DA2-Assignment-1

Gyongyver Kamenar (2103380)

Introduction

I investigated the gender wage gap in 5 similar occupations: 1) computer system analyst, 2) information security analyst, 3) computer programmers 4) software developers, application and system software and 5) web developers. These occupations include 2642 people, from which 1948 are males and 694 females. I decided to take the logarith of wages because financial data is more comparable in percentages, so my dependent variable is the log(earnings per working hours) which I modeled with several explanatory variables.

Data cleaning and descriptive statistics

I created the descriptive statistics table of the most relevant numerical variables. There are also categorical but ordinal variables so I calculated their descriptives too to spot any potential error or interesting feature. I found an error regarding the wage: a 43-years-old female earns 2$ for 40 hours of work, I think it is a measurement error so I excluded it from the data. I also excluded 3 people who have not finished the 12the grade of high school, they are 17 and 18 years olds any people usually do not work at that age. There are also another 3 people who do not have at least high school diploma but they are middle-aged, work full time and have proper earnings so I left them in my dataset.

Modelling

Firstly, I modeled the unconditional wage gap with a linear regression and heteroskedasticity-robust SE. The coefficient is -0.16 meaning that women earn 16% less on average. The coefficient is significant even at 0.1% level. We can say that it is statistically different from zero, the 95% CI is [-0.20; -0.12]. In the second model I included the grade92 variables as factors to show who the wage varies with gender and education. The base value is the highest level (Doctorate degree) so every grade92 coefficients are negative because these are compared to the Phd. Only the 38 and 45 categories are not significant from the grade92 categories, I think because there are too few people in these two cases (3 and 14). Let me interpret the coefficient of grade92=39 (high school graduates), people whose highest education is high school graduation earn 51% less on average within the same sex.

In the 3rd model I also added age because I think it must be a relevant factor in explaining earnings, the correlation between them is 0.26. As expected age is also significant at 0.1%. People aged 1 year more have 1.2% higher wage on average ceteris paribus. However, one can see on the 2nd graph that the relationship between age and wage is not linear, intuitively wages increase more at the beginning of the career but at a point it stops increasing and starts declining or stay constant. So in my 4th model I added age square as well. As we can see the age squared is also significant at 0.1% level and the R^2 increased a lot too. The sex coefficient is still significant in this model, but now females earn 18% less on average holding all the other variables constant.

Another explanatory factor can be the number of children so I added the number of own children (ownchild) variable to the model. It coefficient is significant at 1% level, by interpreting we can sey that people with one more child earn 2.8% more ceteris paribus. In the last model, I used the interaction between sex and education, to see that at different education elvels there is a difference between female and male wages. Actually none of the interaction terms happend to bee significant even at 10% level. In several educational categories there is not enough females to draw any statistical conclusion from it.

Summary

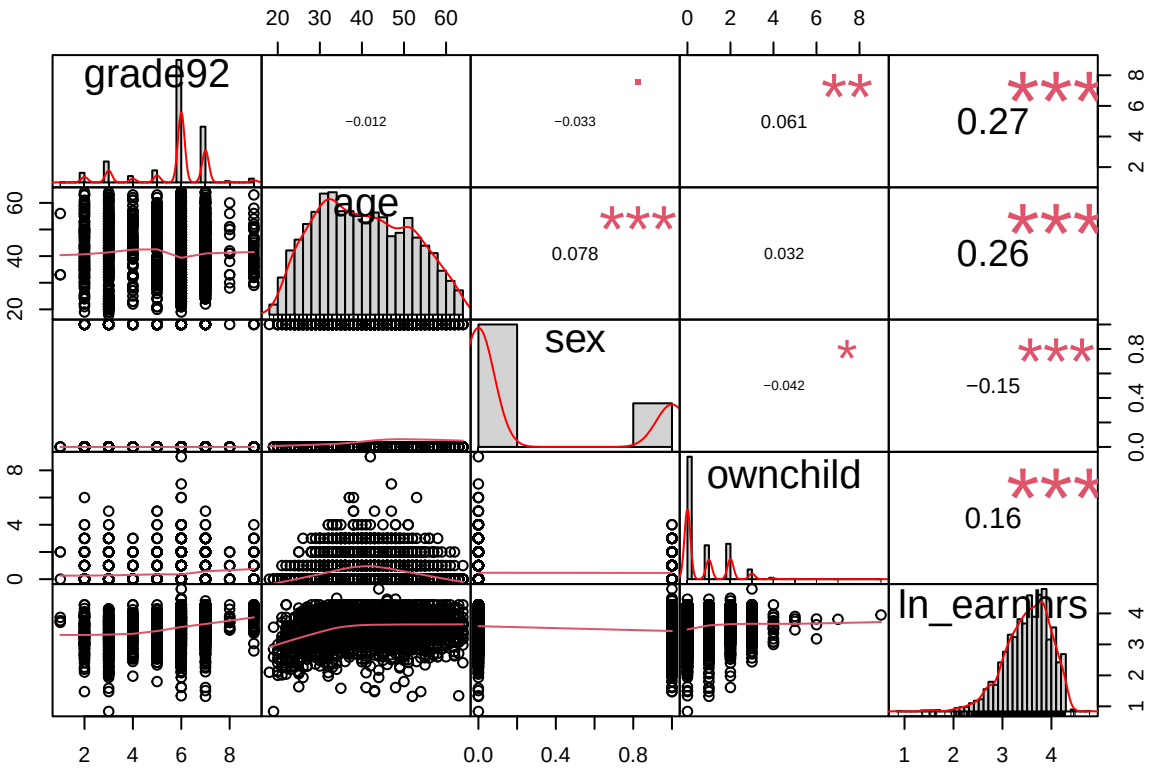
Within five computer and mathematical occupations I found ~16% unconditional gender wage gap. After controlling for age, education level and the number of children I still found that there is a 18% difference on average between male and female wages. In both cases the coefficient si significant at 0.1% level so we can say, that there is a gender wage gap (different from zero) in the population represented by this sample.

