Uppsala University

Applied Statistical Methods

Homework Assignment 3

Gender Discrimination

Claes Kock

Mayara Latrech

Yuchong Wu

January 13, 2020

Abstract

This is a study on the gender discrimination regarding wages in Sweden. Using the data set of Households Economic Living Conditions, we perform a Mincer's earnings equation and then solve the problem of heteroskedasticity using **Box-Cox transformation**.

To figure out if there is a difference in wages between men and women, we use **Non-parametric two-samples Wilcoxon rank test** without regression and **T-test** within regression. Both tests show that it is statistically significant that on average men earn 9.97 Kr more per hour, compared to women.

To investigate if the difference in wages is caused by something other than just discrimination, we taking *education* and *experience* into account. And we found that even we add these two factors into our model, there is still some remaining effect which can be explained by *gender*. After calculating, We conclude that there is an average wage difference of 1.75 Kr per hour due to gender-based discrimination.

There might be a problem that education is correlated with personal ability which is not in the model, nor the data. To solve the problem, we managed to pick out the strongest valid instrument from plenty of possible instruments, and then perform a **two-stage least square regression** instead of **OLS**. After comparison, we conclude that there is a slight difference in estimation between the two models.

After establishing another regression model which is grouped by *gender* and performing a statistical test, we are sure that one more year of *education* is more worth to a man than to a woman. But for *experience*, there is insufficient evidence to support the hypothesis.

Contents

1 Introduction

From a theoretical perspective, this assignment intends to examine the wage gap between men and women in the Swedish economy. By "wage gap" we mean the difference in average salary between men and women.

This is done through the use of linear models. Thus "gender" will be an important variable, but we will also apply adjusted models, where we have applied additional variables which might influence the wage a person receives. These include nationality of parents, if the family spoke Swedish or not, where the person grew up and the general level of education, both for the person in question. but also their parents.

2 Data

Our data comes from the project HUS(Households economic living conditions). HUS is a database about households use of time, money and public services. HUS also contains background variables, e.g. education, working experience and salaries.

The survey is representative for native Swedish speaking families living in Sweden when the survey was conducted. The data set contains 16 variables. Out of these 16 variables, we have renamed 14 variables in order to better reflect what they describe.

3 Preliminary Analysis

In the data set, we have several values which show a null value (due to information being unavailable), and thus we need to clean the data.

Table 1: Preliminary Analysis for the variables

colnames.dat.	means	variances	mins	maxs
birth_year	43.25	141.50	5	66
gender	1.51	0.25	1	2
citizenship	1.30	0.79	1	5
language	1.09	0.08	1	2
lived_with	1.29	1.09	1	9
father_schooling	2.66	2.58	1	8
father_occupation	4.11	6.03	0	9
mother_schooling	2.45	1.66	0	8
mother_occupation	1.88	1.33	0	5
live	2.91	1.38	1	5
county	11.23	55.87	1	25
education	10.96	11.61	0	28
marital	1.46	0.55	1	3
experience	19.75	138.76	0	63
category	0.99	0.13	0	5
wage	46.47	486.50	7	448

As we can see from the above table, the average individual was born in 1943, making the average age 48 years. [1] The average hourly wage is approximately 46.5 Kr, which is more than it looks, due to the value not being adjusted for inflation.

The average years worked for an individual is around 20 years, and the amount of schooling is around 10 years. We can also see that while the education of the parents seems to be of a similar level (on average), the value for the type of occupation is not. [2]

[1] The data set was released in 1991

[2] The outliners might be affecting these values.

4 Regression

4.1 Mincer's earnings equation

For the regression model we look at education and experience, which is the amount of years spent in education and the amount of years spent working respectively.

$$ln(wage) = \beta_0 + \beta_1 education + \beta_2 experience + \beta_3 experience^2 + u$$

We apply this regression to our data set and get the following coefficients.

Table 2: Regression Coefficients for the regression model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.97	0.03	91.01	0
education	0.04	0.00	20.85	0
experience	0.03	0.00	12.29	0
experience_sq	0.00	0.00	-7.10	0

The estimate of the model is: [3]

$$ln(wage) = 2.97 + 0.04\ education + 0.03\ experience - 0.0003\ experience^2 + u$$

To figure out whether the residuals in the wage model we estimated seem to be heteroscedastic or not, we plot the residuals.

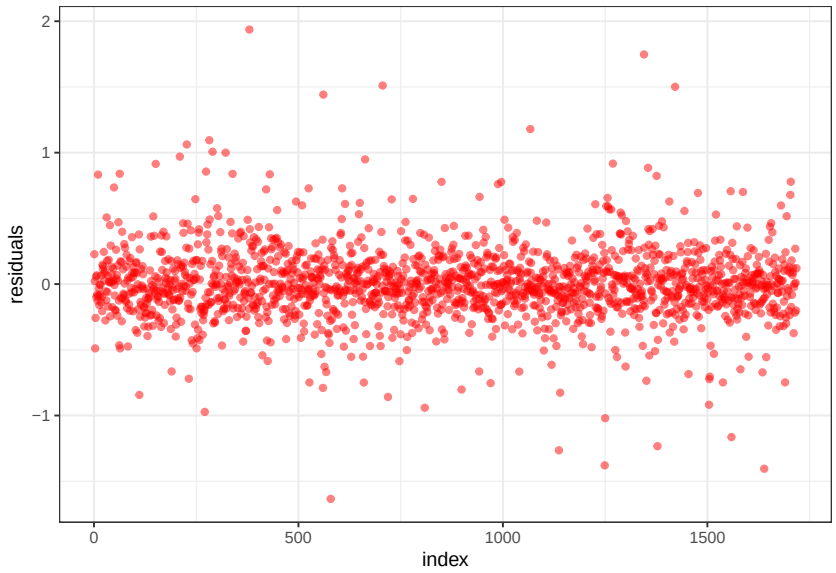

Figure 1: Residuals Plot for the regression model

[3]Detail in APPENDIX: Regression Outcome for the Mincer's earnings equation

As can be seen in the plot, there might be heteroskedasticity in our model, which is not clear enough. To further verify heteroskedasticity, we perform a Brausch-Pagan test.

