Homework 1

Zsiros, Gabriella

2022-11-26

Prep

```
library(tidyverse)
library(modelsummary)
library(stargazer)
library(fastDummies)
library(huxtable)
library(estimatr)
library(knitr)
knitr::opts_chunk$set(fig.pos = "H", out.extra = "")
df <- read.csv('/users/Gabi/Downloads/morg-2014-emp.csv')
```

About the data

Dataset is available at https://osf.io/g8p9j/ . The purpose of this report is to analyse earnings of men and women in a certain occupational sector.

I calculated the hourly earnings as well as its logarithmic values to help with further analysis.

```
df <- df %>%
  mutate(w = earnwke / uhours) %>%
  mutate(lnw = log(w))
```

Which occupation to choose?

I considered that I should have approximately same amount of male data as female, and should have originally more than 500 observations per sex. Based on a short check I have selected the category of *Marketing and sales managers*.

occ2012	Sex1	Sex2	ratio
50	539	494	1.09
2200	696	741	0.939
4760	1632	1594	1.02

Removing extreme values

	sex_factor	Mean	Median	Min	Max	P5	P95	Range
earnwke	male	1667.44	1538.46	192.30	2884.61	610.77	2884.61	2692.31
	female	1254.61	1076.92	1.00	2884.61	404.00	2884.61	2883.61

	sex_factor	Mean	Median	Min	Max	P5	P95	Range
w	male	38.20	37.50	4.81	100.00	14.00	72.12	95.19
	female	29.68	25.85	0.03	73.70	11.00	58.89	73.67

How many hours?

A quick check of the distribution of hours has led me ot narrow it down between 20 and 60 hours per week.

```
df <-  df %>% filter(uhours > 20 & uhours < 60) %>% filter(occ2012 == 50)
df <- df %>%  mutate(df,
                     sex_factor = factor(df$sex,labels = c('male','female')),
                     .after = sex)
df <- df %>% filter(grade92 >38)
```

Hourly earning of men and women

Statistical summary

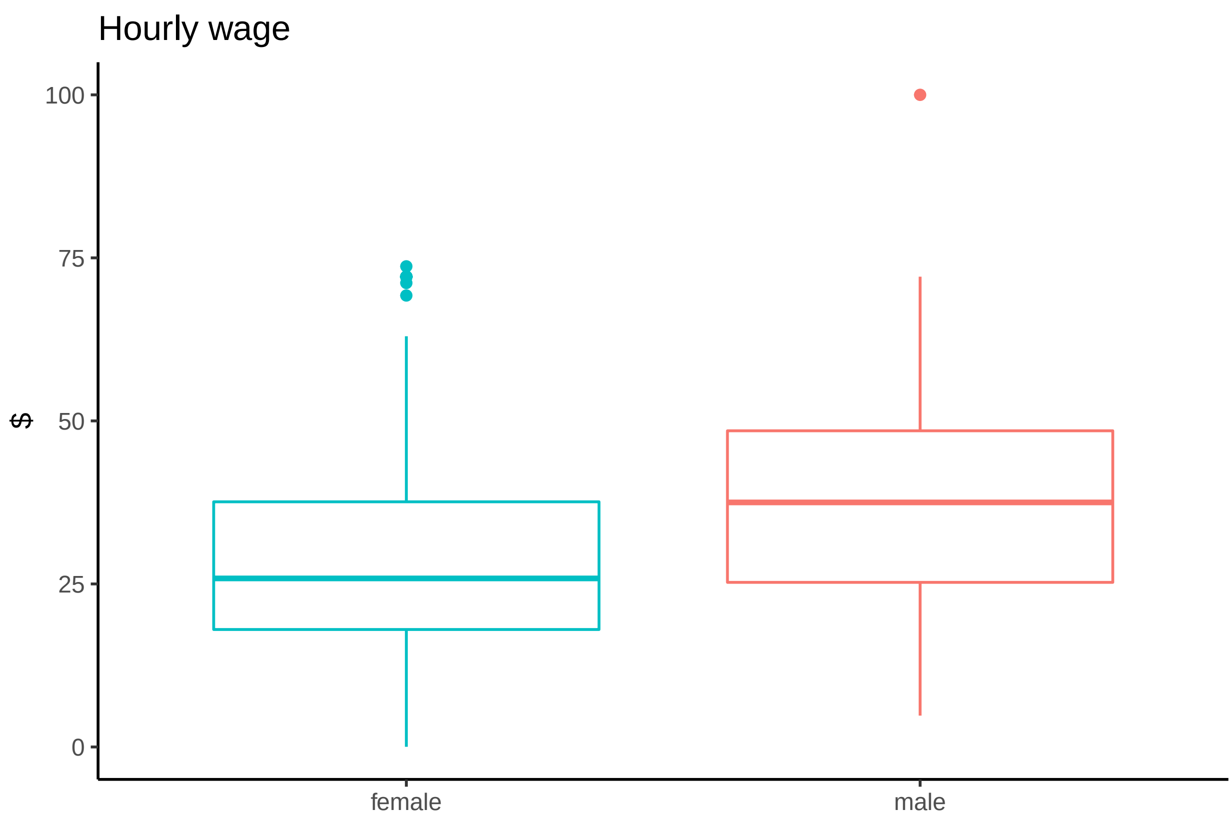
```
datasummary (earnwke * sex_factor ~
               Mean + Median + Min + Max + P5 + P95 + Range,
             data = df)
```

```
datasummary (w * sex_factor ~
               Mean + Median + Min + Max + P5 + P95 + Range,
             data = df)
```

