Data Analysis: Assignment: 1

Ghazal Ayobi and Shah Ali Gardezi

Introduction

For this assignment we use Current Population Survey (CPS) which can be accessed **here**. The occupation that we have selected for this exercise is the ***Accountants and Auditor***, with census occupation code ***0800***.

Data Transformation

As a process of filtering and transforming as we created new variables such as ***female*** which is assigned a binary value of ***1***, hourly wage_**(w)**_ which is calculated by dividing the weekly earnings (earnwke) by the number of hours (uhours), and log of wage ***(lnw)***. We also created a character variable called ***gender*** and using sex variable for MALE and FEMALE. Our sample is varied with high number of females (992) compared to men (573) as shown in Table 1. Descriptive summary of the main variables in our dataset can be found Table 2. From the table we infer that because of the presence of high hourly wage ***(w)*** values like USD 346 per hour, mean tends to be to the right of the median thus making the sample distribution rightly skewed. Similarly, we see that there are certain people who work more than 40 hours (maximum value of 92 hours a week) which is also the cause for skewness. Moreover, there is also the presence of extreme values. For example, the minimum wage value is computed out to be USD 0.01 which is highly unlikely in USA thus we excluded the data points for wage value below USD 1.

Analysis

Before we began our regression, we visualized the wage distribution using ggplot and found that the hourly wage (w) is rightly skewed. Thus, we will be using log of wage ***ln(w)*** for our regression analysis. The distribution curve for ln(w) is shown in Figure 2. We now begin regression analysis in order to find out whether there exists a wage gap based on gender. Table 3 summarizes the results of regression for unconditional gender gap. There are two regressions, the first regression is level-level regression which shows that females on average tend to earn USD 5.5 less than their male counterparts. This wage gap is significant at 1% significant level. The second regression is log-level regression which show that females on average earn 18% less than males and this coefficient is significant with more than 99.9% confidence level.

Our next aim is to find out wage gap based on different levels of education (description of different education level are shown in the Appendix). Table 4 summarizes the result of the multivariate regressions. To start off the model 2 of Table 4 we conditioned gender gap on education and found that females on average earn 14.7 % less than males and this coefficient is significant at 1% significant level. We then proceeded to uncover the wage disparity, based on different levels of education. Model 3, shows the results of comparing employees of same gender with different education levels, using Associate-vocational degree as the baseline variable. We select values with more than 99 % confidence level, for example in the same gender employees with Bachelor degree tend to earn on average 33.3% more than employees with Associate- vocational degree.

To gain a deeper understanding, we now run a regression with interaction terms using the same base variable. In the Table 5, the ALL regression column reveals, that even for women with higher education level, there exists a gender gap with 99% confidence level. For example, women with a Masters degree earns on an average 37.8 % less than males with Masters degree. In order to summarize the findings, we see that with same level of education women earn less than men. We further confirm this by performing bootstrap simulations for 1000 times and the result continued to be similar

Table 1:

data	N
Male	573
Female	992

Table 2:

	Mean	SD	Min	Max	Median	P95	N
Weekly earnings	1250.82	662.50	0.40	2884.61	1071.20	2884.61	1565
Weekly hours worked	40.80	7.41	2.00	92.00	40.00	50.00	1565
w	30.62	16.94	0.01	346.15	26.44	61.71	1565
lnw	3.30	0.55	−4.61	5.85	3.27	4.12	1565

Table 3:

	Wage (LM)	Log Wage (LM)
(Intercept)	34.1834**	3.4242**
	(0.6987)	(0.0195)
female	−5.5519**	−0.1840**
	(0.8775)	(0.0245)
Num.Obs.	1563	1563
R2	0.025	0.035

* p < 0.05, ** p < 0.01

Table 4:

	Model 1	Model 2	Model 3
(Intercept)	3.4242**	−3.1012**	3.0799**
	(0.0196)	(0.7560)	(0.0581)
female	−0.1840**	−0.1471**	−0.1432**
	(0.0246)	(0.0248)	(0.0247)
grade92		0.1510**	
		(0.0175)	
ed_Associate_ap			0.0607
			(0.0621)
ed_BA			0.3335**
			(0.0563)
ed_MA			0.4126**
			(0.0598)
ed_Profess			0.3795**
			(0.1324)
ed_PhD			1.1358**
			(0.3526)
Num.Obs.	1563	1563	1563
R2	0.035	0.083	0.094