Table 1: Descriptive Statistics

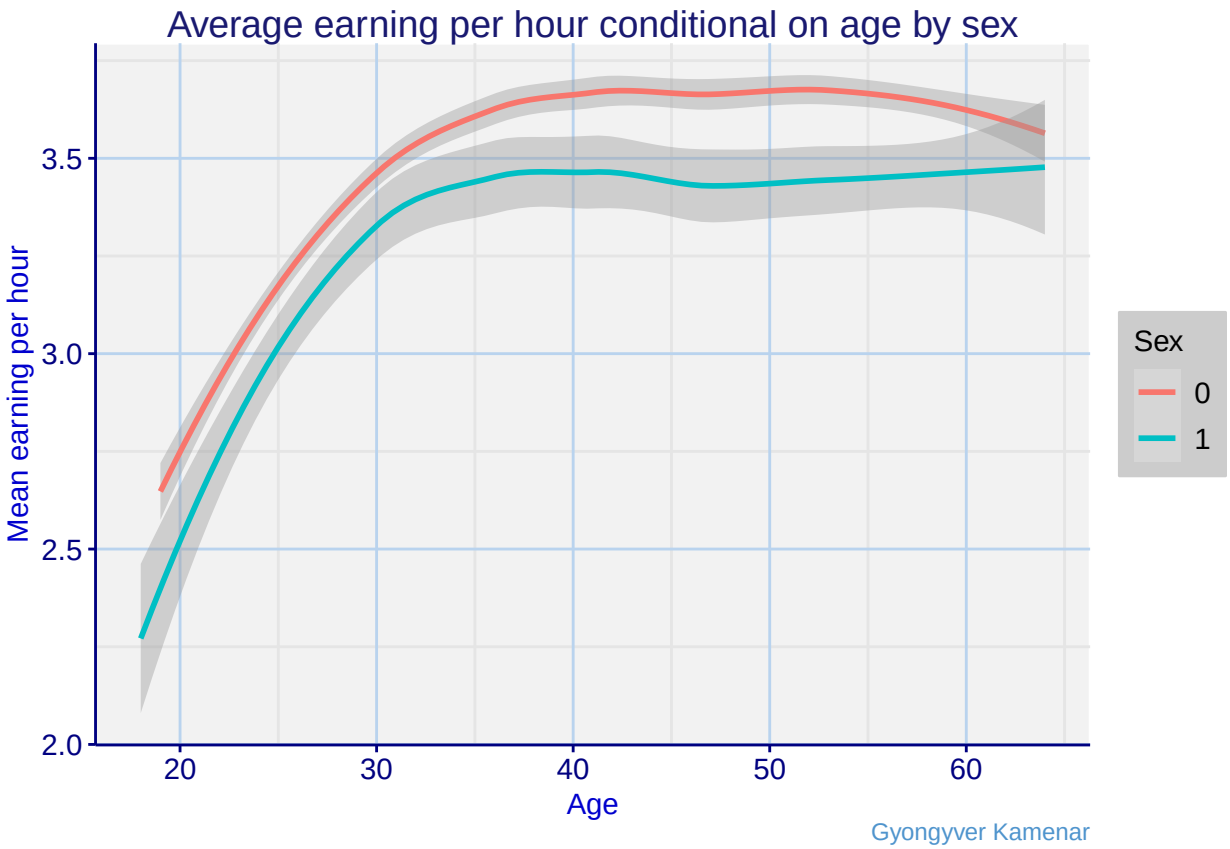
	N	Percent	Mean	SD	Min	Max	Median	P25	P75
earnwke	2642	100.00	1534.96	669.59	37.00	2884.61	1442.30	1001.25	1948.38
earnhrs	2642	100.00	36.97	15.63	2.31	120.19	35.00	25.00	48.08
ln_earnhrs	2642	100.00	3.51	0.48	0.84	4.79	3.56	3.22	3.87
uhours	2642	100.00	41.48	6.43	4	80	40.00	40.00	40.00
as.numeric(grade92)	2642	100.00	5.75	1.42	1.00	9.00	6.00	6.00	7.00
age	2642	100.00	40.56	10.99	18	64	40.00	32.00	50.00
ownchild	2642	100.00	0.70	1.02	0	9	0.00	0.00	1.00
as.numeric(chldpres)	2642	100.00	2.87	2.98	1.00	15.00	1.00	1.00	4.00