It is visible in he summary that both the mean and median show difference between the two sexes.

Visualizing the wage gap

```
ggplot(data = df, aes(x = sex_factor, y= w, color = sex_factor))+
  geom_boxplot() +
  scale_x_discrete(limits=rev)+
  labs(x = '', y = '$', title = "Hourly wage", ) +
  theme_classic() +
  theme(legend.position="none")
```



T-test

```
df50f <- df %>% filter(sex == 2)
df50m <- df %>% filter(sex == 1)
t.test(df50m$w,df50f$w, mu = 0)
```

```
##
##  Welch Two Sample t-test
##
## data:  df50m$w and df50f$w
## t = 7.9448, df = 901.17, p-value = 5.784e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.419754 10.632048
## sample estimates:
## mean of x mean of y
##  38.20496  29.67906
```

T test with value 8.5176 shows with a p -value of 2.2e-16 (very close to zero) that there is a significant difference in the average earning between men and women. Men earn 6.25-10.46 $ more on a weekly basis with 95% CI.

Linear regression

```
reg1 <- lm( w ~ sex, df)
reg2 <- lm(lnw ~ sex, df)
huxreg('wage' = reg1,'ln wage' = reg2)
```

	wage	ln wage
(Intercept)	46.731 ***	3.831 ***
	(1.695)	(0.059)
sex	-8.526 ***	-0.295 ***
	(1.073)	(0.037)
N	906	906
R2	0.065	0.065
logLik	-3805.266	-759.514
AIC	7616.532	1525.028

