198\*\*\* |
| Productivity |  | (0.004) | (0.006) | (0.007) |
|  |  |  |  |  |
| Constant | -0.859\*\*\* | -1.427\*\*\* | -1.616\*\*\* | -1.254\*\*\* |
|  | (0.189) | (0.179) | (0.230) | (0.281) |
|  | | | | |
| Observations | 23,144 | 23,144 | 15,058 | 8,086 |
| R2 | 0.400 | 0.464 | 0.421 | 0.495 |
| Adjusted R2 | 0.400 | 0.463 | 0.421 | 0.495 |
| Residual Std. Error | 0.386 (df = 23138) | 0.365 (df = 23137) | 0.380 (df = 15052) | 0.333 (df = 8080) |
| F Statistic | 3,089.985\*\*\* (df = 5; 23138) | 3,332.956\*\*\* (df = 6; 23137) | 2,191.070\*\*\* (df = 5; 15052) | 1,586.333\*\*\* (df = 5; 8080) |
|  | | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | |

Table 3 –Wage Regression Models (Age, Education) Including Productivity

It is interesting to note that women experience lower returns to productivity (0.198) than men (0.254). Let us run an Oaxaca decomposition based on model M2. Here, the data is for both genders pooled together, so we use the simple version of the Oaxaca when one assumes the estimated variable coefficients are the same for both groups.

Given this, we see that the difference in firm productivity by gender contributes to the gender wage gap, which is a decent amount.

If we consider a causal wage model where part of the ‘residual’ is due to an unobserved characteristic that makes workers more or less productive, then this ‘fixed effect’ will bias the validity of our OLS assumptions as now the residual is correlated to a regressor. We expect that a worker’s productivity is positively related to their wage. Similarly, we expect that higher productivity firms will pay out higher wages to their workers. Hence in this model, if more productive employers hire workers with higher productivity, the fixed effect will get ‘mixed’ into the OLS estimate of , and will be higher than the true causal effect of working at a more profitable employer. In other words, the OLS estimate is accounting for both worker and firm productivity, which will bias it upwards from the true estimate of returns to just employer productivity.

Let us now run Oaxaca decompositions using M3 and M4.

From these two decompositions, we see that the contributions of the education and age covariates, and the contributions of the coefficients on these variables, have remained somewhat similar. We see that the contribution of the difference in means of gender productivity contributes around to to the gender wage gap – similar to what we discovered in M2. Moreover, the difference in gender returns to productivity contributes around to to the gender wage gap, which is extremely large!

Let us now redo the decompositions by renormalizing va. We will achieve this by subtracting , the minimum value of va in the data, from all the relevant data points (i.e. subtracting 1.7 from the va means for each gender). Then, the Oaxaca decomposition is:

Now, the difference in gender returns to productivity contributes around to to the gender wage gap, which is much smaller than the previous result. Hence, we must be cautious in throwing too much weight to this result. One thing is clear – women working at a more productive firm has a smaller effect on their wage than for men.

The **fourth** avenue of exploration will be running a regression based on first differences.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | | |
|  | *Dependent variable:* | | | |
|  |  | | | |
|  | dy (Change in Wage from Period -1 [Old Job] to Period 0 [New Job]) | | | |
|  | C1 | C2 | C3 | C4 |
|  | | | | |
| Age | -0.010\*\*\* | -0.010\*\*\* | -0.010\*\*\* | -0.010\*\*\* |
|  | (0.002) | (0.002) | (0.002) | (0.003) |
|  |  |  |  |  |
| Age^2 | 0.0001\*\*\* | 0.0001\*\*\* | 0.0001\*\*\* | 0.0001\*\*\* |
|  | (0.00002) | (0.00002) | (0.00003) | (0.00003) |
|  |  |  |  |  |
| Change in | 0.076\*\*\* | 0.076\*\*\* | 0.083\*\*\* | 0.062\*\*\* |
| Employer Productivity | (0.003) | (0.003) | (0.004) | (0.004) |
|  |  |  |  |  |
| Female |  | -0.013\*\*\* |  |  |
|  |  | (0.004) |  |  |
|  |  |  |  |  |
| Constant | 0.258\*\*\* | 0.262\*\*\* | 0.276\*\*\* | 0.227\*\*\* |
|  | (0.034) | (0.034) | (0.044) | (0.050) |
|  |  |  |  |  |
|  | | | | |
| Observations | 23,144 | 23,144 | 15,058 | 8,086 |
| R2 | 0.035 | 0.035 | 0.037 | 0.031 |
| Adjusted R2 | 0.035 | 0.035 | 0.037 | 0.030 |
| Residual Std. Error | 0.277 (df = 23140) | 0.277 (df = 23139) | 0.295 (df = 15054) | 0.240 (df = 8082) |
| F Statistic | 279.502\*\*\* (df = 3; 23140) | 212.483\*\*\* (df = 4; 23139) | 195.142\*\*\* (df = 3; 15054) | 84.819\*\*\* (df = 3; 8082) |
|  | | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | | |

