Data Analysis 2 - Assignment 1 - Haaris Cheema

Introduction and Exploratory Data Analysis

The aim of this report is to analyze the wage gap between male and female finance specialists in the US in 2014. Multiple occupations have been clubbed together under the umbrella of financial specialists. The level of education has been restricted to four categories, namely bachelors, masters, professional degrees (Prof) and PHDs.

The preliminary analysis consists of analyzing the distribution of key variables such as weekly earnings, usual work hours, weekly earning per hour and the level of education. Our filtered data contains 52% females (please see appendix). Weekly earning per hour was a more meaningful indicator to analyze the wage gap and hence we created this variable (w). The summary table indicates that it has right-tailed distribution (please also see fig.). This will be a key consideration when we choose the most appropriate regression models.

Regressions, Interpretations and Findings

We modelled different options in our regressions and the outputs can be seen (fig.). To model the unconditional wage gap, we tested the log-level as well as the level-level model. The regression output is shown (fig.). The first model's beta coefficient indicates that females earn, on average, 19 log points (19%) lesser than males,per hour. The second model's coefficient indicates that females earn, on average, 6.47 dollars lesser than males per, hour. Gender is statistically significant even when tested at a level of significance of 0.001. These results also show that the r-squared does not vary a lot between level-level and log-level models, and considering the skew in the dependent variable, we will proceed with the log-level model.

We performed two additional regressions as we analyzed the gender wage gap conditioned upon the level of education. The regression outputs have been synthesized in fig.. We can see from the second regression (column (2)), that comparing employees of the same gender, those individuals with a professional degree earn, on average, 24% more than those with a bachelors degree. Similarly, those with a Masters degree and those with a PHD, tend to earn, on average, 12.8% and 50.6% more than those with a bachelors degree, respectively. The coefficient on female is smaller when education is included in the regression (comparing column 1 and 2). This suggests that part of the gender wage gap is due to the fact that women are somewhat more likely to be in the lower earner BA group (our reference group), as opposed to higher earning PHDs for instance.

The third regression (column (3)) indicates that when the gender wage gap is conditioned upon age as well, the coefficient on female becomes larger. Like in the case of the second regression, all variables are statistically significant at a level of significance of 0.01. However, in the third regression, the R-squared value increases to 9%. Therefore in this particular analysis, this regression seems to be the most appropriate one to analyze wage gap conditioned one education.

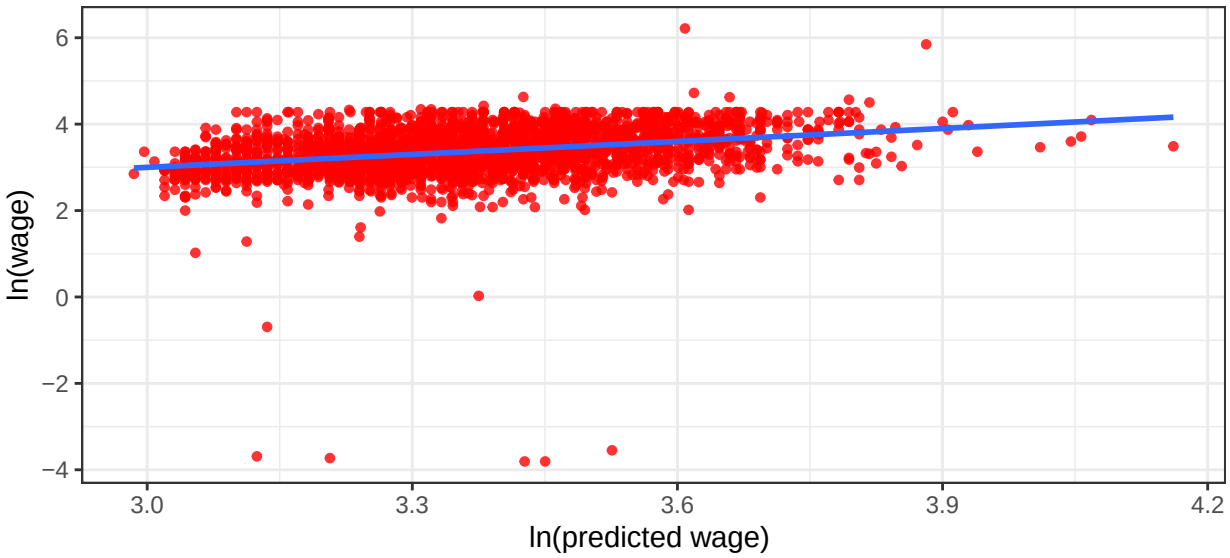
To solidify our model choice, we have plotted the predicted values against the actual values as well as the residuals. (see fig.). The predicted values of the log of wage were plotted against the actual log of wage. We can verify the accuracy of these predictions as there seems to be a strong correlation between the predicted values and the actual values. Also, when we plotted the predicted values against the residuals, we saw that a majority of the values are clustered around 0, which indicates a strong accuracy of the predicted values.This shows that our choice regarding the regression model is an appropriate one.

	Mean	Median	SD	Min	Max	P95	N
earnwke	1373.75	1173.07	710.54	1.00	2884.61	2884.61	2449
uhours	41.68	40.00	7.58	1.00	92.00	55.00	2449
w	33.10	28.85	19.56	0.02	500.00	64.10	2449
grade92	43.28	43.00	0.53	43.00	46.00	44.00	2449

Appendix

	(1)	(2)
(Intercept)	3.465 ***	36.531 ***
	(0.018)	(0.636)
female	-0.190 ***	-6.471 ***
	(0.024)	(0.792)
N	2449	2449
R2	0.026	0.027

*** p < 0.001; ** p < 0.01; * p < 0.05.



	(1)	(2)	(3)
(Intercept)	3.465 ***	3.422 ***	2.952 ***
	(0.018)	(0.021)	(0.048)
female	-0.190 ***	-0.179 ***	-0.187 ***
	(0.024)	(0.024)	(0.023)
ed_Profess		0.244 ***	0.218 **
		(0.073)	(0.073)
ed_MA		0.128 ***	0.123 ***
		(0.028)	(0.027)
ed_PHD		0.506 ***	0.491 ***
		(0.139)	(0.140)
age			0.012 ***
			(0.001)
N	2449	2449	2449
R2	0.026	0.041	0.090

*** p < 0.001; ** p < 0.01; * p < 0.05.