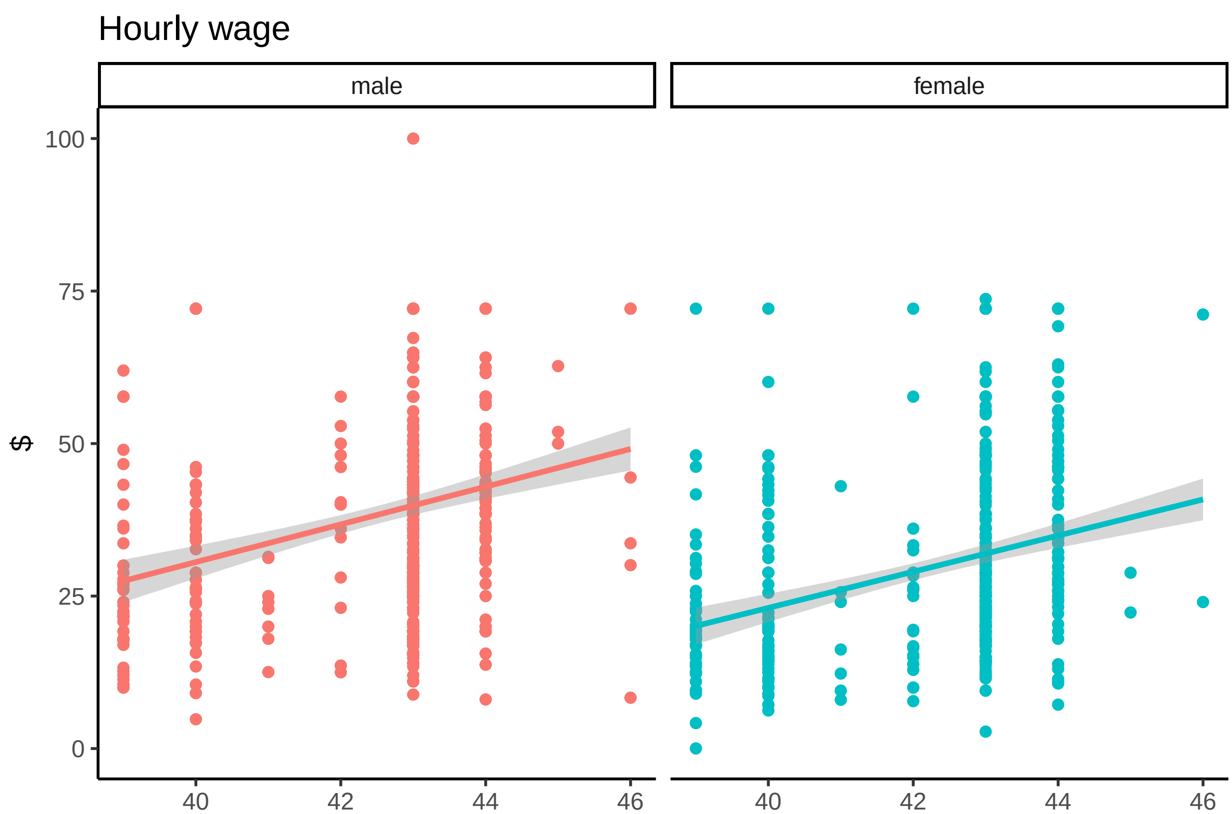
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Applying simple regression analysis shows that women earn $8.5, i.e. 29% less on average on a weekly basis