Table 4 –First Differences Wage Regression Models (Age) Including Change in Productivity

The most striking feature of these first differences regressions is that the estimated coefficient on dva is positive and *smaller* than the estimated coefficients on va in models M1 – M4. We calculate coefficients of 0.076, 0.076, 0.083 and 0.062 on dva, which are much smaller than the coefficients 0.233, 0.254, 0.198 calculated for va in M1-M4.

This corroborates the idea that there is some unmeasurable fixed effect hiding in the residual that is correlated to the productivity variable. By taking the first-differences version of the causal model we have estimated, we eliminate the fixed effect that does not vary across time and hence remove the overestimating bias it induces on the coefficient for Productivity.

An interesting feature to note is that the coefficient for dva for females is 0.062, which is smaller than the coefficient for dva for males (0.083). Some possible reasons for these lower returns to productivity for women could include workplace discrimination, taking breaks from careers to raise a family, or the common occurrence of some women working part-time.

For our **fifth** avenue of exploration, we shall conduct an event study of wages as workers move between different groups of employers.

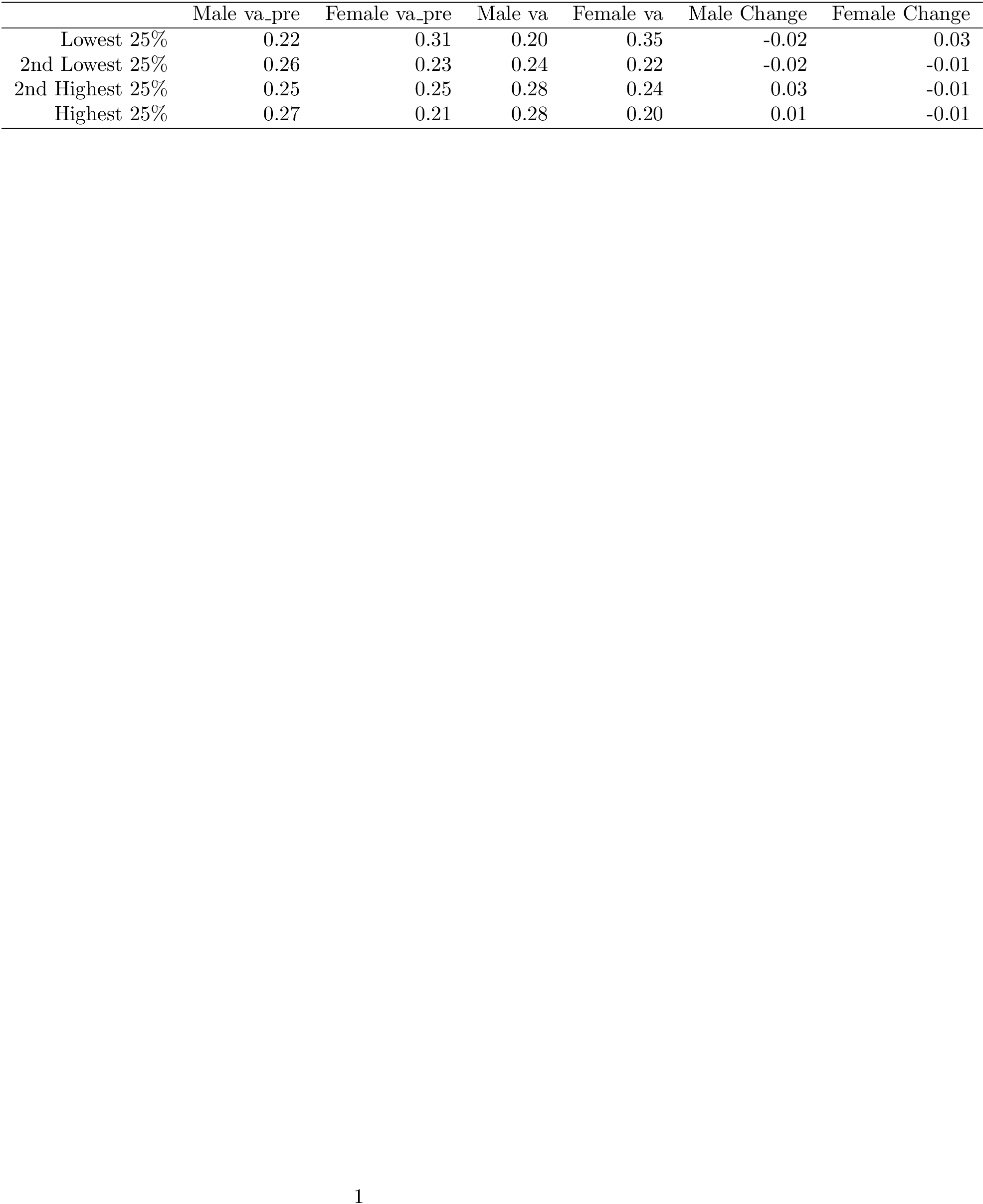


Table 5 –Proportion of Workers in each Employer Productivity Quartile (for Old and New Job)

We see that men are reasonably distributed across firms in both their old and new job. Unsurprisingly, the average change for a male is to move to a higher productivity firm. For women, the results are quite intriguing. Women occupy a large percentage of the bottom 25% of firms, and on average move *down* to these low productivity firms for their new job.

Let us now consider wage profiles of workers who start in a specified origin quartile and end up in a new destination quartile of firms ranked by productivity. If employer productivity has a positive causal effect on wages, we expect that a worker who moves from a firm in Quartile 1 to Quartile 4 will have a higher new wage, while a worker who moves from a Quartile 4 to a Quartile 1 firm will experience a pay cut.

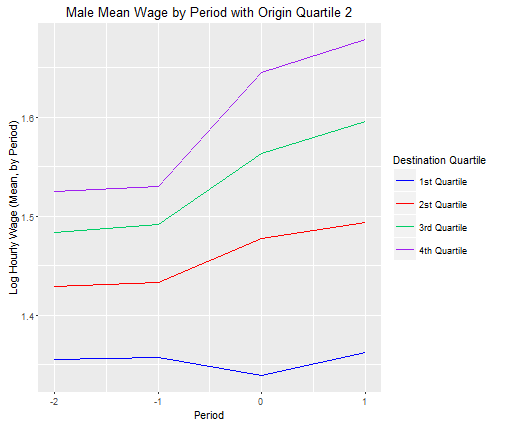
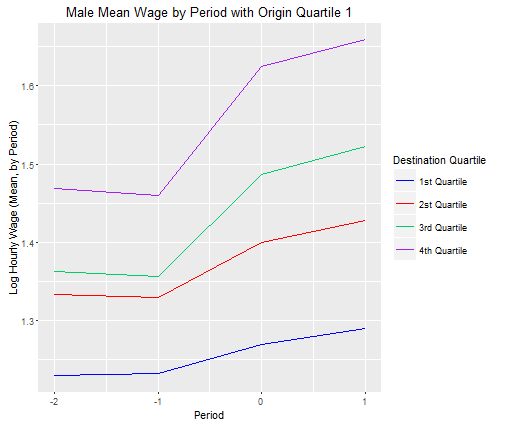
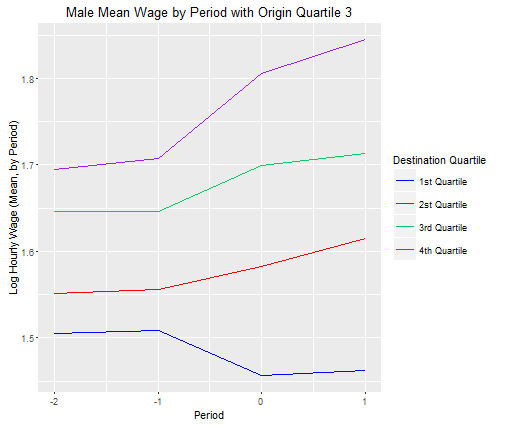
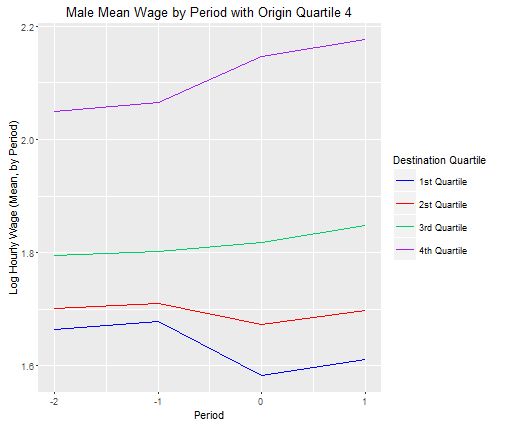
  