Table 3: Brausch-Pagan test

	BP	df	p_value
bptest	12.66	3	0.01

Since the p-value of the test is lower than 0.05, we **reject** the null hypothesis that we do not have a heteroskedacticity in our model, which we of course don't want to have.

4.2 Box-Cox transformation

Having heteroskedacticy weakens OLS (not best anymore) and we also risk getting incorrect standard errors. In order to rectify this problem we perform a **Box-Cox transformation** on our dependent variable. This means that we transform the non-normal dependent variable into a normal dependent variable.

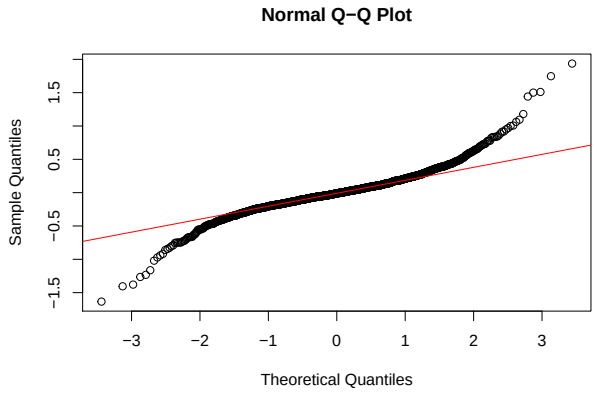

Figure 2: QQ-plot for the model before Box-Cox transformation

Here we can see the spread of the residuals of the model before the Box-Cox transformation.

We transform our dependent variable into a normal variable and add it to our data set. Following this, we construct a new model where the dependent variable is the transformed dependent variable. [4] After doing this we perform a Brausch-Pagan test to see if we still have a problem with heteroscedasticity.

Table 4: Brausch-Pagan test for the model after Box-Cox transformation

	BP	df	p_value
bptest	4	3	0.26

From the result of our test, we can observe a p-value of 0.2618, which is smaller than 0.05. We thus **can not reject** the null hypothesis that we have homoscedasticity. Thus our problem is solved.

[4] Detail in APPENDIX: Regression Outcome after Box-Cox transformation on our dependent variable

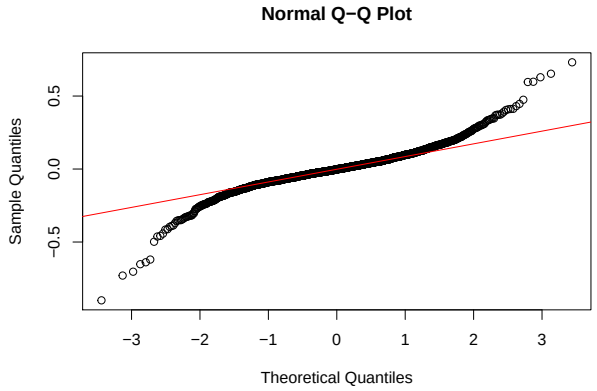

Figure 3: QQ-plot for the model after Box-Cox transformation

This QQ-plot shows the distribution of the residuals of our new model. We can see that the spread is much smaller compared to the first model, showing that the spread of the residuals has decreased.

Table 5: Regression Coefficients for the model after Box-Cox transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.39	0.01	94.77	0
dat$education	0.02	0.00	20.51	0
dat$experience	0.01	0.00	12.43	0
dat$experience_sq	0.00	0.00	-7.30	0

Our new estimated model after applying the BoxCox-test:

$$ln(wage) = 1.39 + 0.02\ education + 0.01\ experience - 0.0001\ experience^2$$

By applying the BoxCox-test, we force the dependent variable to be normal, which in turn affects the behavior of the residuals, making them even more spread.

5 Difference in Wages between men and women

5.1 Difference in means

First we compare the difference in means between male and female workers:

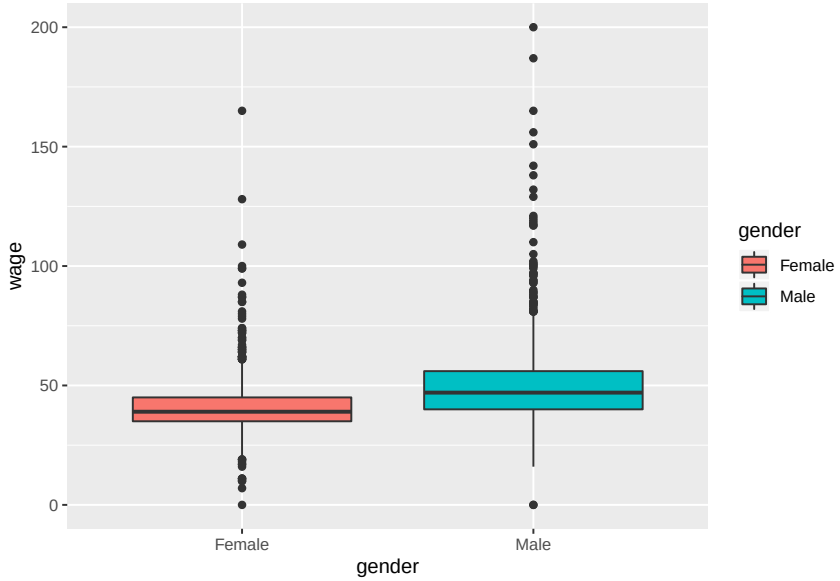

Figure 4: Difference in Wages between men and women

Table 6: Comparison in mean of wages between male and female workers

gender	wage
Male	51.54
Female	41.57

The average difference is approximately 9.97 Kr/hour.

