DA2 - Assingment 1

Oszkar Egervari

11/27/2021

Introduction

In this assignment I examine the gender wage gap difference in the cps-earningsdatase dataset. First I take a look at the unconditional gender gap, then I will attempt to show how the gender gap varies with the level education.

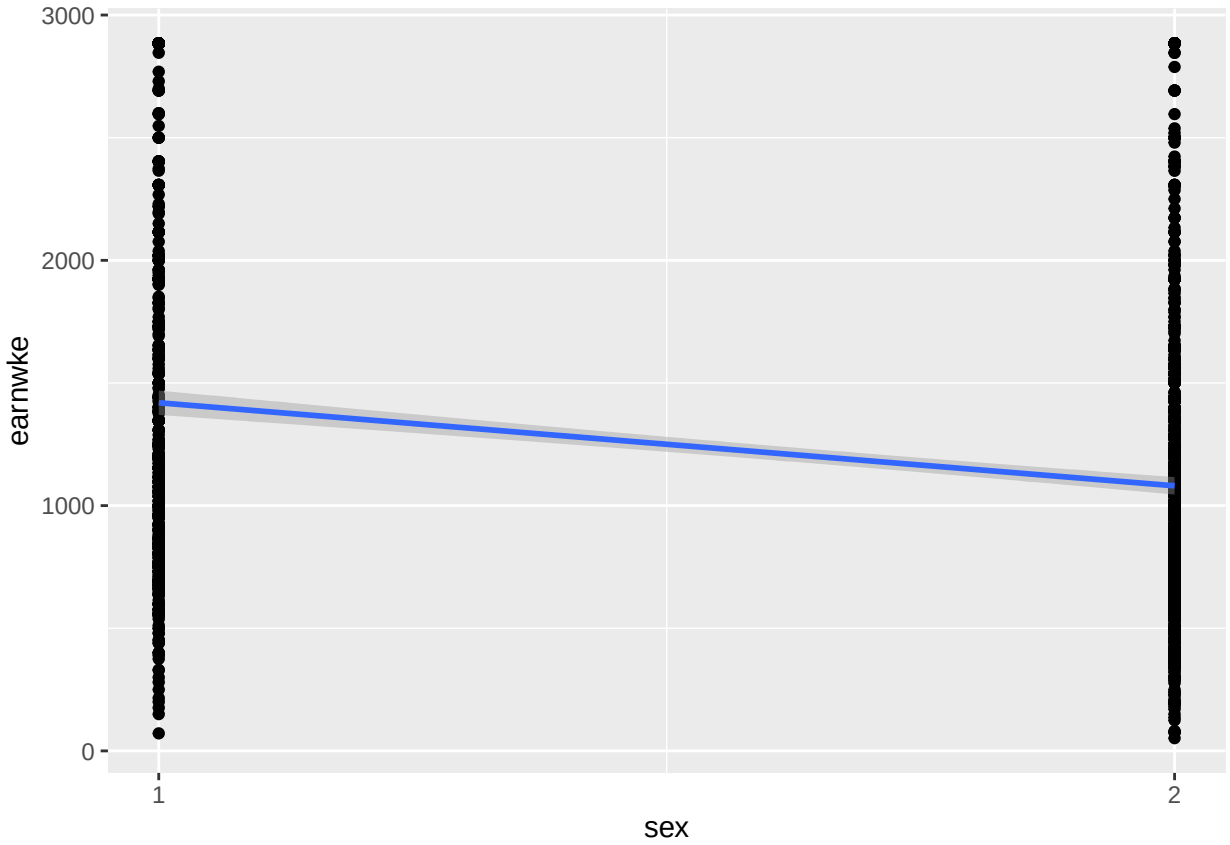
From the dataset I chose the 'Accountants and auditors' profession with occupation code 800.