Introducing grade variable

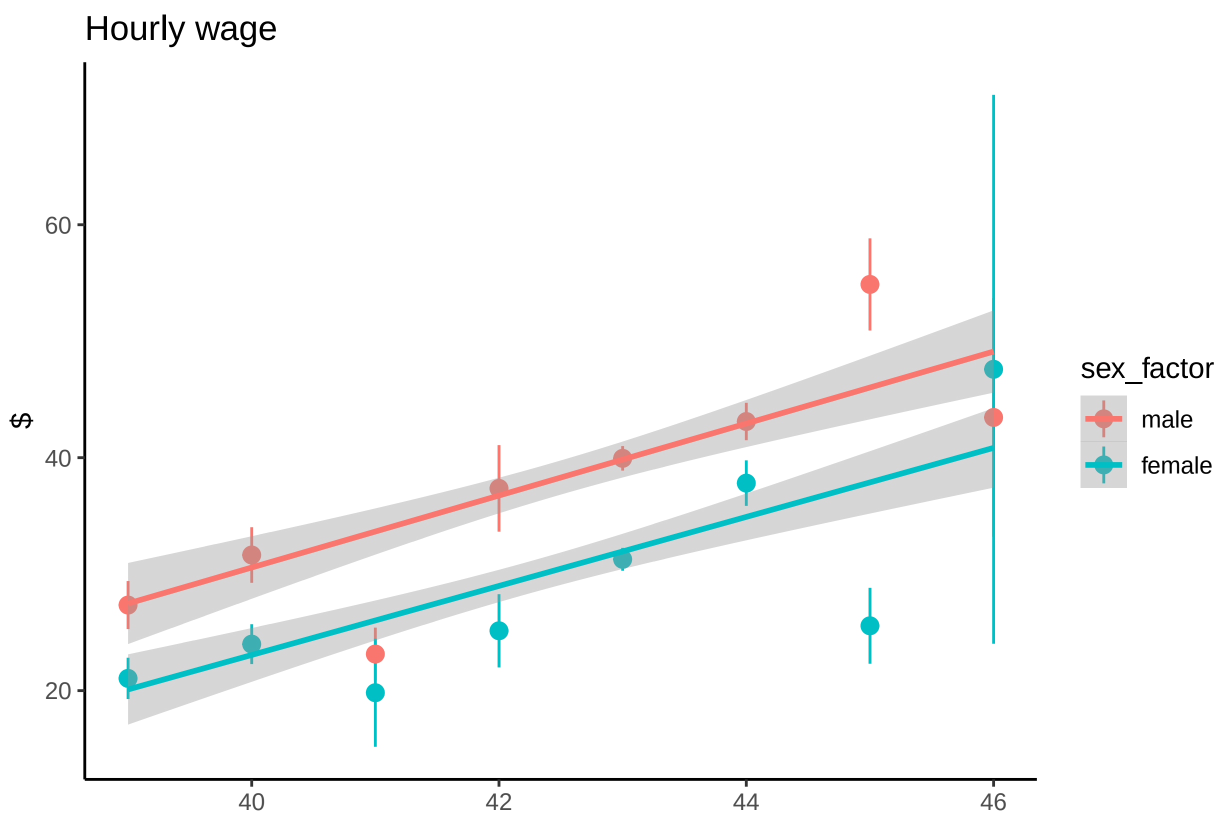
Scatter plot with regression

```
ggplot(data = df, aes(x = grade92, y=w, color = sex_factor))+
  geom_point()+
  geom_smooth(method = 'lm')+
  labs(x = '', y = '$', title = "Hourly wage", ) +
  facet_wrap(~sex_factor)+
  theme_classic() +
  theme(legend.position="none")
```



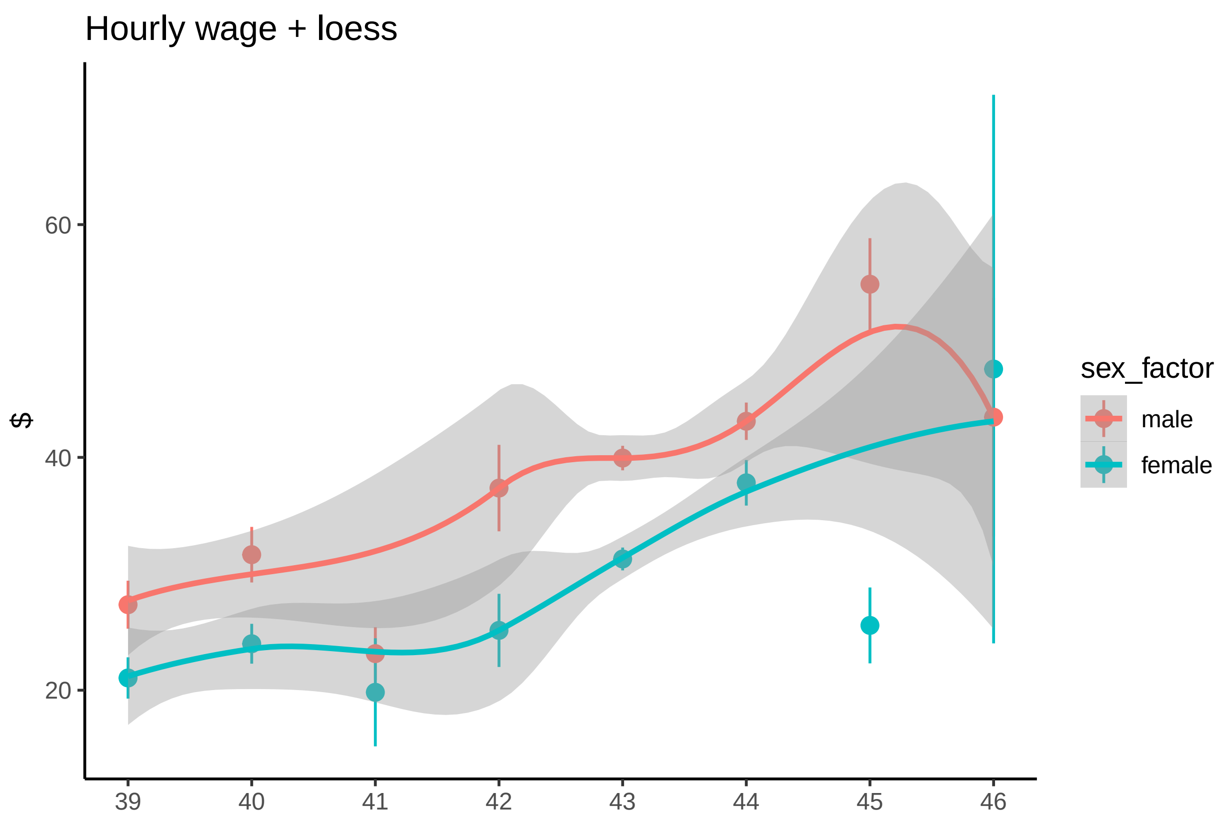
Summary plot with regression

```
ggplot(data = df, aes(x = grade92, y=w, color = sex_factor))+
  stat_summary()+
  scale_x_continuous(breaks = c(39:46))+
  geom_smooth(method = 'lm') +
  labs(x = '', y = '$', title = "Hourly wage", ) +
  xlim(39,46)+
  theme_classic()
```