Figure 5 –Wage Event Time Study of Male Workers with Productivity Origin Quartile 1-4 and Destination Quartile 1-4

Let us examine men first in Figure 5. The results seem to corroborate the idea of a positive causal effect of employer productivity. If a male starts in the first Quartile, he experiences a wage boost at his new job no matter what, especially from period -1 to period 0 when he switches jobs. If a male starts in Quartile 2 or Quartile 3, he takes a pay cut at first when moving to a Quartile 1 firm. The results are starkest for a male starting in Quartile 4, whose wage drops if he moves to a Quartile 1 or 2 firm. It is interesting to note that even before a worker gets to a higher Destination quartile firm, they already have a higher wage differential than someone who is destinted to go to a lower Destination quartile firm. Possibly a worker, even if at a low Origin Quartile firm, is already ‘set’ to go to a high Destination quartile firm due to high education level, special skills or a pre-determined career path. To reward these attributes, this worker is paid higher than a worker with the same Origin quartile but lower Destination Quartile. As expected, if a worker makes a higher move up in quartiles, his wage change (as seen by the slope of the line) is steeper. In order words, the slope (or change) of a male worker’s wage is positivley correlated with the destination quartile – this can be seen in all 4 panels. Interestingly, origin quartile only really matters if your origin quartile is the top 25% of firms. Otherwise, it is mostly your destination quartile which dictates your starting wage and your wage trajectory.

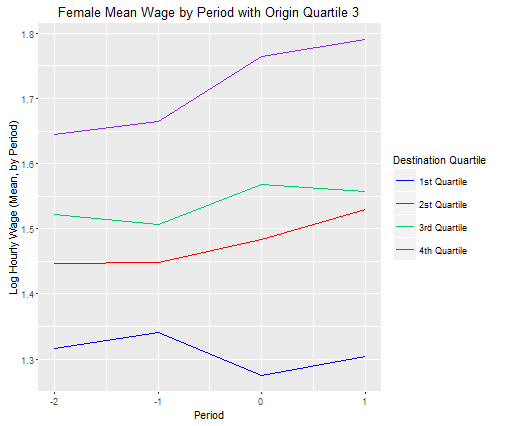
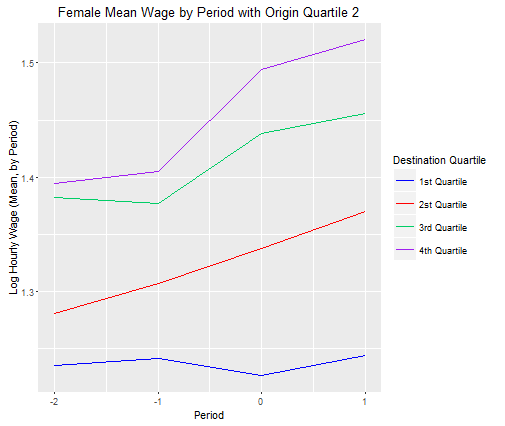
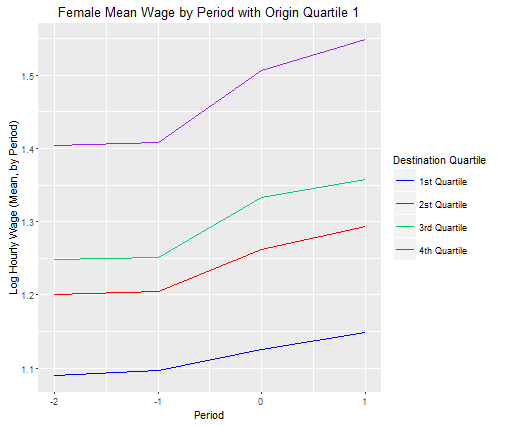
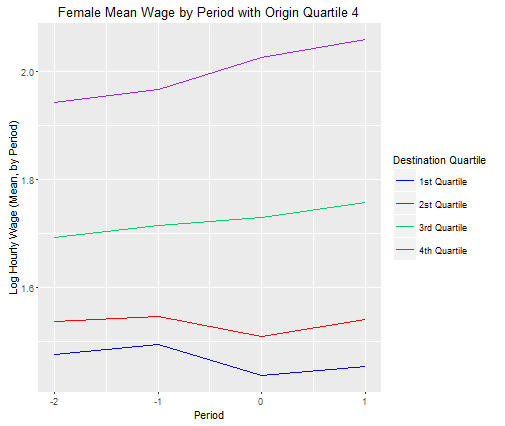
 