5.2 Test of difference in means

In this case, we have two unrelated [5] groups of samples. Therefore, we use an **independent t-test** to evaluate whether the means are different.

5.2.1 Assumptions

- two samples are independent

Samples from men and women are not related.

[5] independent and unpaired

- The data from each of the 2 groups follow a normal distribution

To verify this assumption, we use **Shapiro-Wilk Normality Test**.

Table 7: Shapiro-Wilk normality test for Men's wage

	method	statistic	p_value
Male	Shapiro-Wilk normality test	0.8	0

Table 8: Shapiro-Wilk normality test for woman's wage

	method	statistic	p_value
Female	Shapiro-Wilk normality test	0.83	0

According to the outcome, we **reject** the null hypothesis that the data are normally distributed, and conclude that the data are not normally distributed.

Thus we use **non parametric two-samples Wilcoxon rank test** instead of Shapiro-Wilk Normality Test.

5.2.2 Non parametric two-samples Wilcoxon rank test

Table 9: Non parametric two-samples Wilcoxon rank test for woman's wage

method	statistic	p_value
Wilcoxon rank sum test with continuity correction	0.834	0

According to the outcome, we **reject** the null hypothesis that the means of wages are equal between male and female, and conclude that there is gender discrimination.

5.3 Test using regression model

Then we formulate a regression model which allow us to test if there is a difference in wage.

In this model, wage is dependent on the dummy variable "Male", which is 1 if the individual is male and 0 otherwise. Our model for examining the wage difference between genders is:

$$Wage = \beta_0 + \beta_1 \ gender + u$$

Table 10: Regression Coefficients for the model including dummy variable

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	41.57	0.73	57.15	0
gender_1	9.98	1.04	9.62	0

From our results [6], we can see that the parameter of gender is non-zero and also **significant**, since the p-value is lower than 0.05. Thus there is a difference in wage when the individual is male, compared to when the individual is a female. However, do note that the R-squared is very low, so only a small part of the total wage is explained by gender.

Table 11: Wage difference between male and female

Method	Result	Test	Conclusion
Difference of Means	9.97	Non parametric two-samples Wilcoxon rank test	significant
Regression	9.97	T-test	significant

From the above table, we can see that the result for both tests are the same. Both tests show that it is **statistically significant** that on average men earn 9.97 Kr more per hour, compared to women in our data set. From these tests, we would conclude that there is gender discrimination.

[6] Detail in APPENDIX: Regression Outcome for the regression model including dummy variable

6 Taking *education* and *experience* into account

6.1 Taking *education* into account

We want to investigate if the difference in wages is caused by something other than just discrimination. For example, does education and experience play a role, and if so, is there a difference between men and women for these variables? To investigate this, we start by adding education as a variable to our model.

Table 12: Regression Coefficients for the model containing education variable

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.88	1.71	12.79	0
gender_1	8.95	1.00	8.99	0
education	1.84	0.15	12.60	0

In our new model [7], we can see the effect of gender, as well as the effect of education. If we hold education constant, the effect of gender on wage is 8.94, meaning an increase of 8.94 Kr per hour on average if the individual is male, when holding education constant. This is slightly different compared to the previous model, which had an increase of 9.97 Kr per hour. We can also see that this model has a higher R-square value compared to our previous model. However, it is still fairly low. We can see that education is non-zero and significant, meaning that it has an effect on an individuals hourly wages.

We perform an F-test to see if education is the variable affecting the difference in the wage rather than gender.

$$wage = \beta_0 + \beta_1\ gender + \beta_2\ education$$

1. **Hypothesis:**

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

2. **Significance level:**

$$\alpha = 0.05$$

3. **Estimators:**

$$\hat{\beta}_1,\ R^2_{UR},\ R^2_R$$

4. **Assumptions:**

$$Large\ n$$

5. **Test statistic:**

$$F_{obs} = \frac{(R^2_{UR} - R^2_R)/m}{(1 - R^2_{UR})/(n-k)} \sim F_{m,n-k}$$

[7] Detail in APPENDIX: Regression Coefficients for the model containing education variable

where,

- R^2_{UR} is the coefficient of determination for the Unrestricted model
- R^2_R is the coefficient of determination for the Restricted model
- m is the number of (linear) restrictions (in the null hypothesis)
- n is the number of observations
- k is the number of regression coefficients (parameters) in the regression line of the Unrestricted model (including the intercept)

6. **Figure under the null and decision rule:**

We reject the null if $F_{obs} > F_{m,n-k}$ or if p-value<0.05.

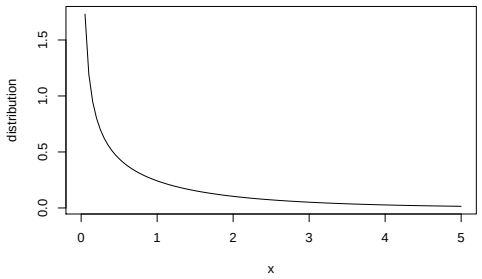

Figure 5: F distribution

7. **Calculations and decision:**

p-value = 2.2e-16 < 0.05 so we reject the null.

8. **Conclusion:**

Under the 5% significance level, we reject the null that gender is not relevant for hourly earnings. Hence, with this model we can see that there is gender discrimination.

6.2 Taking *education* and *experience* into account

We continue by adding experience to the model.

Table 13: Regression Coefficients for the model containing variable education and experience

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.85	2.05	3.35	0
gender_1	5.61	0.99	5.65	0
education	2.40	0.15	16.27	0
experience	0.54	0.04	12.28	0

In this model [8] we can see the effect of education and experience together. If we hold them both constant, the difference in hourly wage between men and women is 5.612 Kr per hour, which is a fairly large decrease compared to the other models. The clear difference is still there, however. We have an increase in the value of R-square again, the model explains around 20% of the variation of hourly

[8] Detail in APPENDIX: Regression Coefficients for the model containing variable education and experience

earnings. From our estimated model, we can see that one year of school increases the wage per hour by 2.39 Kr, holding all other variables constant. Similarly, one year of working increases the wage per hour by 0.53 Kr, holding all other variables constant.