Loess

```
ggplot(data = df, aes(x = grade92, y=w, color = sex_factor))+
  stat_summary()+
  geom_smooth(method = 'loess') +
  scale_x_continuous(breaks = c(39:46))+
  labs(x = '', y = '$', title = "Hourly wage + loess", ) +
  theme_classic()
```



Lowess method in this case does not seem to be sensible, as the grade variable is a factor, rather than a numerical value.

Multivariate regression

```
reg4 <- lm( w ~ sex + grade92, df)
reg5 <- lm( lnw ~ sex + grade92, df)
reg6 <- lm_robust(lnw ~ sex + grade92, data = df, se_type = "HC1")
huxreg('wage'=reg4,'ln wage'= reg5,'ln wage robust' = reg6)
```

```
knitr::opts_chunk$set(fig.pos = "H", out.extra = "")
```

Log-level transformation seems to be a more accurate model, with lower SE-s, and higher R2. In this case robust SE does not show great decrease of SE, so the second model (`reg5`) will be used to final summary.

We can see a greater statistical significance in Bachelor's and Master's degree

	wage	ln wage	ln wage robust
(Intercept)	-82.492 ***	-0.926 *	-0.926
	(13.474)	(0.464)	(0.573)
sex	-7.807 ***	-0.268 ***	-0.268 ***
	(1.025)	(0.035)	(0.035)
grade92	3.026 ***	0.111 ***	0.111 ***
	(0.313)	(0.011)	(0.014)
N	906	906	906
R2	0.153	0.164	0.164
logLik	-3760.714	-708.978	
AIC	7529.428	1425.957	

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

Summary

```
reg7 <- lm( grade92 ~ sex, df)
huxreg('ln wage' = reg2, 'ln wage' = reg5, 'grade' = reg7,  statistics = c(N = "nobs", R2 = "r.squared")
```

	ln wage	ln wage	grade
(Intercept)	3.831 *** (0.059)	-0.926 * (0.464)	42.708 *** (0.171)
sex	-0.295 *** (0.037)	-0.268 *** (0.035)	-0.238 * (0.109)
grade92		0.111 *** (0.011)	
N	906	906	906
R2	0.065	0.164	0.005

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

Comparing men and women in *Marketing and Sales manager* occupational sector, analysis shows an approximate 30% difference in average salaries, considering a 20-60 work week. The second model introduces the education level, where comparing men and women in the same education level, we get a 26.8 log point difference, which here I will interpret as 27%. Relation between grade and sex is not to be interpreted in this case, since the education level is a factor.