Unconditional gender gap

```
## OLS estimation, Dep. Var.: earnwke
## Observations: 1,811
## Standard-errors: IID
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 1757.979    53.8152  32.6670 < 2.2e-16 ***
## sex         -338.624    31.1679 -10.8645 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 627.8   Adj. R2: 0.060734
```

The table shows us that the Intercept is 1757.979, which in this case means the expected male weekly earnings. The coefficient 'sex' has a value of -338.624, which means that females are expected to earn less than males by this amount. The t values of the coefficients are very small, which means that the probability that the observed wage gap difference between sexes is due to chance, is unlikely. The adjusted R2 score of this model is around 0.06, which means that 6% of the wage difference is explained by sex.

```
## `geom_smooth()` using formula 'y ~ x'
```



Gender gap with education level

In this section I examine the gender wage gap with the added factor of education level. In the original dataset, the level of education is a number, so as a first step I added a new column to the dataset with the text forms of the education level.

```
f <- df %>%
  mutate(education_lvl = case_when(grade92 == 34 ~ '7th or 8th',
                                   grade92 == 37 ~ '11th',
                                   grade92 == 38 ~ '12th grade NO DIPLOMA',
                                   grade92 == 39 ~ 'High school graduate, diploma or GED',
                                   grade92 == 40 ~ 'Some college but no degree',
                                   grade92 == 41 ~ 'Associate degree -- occupational/vocational',
                                   grade92 == 42 ~ 'Associate degree -- academic program',
                                   grade92 == 43 ~ 'Bachelors degree (e.g. BA,AB,BS)',
                                   grade92 == 44 ~ 'Masters degree (e.g. MA,MS,MEng,Med,MSW,MBA)',
                                   grade92 == 45 ~ 'Professional school deg. (e.g. MD,DDS,DVM,LLB,JD)',
                                   grade92 == 46 ~ 'Doctorate degree (e.g. PhD, EdD)'))
```

Before the regression model, I'd like to take a look at the dataset with the help of the datasummary function, to see if there are some differences, that are immediately obvious:

We can see, that even though the education level of males and females are basically the same, moreover females are a little older than males, the average weekly earnings of males are almost 400 $ greater than the female earnings. We can also see that the male mean is 220 dollars greater than the male median, which

as.factor(sex)		Mean	SD	Min	Max	Median	P95	N
1	age	40.78	11.89	20.00	64.00	40.00	60.30	615
	earnwke	1419.35	717.70	71.53	2884.61	1200.00	2884.61	615
	grade92	42.97	1.19	34.00	46.00	43.00	44.00	615
2	age	42.98	11.10	17.00	64.00	43.00	60.00	1196
	earnwke	1080.73	576.73	52.00	2884.61	960.00	2403.00	1196
	grade92	42.39	1.49	34.00	46.00	43.00	44.00	1196

indicates a model that is skewed to the left side (there are outlier values). This difference is only 120 dollars in case of females.

Now let's take a look at the regression table:

```
## OLS estimation, Dep. Var.: earnwke
## Observations: 1,811
## Standard-errors: IID
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) -3276.160   449.0883 -7.29513 4.4443e-13 ***
## sex          -271.813    30.7092 -8.85119  < 2.2e-16 ***
## grade92       115.611    10.2440 11.28568  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 606.8   Adj. R2: 0.122062
```

The intercept shows the case when sex is male and the education level is 0 (of course that is not realistic, in our data the lowest level is 34 which is 7th or 8th grade). Sex means the difference in weekly earnings accounted by the sex variable and grade92 means the same thing for education. We can see that the relationship of weekly earnings and education is positive. The adjusted R2 score is 12% now and the coefficients are again significant on a high level.

I was curious to see a model, where the education levels are aggregated to a higher level. So I made three groups:

```
df1 <- df %>%
  mutate(education_lvl_2 = case_when(grade92 >= 34  & grade92 <=39 ~ 'High School diploma or lower',
                                     grade92 >= 40  & grade92 <=43 ~ 'Bachelor or lower',
                                     grade92 >= 44  & grade92 <=46 ~ 'Post graduate diploma and Doctora
```