We perform an F-test to see if education and experience is the one affecting the difference in the wage rather than gender.

$$wage = \beta_0 + \beta_1\ gender + \beta_2\ education + \beta_3\ experience$$

1. **Hypothesis:**

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

2. **Significance level:**

$$\alpha = 0.05$$

3. **Estimators:**

$$\hat{\beta}_1, R^2_{UR}, R^2_R$$

4. **Assumptions:**

$$Large\ n$$

5. **Test statistic:**

$$F_{obs} = \frac{(R^2_{UR} - R^2_R)/m}{(1 - R^2_{UR})/(n-k)} \sim F_{m,n-k}$$

where, - R^2_{UR} is the coefficient of determination for the Unrestricted model - R^2_R is the coefficient of determination for the Restricted model - m is the number of (linear) restrictions (in the null hypothesis) - n is the number of observations - k is the number of regression coefficients (parameters) in the regression line of the Unrestricted model (including the intercept)

6. **Figure under the null and decision rule:**

We reject the null if $F_{obs} > F_{m,n-k}$ or if p-value<0.05.

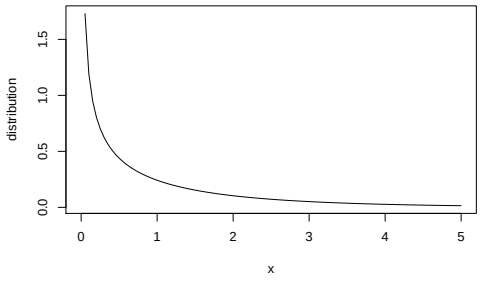

Figure 6: F distribution

7. **Calculations and decision:**

p-value = 1.845e-08 < 0.05 so we reject the null.

8. **Conclusion:**

Under the 5% significance level, we reject the null that gender is not relevant for hourly earnings. Hence, with this model we cam see that there is gender discrimination.

6.3 Quantify the effect of difference in education

We now make the assumption that experience and education are the only variables that matter, meaning that any difference after we have controlled for these variables is the result of discrimination. Thus we use the same model as the previous Mincer's earnings equation.

$$Y_i^M = \alpha_0 + \alpha_1 Edu_i^M + \alpha_2 Exp_i^M \quad + u_i^M$$

$$Y_i^F = \beta_0 + \beta_1 Edu_i^F + \beta_2 Exp_i^F \quad + u_i^F$$

$$\begin{aligned}
\text{Diff} &= \frac{\text{Difference due to gender only}}{\text{Total Difference}} \\
\text{Diff} &= \frac{\text{Discr. levelt}}{\text{Discr. level + Diff Edu.t Discr. effect of Edu+Diff Exp.t Discr effect of Exp.}} \\
\text{Diff} &= \frac{(\alpha_0-\beta_0)+(\alpha_1-\beta_1)\times\overline{Edu^F}+(\alpha_2-\beta_2)\times\overline{Exp^F}}{(\alpha_0-\beta_0)+\alpha_1(\overline{Edu^M}-\overline{Edu^F})+(\alpha_1-\beta_1)\times\overline{Edu^F}+\alpha_2(\overline{Exp^M}-\overline{Exp^F})+(\alpha_2-\beta_2)\times\overline{Exp^F}}
\end{aligned}$$

According to our calculations, the difference in the log of the hourly wage due to gender discrimination is 0.56 Kr per hour. Thus, the difference in the hourly wage would be the exponential of that difference. In our case, this is $e^{0.56}$, or approximately 1.75 Kr per hour. We can conclude that there is an average wage difference of 1.75 Kr per hour due to gender-based discrimination.

7 Instrumental variable method

Let's assume that education is correlated with personal ability, basically how intelligent/capable a person is. Ability as a variable is not in the model, nor the data. This would make education correlated with the error term, thus making OLS biased and inconsistent for estimation. In order to combat this, we could use an instrument as a proxy for education. This proxy needs to be uncorrelated with the error term, and correlated with education.

We want the strongest possible instrument, so we want the highest possible correlation between the instrument and education. We can look to our data set in order to find possible variables for the instrument. For example the education of an individual's parents could be correlated with the individual's own level of education. This is because education is generally correlated with higher levels of income, which in turn can support higher levels of education. Therefore, children of highly educated parents is more likely to be highly educated themselves. These would be the most obvious variables to use. However, there might be other variables, such as parents nationality, in our data set also correlated with education. We will perform tests to check which of these variables would be strong instruments for education.

Table 14: Regression Coefficients for the model: education ~ IV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.467	0.266	28.090	0.000
mother_schooling	0.460	0.069	6.622	0.000
father_schooling	0.289	0.055	5.233	0.000
live	0.525	0.070	7.533	0.000
father_occupation	0.054	0.033	1.657	0.098
mother_occupation	0.186	0.069	2.678	0.007
lived_with	-0.386	0.073	-5.277	0.000

Above, we can see the results from our test. As expected, both parents education is relevant. Variable *live* measures the urban level of where the individual grew up, where 1 is rural and 5 is living in a major city. Living in a city seems to have a positive effect of 0.51 per point, so a person living in a big city (5 points) would have 2.5 more years of education, compared to someone that lived in the countryside. We also looked at the occupation of both parents. The reason why the mothers occupation is significant, but not the fathers, is because the questions are phrased quite differently in the data: Mothers occupation has to do with how long she's worked outside the home, where as the fathers education has to do with what type of occupation he has. Variable *lived_with* measures if an individual lived with both of his/her parents of whether they lived with only one, with the highest numbers meaning that the individual lived in a foster home or an orphanage. One point of *Lived_with* removes on average 0.365 years of education, so an individual growing up in an orphanage (9 "points") would have on average 3.285 less years of education, compared to someone that grew up with both of their parents.