Figure 6 –Wage Event Time Study of Female Workers with Productivity Origin Quartile 1-4 and Destination Quartile 1-4

The women in Figure 6 tell a similar story of a positive causal effect of employer productivity. The shape of all 4 panels are eerily similar to their male counterparts – the slope of wage from period -1 to period 0 is higher for a higher destination quartile. In every panel, moving to a higher Quartile firm does increase your wage. Moving to a lower Quartile firm causes your wage to initially dip, but then recover as you gain more experience on the job. The main difference for females is that the effect of moving is less pronounced. In general, women start out with slightly lower wages in period -2, and then their wages increase by less than their male counterparts. Hence a move to a more productive employer benefits both genders, but it benefits men more than women.

For our **sixth** avenue of exploration, we will run a special regression using the causal effect of productivity we estimated in C3-C4 to modify the regressions M3-M4. Specifically, we will crease a special set of data , and use the remaining variables in M3/M4 to regress onto this special variable. This differs from just running M3 and M4 like before, because we are going to specify a more accurate estimate of using our first-differences model estimates from C3-C4.

|  |  |  |
| --- | --- | --- |
|  | | |
|  | *Dependent variable:* | |
|  |  | |
|  |  | |
|  | N1 | N2 |
|  | | |
| Education | 0.082\*\*\* | 0.085\*\*\* |
|  | (0.001) | (0.001) |
|  |  |  |
| Age | 0.110\*\*\* | 0.068\*\*\* |
|  | (0.020) | (0.024) |
|  |  |  |
| Age^2 | -0.002\*\*\* | -0.001 |
|  | (0.001) | (0.001) |
|  |  |  |
| Age^3 | 0.00001\*\* | 0.00000 |
|  | (0.00000) | (0.00001) |
|  |  |  |
| Constant | -1.090\*\*\* | -0.719\*\* |
|  | (0.245) | (0.297) |
|  |  |  |
|  | | |
| Observations | 15,058 | 8,086 |
| R2 | 0.350 | 0.436 |
| Adjusted R2 | 0.350 | 0.436 |
| Residual Std. Error | 0.406 (df = 15053) | 0.352 (df = 8081) |
| F Statistic | 2,023.633\*\*\* (df = 4; 15053) | 1,561.205\*\*\* (df = 4; 8081) |
|  | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | |

Figure 6 –‘Special’ Wage Regression Models (Age, Education)

Now, let us create Oaxaca decompositions from the N1-N2 models.

From this, we estimate the difference in mean value of va contributes to to the gender wage gap. Meanwhile, the difference in returns to working at a higher productivity firm contribute to to the gender wage gap.

In conclusion, we have explored the various extents to which age, education and firm productivity can influence a worker’s wage, and how these effects may differ by gender. Regarding education, we discovered that women are more educated on average, and experience a higher return to education than males. In other words, education actually helps close the gender wage gap, so the onus lies on other variables. Regarding age, which can be seen as a proxy for experience, women are roughly the same age as men, but experience a much smaller return for higher ages than men. Essentially, becoming older rewards men much higher than women, and this phenomenon contributes a significant amount to the gender wage gap.

Regarding productivity, we discovered that women have a slight tendency to move ‘down’ firms, compared to men who move up – likely due to taking time off for family or becoming part-time. After establishing that working at a higher productivity firm should command a higher wage, we discovered that women often working at less productive firms did not contribute much to the wage gap, but the difference in returns to firm productivity for both genders contributed significantly to the wage gap – possibly due to workplace discrimination. Using splines, regressions, Oaxaca decompositions and more, we have come one step closer towards understanding the gender wage gap.