* p < 0.05, ** p < 0.01

Table 5:

	Women (log Wage)	Men (log Wage)	All (log Wage)
grade92	0.0733**	0.0699**	0.0699**
	(0.0014)	(0.0030)	(0.0030)
ed_Associate_ap	−0.0746	0.1662	0.1662
	(0.0652)	(0.1651)	(0.1649)
ed_BA	0.1247*	0.3967**	0.3967**
	(0.0615)	(0.1318)	(0.1316)
ed_MA	0.0787	0.4672**	0.4672**
	(0.0684)	(0.1373)	(0.1371)
ed_Profess	−0.2145	0.4678*	0.4678*
	(0.1518)	(0.2003)	(0.2000)
ed_PhD	1.5368*	0.5811**	0.5811**
	(0.6692)	(0.2203)	(0.2200)
female			0.1369
			(0.1355)
female × ed_Associate_ap			−0.2375
			(0.1747)
female × ed_BA			−0.2653
			(0.1388)
female × ed_MA			−0.3785**
			(0.1438)
female × ed_Profess			−0.6689**
			(0.2436)
female × ed_PhD			0.9724
			(0.7016)
Num.Obs.	991	572	1563
R2	0.074	0.073	0.106

* $p < 0.05$, ** $p < 0.01$

Figure 1

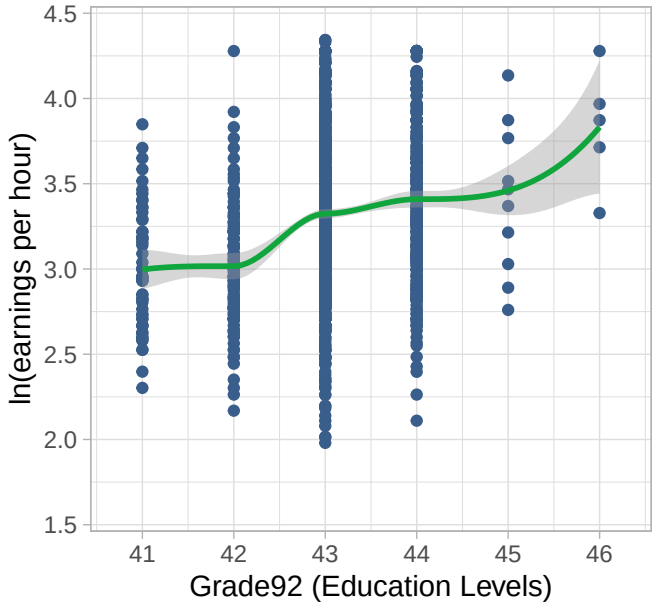
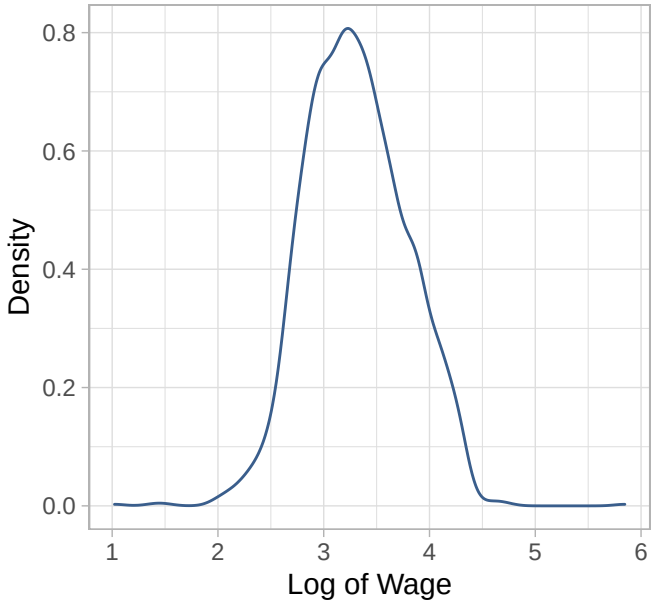


Figure : 2



Appendix

Degree names of grade92 education levels and their variables names used in our code were as follows; Associate- vocational (ed_Associate_voc), Associate-academic program(ed_Associate_ap), Bachelors(ed_BA), Masters(ed_MA), Professional(ed_Profess), PhD (ed_PhD)