All of our variables except Father_occupation are significant and are thus relevant instruments. However, we need to investigate if the variables are correlated with the error term, which is the unexplained variation. This is what determines the validity of the instruments.

Table 15: Regression Coefficients for the model: residuals ~ IV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.021	0.010	-2.100	0.036
mother_schooling	-0.001	0.003	-0.284	0.777
father_schooling	0.000	0.002	0.088	0.930
live	0.011	0.003	4.020	0.000
mother_occupation	-0.005	0.003	-1.914	0.056
lived_with	0.001	0.003	0.384	0.701

Here we can see that variables *live* and *mother_occupation* are significant, meaning that they are highly correlated with the error term. This would mean that they are **not valid** instruments for estimating education. In conclusions, *mother_schooling*, *Father_schooling* and *lived_with* are all **valid** instruments.

Table 16: Regression Coefficients for the model: education ~ IV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.180	0.196	46.910	0
mother_schooling	0.563	0.070	8.059	0
father_schooling	0.340	0.056	6.065	0
lived_with	-0.389	0.075	-5.212	0

8 Two-Stage least squares

8.1 Taking *education* into account

Table 17: Regression Coefficients for the Two-Stage least squares model containing education variable

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	30.466	5.026	6.062	0.000
gender_1	10.010	1.036	9.662	0.000
IV	1.011	0.453	2.232	0.026

We convert our the fitted values of our instrument into a variable called *IV*. We then replace *education* with the *IV* to create a 2 Stage Least Squares regression. This is to protect the variable *education* from being correlated with the error term, instead replacing it with *IV*. Compared to when we controlled for education, we can see that the difference in hourly wage due to discrimination has increased. In the OLS version, male individuals earned on average 8.94 Kr per hour more than female individuals, holding all other values constant. Now this difference has increased to approximately 10 Kr per hour, when we hold *IV* to be constant. We can also see that the impact of education has decreased, from 1.84 Kr per hour to approximately 1 Kr per hour, holding all other variables constant.

We perform an F-test to see if the instrument for education is the variable affecting the difference in the wage rather than gender.

$$wage = \beta_0 + \beta_1\ gender + \beta_2\ \hat{education}$$

$$\hat{education} = \hat{\alpha}_0 + \hat{\alpha}_1 Mother\_schooling + \hat{\alpha}_2 Father\_schooling + \hat{\alpha}_3 Lived\_with$$

1. **Hypothesis:**

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

2. **Significance level:**

$$\alpha = 0.05$$

3. **Estimators:**

$$\hat{\beta}_1, R^2_{UR}, R^2_R$$

4. **Assumptions:**

$$Large\ n$$

5. **Test statistic:**

$$F_{obs} = \frac{(R^2_{UR} - R^2_R)/m}{(1 - R^2_{UR})/(n-k)} \sim F_{m,n-k}$$

where,

- R^2_{UR} is the coefficient of determination for the Unrestricted model
- R^2_R is the coefficient of determination for the Restricted model
- m is the number of (linear) restrictions (in the null hypothesis)
- n is the number of observations
- k is the number of regression coefficients (parameters) in the regression line of the Unrestricted model (including the intercept)

6. **Figure under the null and decision rule:**

We reject the null if $F_{obs} > F_{m,n-k}$ or if p-value<0.05.

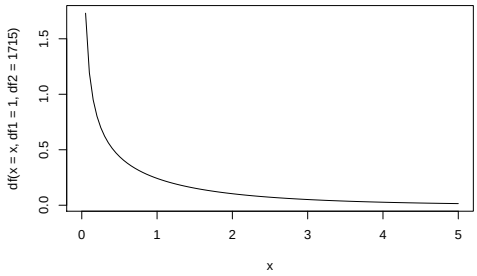

Figure 7: F distribution

7. **Calculations and decision:**

p-value=2.2e-16<0.05 so we reject the null.

8. **Conclusion:**

Under the 5% significance level, we reject the null that gender is not relevant for hourly earnings. Hence, with this model we cam see that there is gender discrimination.

8.2 Taking *education* and *experience* into account

Table 18: Regression Coefficients for the Two-Stage least squares model containing education variable

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.965	5.348	2.611	0.009
gender_1	7.965	1.049	7.592	0.000
IV	1.944	0.460	4.229	0.000
experience	0.368	0.046	8.012	0.000

Now we control for both experience and education, where IV is the instrument of estimation for *education*. According to the estimated model, a male individual would earn 7.96 Kr per hour more than a female individual, holding all other variables constant. Similarly education increases hourly wage by 1.94 Kr per year studied, holding all other variables constant. One extra year of working provides 0.368 Kr per hour, holding all other variables constant.

We perform an F-test to see if the instrument for education and experience are the variables affecting the difference in the wage rather than gender.

$$wage = \beta_0 + \beta_1\ gender + \beta_2\ \hat{education} + \beta_3\ experience$$

$$\hat{education} = \hat{\alpha}_0 + \hat{\alpha}_1 Mother\_schooling + \hat{\alpha}_2 Father\_schooling + \hat{\alpha}_3 Lived\_with$$

1. **Hypothesis:**

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

2. **Significance level:**

$$\alpha = 0.05$$

3. **Estimators:**

$$\hat{\beta}_1, R^2_{UR}, R^2_R$$

4. **Assumptions:**

$$Large\ n$$

5. **Test statistic:**

$$F_{obs} = \frac{(R^2_{UR} - R^2_R)/m}{(1 - R^2_{UR})/(n-k)} \sim F_{m,n-k}$$

where,

- R^2_{UR} is the coefficient of determination for the Unrestricted model
- R^2_R is the coefficient of determination for the Restricted model
- m is the number of (linear) restrictions (in the null hypothesis)
- n is the number of observations
- k is the number of regression coefficients (parameters) in the regression line of the Unrestricted model (including the intercept)

6. **Figure under the null and decision rule:**

We reject the null if $F_{obs} > F_{m,n-k}$ or if p-value<0.05.