Appendix

Correlation plot



	reg1	reg2	reg3	reg4	reg5	reg6
Dependent Var.:	**ln_earnhrs**	**ln_earnhrs**	**ln_earnhrs**	**ln_earnhrs**	**ln_earnhrs**	**ln_earnhrs**
(Intercept)	3.552*** (0.0103)	3.784*** (0.0389)	3.297*** (0.0957)	2.076*** (0.1375)	2.197*** (0.1531)	3.784*** (0.0390)
sex	-0.1623*** (0.0219)	-0.1536*** (0.0210)	-0.1761*** (0.0203)	-0.1837*** (0.0200)	-0.1800*** (0.0200)	-0.1039 (0.1554)
grade9239		-0.4286*** (0.0600)	-0.4407*** (0.1000)	-0.4341*** (0.0791)	-0.4170*** (0.0931)	-0.3872*** (0.0631)
grade9240		-0.5498*** (0.0539)	-0.5326*** (0.0963)	-0.4874*** (0.0734)	-0.4733*** (0.0884)	-0.5433*** (0.0570)
grade9241		-0.3627*** (0.0686)	-0.4013*** (0.1039)	-0.3998*** (0.0834)	-0.3801*** (0.0973)	-0.3347*** (0.0742)
grade9242		-0.4053*** (0.0541)	-0.4327*** (0.0968)	-0.4272*** (0.0745)	-0.4111*** (0.0893)	-0.3955*** (0.0568)
grade9243		-0.2074*** (0.0410)	-0.1902* (0.0911)	-0.1828** (0.0672)	-0.1662* (0.0834)	-0.2144*** (0.0413)
grade9244		-0.1164** (0.0428)	-0.1170 (0.0919)	-0.1305. (0.0684)	-0.1155 (0.0843)	-0.1144** (0.0435)
grade9245		-0.0112 (0.0737)	-0.0375 (0.1140)	-0.0554 (0.0899)	-0.0361 (0.1016)	-0.0074 (0.0779)
grade9246		0.0821 (0.0668)	0.0735 (0.1052)	0.0338 (0.0866)	0.0466 (0.0994)	0.0775 (0.0699)
age			0.0120*** (0.0008)	0.0750*** (0.0061)	0.0669*** (0.0067)	
age_square				-0.0008*** (7.23e-5)	-0.0007*** (8.02e-5)	
ownchild					0.0276** (0.0085)	
sex x grade9239						-0.1625 (0.1857)
sex x grade9240						-0.0763 (0.1797)
sex x grade9241						-0.1616 (0.2064)
sex x grade9242						-0.0810 (0.1786)
sex x grade9243						-0.0224 (0.1579)
sex x grade9244						-0.0571 (0.1607)
sex x grade9245						-0.1035 (0.1694)
S.E. type	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedasti.-rob.	Heteroskedasti.-rob.	Heteroskedast.-rob.
Observations	2,642	2,642	2,642	2,642	2,642	2,642
R2	0.02254	0.10273	0.17777	0.21387	0.21671	0.10401
Adj. R2	0.02217	0.09966	0.17465	0.21058	0.21313	0.09855


Average earning per hour conditional on age by sex
Mean earning per hour
3.5
3.0
2.5
2.0
20
30
40
50
60
Age
Sex
0
1
Gyongyver Kamenar