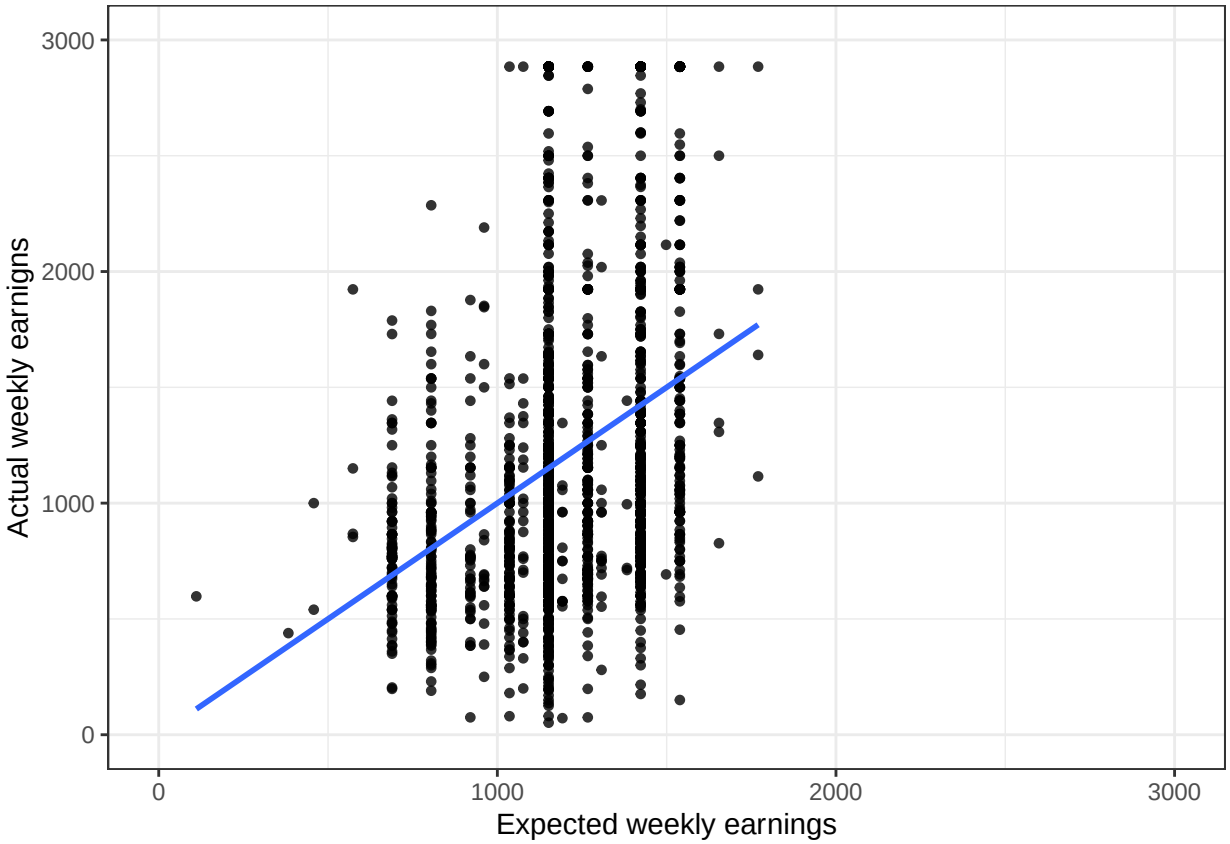
The groups are arbitrary. I added 'Doctorate degree (e.g. PhD, EdD)' to the 'Post graduate diploma' level, because in the previous group there were only 6 employees.

```
## OLS estimation, Dep. Var.: earnwke
## Observations: 1,811
## Standard-errors: IID
##                                                    Estimate Std. Error  t value
## (Intercept)                                        1673.928    54.8998 30.49059
## sex                                                -301.280    31.0940 -9.68934
## education_lvl_2High School diploma or lower        -288.121    62.2766 -4.62648
## education_lvl_2Post graduate diploma and Doctorate  209.328    37.9122  5.52140
##                                                     Pr(>|t|)
## (Intercept)                                         < 2.2e-16 ***
## sex                                                 < 2.2e-16 ***
```

```
## education_lvl_2High School diploma or lower        3.9837e-06 ***
## education_lvl_2Post graduate diploma and Doctorate 3.8508e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 617.9   Adj. R2: 0.089003
```

The Intercept this time shows the males in the 'Bachelor or lower' group. Compared to this value, the female expected are earnings lower by -301.28 dollars. We can see, that in case of a 'High school diploma or lower' level education, the expected earnings are lower by 288.121 (compared to the 'Bachelor or lower' group), while the post graduate and doctorate group is expected to earn 209.328 dollars more as the base group. This model has a slightly less R2 value than the previous with around 0.089.

Lastly, I'd like to take a look at the yhat - y graph:



We can see based on the graph, that the model doesn't predict the weekly earnings above the 2000 dollar level.

Conclusion

It can be concluded based on the regression models, that there is a difference (significant on a high level) between the weekly earnings of the sexes in the dataset used, among the 'Accountants and auditors' profession. Furthermore, we can say, that if the level of education is higher, the expected earnings are also higher.