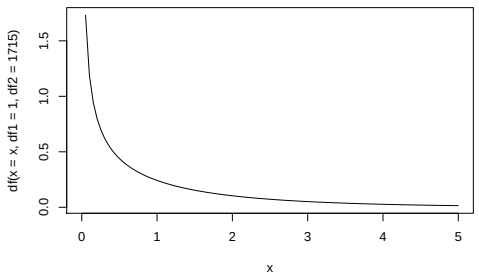

Figure 8: F distribution

7. **Calculations and decision:**

p-value=2.2e-16<0.05 so we reject the null.

8. **Conclusion:**

Under the 5% significance level, we reject the null that gender is not relevant for hourly earnings. Hence, with this model we can see that there is gender discrimination.

8.3 Quantify the effect of difference in gender discrimination with 2SLS

$$Y_i^M = \alpha_0 + \alpha_1 IV_i^M + \alpha_2 Exp_i^M \quad + u_i^M$$

$$Y_i^F = \beta_0 + \beta_1 IVi^F + \beta_2 Exp_i^F \quad + u_i^F$$

$$\begin{aligned}
\text{Diff} &= \frac{\text{Difference due to gender only}}{\text{Total Difference}} \\
\text{Diff} &= \frac{\text{Discr. levelt}}{\text{Discr. level + Diff IV.t Discr. effect of IV+Diff Exp.t Discr effect of Exp.}} \\
\text{Diff} &= \frac{(\alpha_0-\beta_0)+(\alpha_1-\beta_1)\times\overline{IV^F}+(\alpha_2-\beta_2)\times\overline{Exp^F}}{(\alpha_0-\beta_0)+\alpha_1(IV^M-IV^F)+(\alpha_1-\beta_1)\times\overline{IV^F}+\alpha_2(\overline{Exp^M}-\overline{Exp^F})+(\alpha_2-\beta_2)\times\overline{\overline{Exp^F}}}
\end{aligned}$$

According to our calculations, the difference in the log of the hourly wage due to gender discrimination is 0.47 Kr per hour. Thus, the difference in the hourly wage would be the exponential of that difference. In our case, this is $e^{0.47}$, or approximately 1.27 Kr per hour. We can conclude that there is an average wage difference of 1.27 Kr per hour due to gender-based discrimination.

9 Comparison between OLS and 2SLS

Table 19: Table for comparison, task 4a and 6a

Model	Gender	Education
OLS	8.95	1.84
2SLS	10.01	1.01
Difference	1.06	-0.83

In the above table, we can see the differences in estimation between OLS and 2SLS when taking controlling for education. The difference in hourly wage is 1.06 Kr more for 2SLS and the difference in the estimated hourly wage due to one year of education is 0.83 Kr less for 2SLS. The difference in numbers is quite small. 2SLS seems to increase the effect of gender and decrease the effect of education.

Table 20: Table for comparison, task 4b and 6b

Model	Gender	Education	Experience
OLS	5.61	2.39	0.53
2SLS	7.96	1.94	0.36
Difference	2.35	-0.45	-0.16

In this table, we compare OLS and 2SLS when controlling for experience and education. We can see the same trend we saw in table 5: Increase hourly wage due to gender difference and decrease in hourly wage due to experience and education. The difference in gender is larger compared to model 5, but the difference in education is smaller. Hourly wage due to gender is 2.35 Kr more when using 2SLS, holding all other variables constant. Hourly wage due from one year of education is 0.45 Kr less for 2SLS and hourly wage due to one year of work is 0.16 Kr less.

Table 21: Table for comparison, task 4c and 6c

Model	Log_wage	Wage
OLS	0.44	1.19
2SLS	0.47	1.27
Difference	0.03	0.08

From the above table, both models show that there is a difference in hourly wage due to gender based discrimination, only differing slightly in values. According to OLS, the log difference in hourly wage due to gender is 0.44 Kr per hour, and the exponential difference is 1.19 Kr per hour. For 2SLS the log difference in hourly wage due to gender is 0.47 Kr per hour, and the exponential difference is 1.27 Kr per hour. We can see that the difference in log hourly wage is 0.03 Kr per hour. The exponential difference is 0.08 Kr per hour. We conclude that there is a slight difference in estimation between the two models.

10 One more year of *education* and *experience*

To figure out if one more year of education is more worth to a man than to a woman, we established a two stage least square regression model which is grouped by gender such that it is possible to test this hypothesis.

$$ln(wage) = \beta_0 + \beta_1^{Male}\, IV(education) + \beta_2^{Female}\, IV(education) + \beta_3^{Male}\, experience + \beta_4^{Female}\, experience + u$$

Table 22: Regression Coefficients for the Two-Stage least squares model grouped by gender

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.303	0.078	42.584	0
IV:genderFemale	0.024	0.007	3.500	0
IV:genderMale	0.038	0.007	5.703	0
genderFemale:experience	0.007	0.001	6.708	0
genderMale:experience	0.007	0.001	8.418	0

10.1 Test one more year of *education*

Hypothesis:

$$H_0 : \beta_1 = \beta_2$$

$$H_1 : \beta_1 \neq \beta_2$$

Table 23: Linear hypothesis test for education

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1714	162.14				
1713	159.11	1	3.02	32.55	0

Under the 5% significance level, we **reject** the null hypothesis that one more year of education is equally worth to a man than to a woman. Thus we conclude that one more year of education is more worth to a man than to a woman. [9]

10.2 Test one more year of *experience*

Also, given this model, we can also test if one more year of experience is more worth to a man than a woman.

Hypothesis:

[9] We hide the test template here since it is a F-test of regression coefficient similar to the previous.

$$H_0 : \beta_3 = \beta_4$$

$$H_1 : \beta_3 \neq \beta_4$$

Table 24: Linear hypothesis test for experience

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1714	159.12				
1713	159.11	1	0.01	0.12	0.73

Under the 5% significance level, we **can not reject** the null hypothesis that one more year of experience is equally worth to a man than to a woman.

11 Conclusion

In this article, we get into a deep study on the gender discrimination regarding wages in Sweden. Using the data set of Households Economic Living Conditions, we perform a Mincer's earnings equation and then solve the problem of heteroskedasticity using **Box-Cox transformation**.

To figure out if there is a difference in wages between men and women, we use **Non-parametric two-samples Wilcoxon rank test** without regression and T-test within regression. Both tests show that it is statistically significant that on average men earn 9.97 Kr more per hour, compared to women.

To investigate if the difference in wages is caused by something other than just discrimination, we taking *education* and *experience* into account. And we found that even we add these two factors into our model, there is still some remaining effect which can be explained by *gender*. After calculating, We conclude that there is an average wage difference of 1.75 Kr per hour due to gender-based discrimination.

There might be a problem that education is correlated with personal ability, basically how intelligent/capable a person is. Ability as a variable is not in the model, nor the data. To solve the problem, we managed to pick out the strongest valid instrument from plenty of possible instruments, and then perform a **two-stage least square regression** instead of OLS. After comparison, we conclude that there is a slight difference in estimation between the two models.

To figure out if one more year of *education* or *experience* is more worth to a man than to a woman, we established another regression model which is grouped by gender. After performing a statistical test, we are sure that one more year of *education* is more worth to a man than to a woman. But for *experience*, there is insufficient evidence to support the opinion.

12 Appendix

12.1 Regression Outcome for the Mincer's earnings equation

```
Call:
lm(formula = log(wage) ~ education + experience + experience_sq,
    data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63439 -0.14027 -0.01093  0.12203  1.93652

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.969e+00  3.262e-02  91.008  < 2e-16 ***
education      4.297e-02  2.061e-03  20.850  < 2e-16 ***
experience     2.510e-02  2.042e-03  12.293  < 2e-16 ***
experience_sq -3.005e-04  4.229e-05  -7.105 1.76e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2792 on 1714 degrees of freedom
Multiple R-squared:  0.2855,    Adjusted R-squared:  0.2842
F-statistic: 228.3 on 3 and 1714 DF,  p-value: < 2.2e-16
```

12.2 Regression Outcome after Box-Cox transformation on our dependent variable

```
Call:
lm(formula = dat$log_wage_BC ~ dat$education + dat$experience +
    dat$experience_sq)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89945 -0.06026 -0.00250  0.05718  0.73103

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.390e+00  1.467e-02  94.769  < 2e-16 ***
dat$education      1.900e-02  9.264e-04  20.511  < 2e-16 ***
dat$experience     1.141e-02  9.178e-04  12.434  < 2e-16 ***
dat$experience_sq -1.388e-04  1.901e-05  -7.299 4.41e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1255 on 1714 degrees of freedom
Multiple R-squared:  0.2816,    Adjusted R-squared:  0.2804
F-statistic:   224 on 3 and 1714 DF,  p-value: < 2.2e-16
```

12.3 Regression Outcome for the regression model including dummy variable

```
Call:
lm(formula = wage ~ gender_1, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-35.54  -9.54  -3.57   4.43 406.43

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.5670     0.7274  57.147   <2e-16 ***
gender_1      9.9750     1.0372   9.618   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.49 on 1716 degrees of freedom
Multiple R-squared:  0.05115,   Adjusted R-squared:  0.05059
F-statistic:  92.5 on 1 and 1716 DF,  p-value: < 2.2e-16
```

12.4 Regression Outcome for the regression taking *education* into account

```
Call:
lm(formula = wage ~ gender_1 + education, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-45.87  -8.98  -2.48   4.71 389.29

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.8791     1.7105  12.791   <2e-16 ***
gender_1      8.9499     0.9959   8.987   <2e-16 ***
education     1.8416     0.1462  12.600   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.57 on 1715 degrees of freedom
Multiple R-squared:  0.1315,    Adjusted R-squared:  0.1305
F-statistic: 129.9 on 2 and 1715 DF,  p-value: < 2.2e-16
```

12.5 Regression Outcome for the regression taking *education* and *experience* into account

```
Call:
lm(formula = wage ~ gender_1 + education + experience, data = dat)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-44.56  -7.72  -2.20   4.41 384.12

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.84741    2.04673   3.346 0.000839 ***
gender_1     5.61287    0.99292   5.653 1.84e-08 ***
education    2.39506    0.14723  16.268  < 2e-16 ***
experience   0.53688    0.04372  12.281  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.72 on 1714 degrees of freedom
Multiple R-squared:  0.2018,    Adjusted R-squared:  0.2004
F-statistic: 144.4 on 3 and 1714 DF,  p-value: < 2.2e-16
```

12.6 Test one more year of *education*

```
Linear hypothesis test

Hypothesis:
IV:genderFemale - IV:genderMale = 0

Model 1: restricted model
Model 2: log_wage ~ IV:gender + experience:gender

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1714 162.13
2   1713 159.11  1    3.0233 32.549 1.367e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12.7 Test one more year of *experience*

```
Linear hypothesis test

Hypothesis:
genderFemale:experience - genderMale:experience = 0

Model 1: restricted model
Model 2: log_wage ~ IV:gender + experience:gender

  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1   1714 159.12
2   1713 159.11  1  0.010868 0.117 0.7324
```

12.8 Code for the whole study

```
## ----setup, include=FALSE, message=FALSE----------------------------------------------------------
knitr::opts_chunk$set(fig.pos = 'H', echo = FALSE, warning = FALSE, comment = "")
library(knitr)
library(ggplot2)
library(tidyverse)
library(caret)
library(fastDummies)
library(car)
options(knitr.kable.NA = '')
dat = read.csv("b4 - hwa - wage - data.csv")


## -------------------------------------------------------------------------------------------------
names(dat) = c("birth_year", "gender", "citizenship", "language", "lived_with", "father_schooling", "


## -------------------------------------------------------------------------------------------------
dat[is.na(dat)] = 0
means = c()
variances = c()
mins = c()
maxs = c()
for(i in 1:length(colnames(dat))){
   means[i] = mean(dat[[colnames(dat)[i]]])
   variances[i] = var(dat[[colnames(dat)[i]]])
   mins[i] = min(dat[[colnames(dat)[i]]])
   maxs[i] = max(dat[[colnames(dat)[i]]])
}

stats = data.frame(colnames(dat), means, variances, mins, maxs)
kable(stats, caption = "Preliminary Analysis for the variables", digits = 2)


## -------------------------------------------------------------------------------------------------
dat = dat %>%
   mutate(log_wage = log(wage)) %>%
   mutate(experience_sq = experience^2)
m1 = lm(log(wage) ~ education + experience + experience_sq, data = dat)
sum = summary(m1)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the regression model", digits = 2)


## ----out.width="70%", fig.align="center", fig.cap="Residuals Plot for the regression model"--------
data.frame(index = 1:length(m1$residuals), residuals = m1$residuals) %>%
   ggplot(aes(x = index, y = residuals)) +
   geom_point(color = "red", alpha = 0.5) +
   theme_bw()
```

```r
## ------------------------------------------------------------------------------------------------
bptest = lmtest::bptest(m1)
table = data.frame(BP = bptest$statistic, df = bptest$parameter, p_value = bptest$p.value)
rownames(table) = "bptest"
kable(table, caption = "Brausch-Pagan test", digits = 2)


## ----out.width="50%", fig.align="center", fig.cap="QQ-plot for the model before Box-Cox transformat
qqnorm(m1$residuals)
qqline(m1$residuals, col = "red")


## ------------------------------------------------------------------------------------------------
BClog_wage <- BoxCoxTrans(dat$log_wage)
dat <- cbind(dat, log_wage_BC = predict(BClog_wage, dat$log_wage))
m2 <- lm(dat$log_wage_BC ~ dat$education + dat$experience + dat$experience_sq)


## ------------------------------------------------------------------------------------------------
bptest = lmtest::bptest(m2)
table = data.frame(BP = bptest$statistic, df = bptest$parameter, p_value = bptest$p.value)
rownames(table) = "bptest"
kable(table, caption = "Brausch-Pagan test for the model after Box-Cox transformation", digits = 2)


## ----out.width="50%", fig.align="center", fig.cap="QQ-plot for the model after Box-Cox transformati
qqnorm(m2$residuals)
qqline(m2$residuals, col = "red")


## ------------------------------------------------------------------------------------------------
sum = summary(m2)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the model after Box-Cox transformation", digits = 2)


## ----out.width="70%", fig.align="center", fig.cap="Difference in Wages between men and women"------
dat_gender = dat
for(i in 1:nrow(dat_gender))
{
   if(dat_gender$gender[i] == 1)
   {
      dat_gender$gender[i] = "Male"
   }
   if(dat_gender$gender[i] == 2)
   {
      dat_gender$gender[i] = "Female"
   }
   if(dat_gender$wage[i] > 200)
   {
      dat_gender$wage[i] = 0
```

```
   }
}
dat_gender %>%
   ggplot(aes(x = gender, fill = gender, y = wage)) +
   geom_boxplot()


## ------------------------------------------------------------------------------------------------------
m = aggregate(wage ~ gender, dat, mean)
m$gender = c("Male", "Female")
kable(m, caption = "Comparison in mean of wages between male and female workers", digits = 2)
difference = m$wage[1]-m$wage[2]


## ------------------------------------------------------------------------------------------------------
# Shapiro-Wilk normality test for Men's wage
shapiro = with(dat_gender, shapiro.test(wage[gender == "Male"]))
table = data.frame(method = shapiro$method, statistic = shapiro$statistic, p_value = shapiro$p.value)
rownames(table) = "Male"
kable(table, caption = "Shapiro-Wilk normality test for Men's wage", digits = 2)

# Shapiro-Wilk normality test for Women's wage
shapiro = with(dat_gender, shapiro.test(wage[gender == "Female"]))
table = data.frame(method = shapiro$method, statistic = shapiro$statistic, p_value = shapiro$p.value)
rownames(table) = "Female"
kable(table, caption = "Shapiro-Wilk normality test for woman's wage", digits = 2)


## ------------------------------------------------------------------------------------------------------
wilcox = wilcox.test(wage ~ gender, data = dat_gender, exact = FALSE)
table = data.frame(method = wilcox$method, statistic = shapiro$statistic, p_value = shapiro$p.value)
rownames(table) = NULL
kable(table, caption = "Non parametric two-samples Wilcoxon rank test for woman's wage", digits = 3)


## ------------------------------------------------------------------------------------------------------
dat <- fastDummies::dummy_cols(dat, select_columns = "gender")
m3 <- lm(wage ~ gender_1, dat)
sum = summary(m3)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the model including dummy variable", digits = 2)


## ------------------------------------------------------------------------------------------------------
table = data.frame(Method = c("Difference of Means","Regression"),
                   Result = c("9.97", "9.97"),
                   Test = c("Non parametric two-samples Wilcoxon rank test", "T-test"),
                   Conclusion = c("significant", "significant"))


# Then for PDF:
```

```r
kable(table, caption = "Wage difference between male and female", digits = 2)


## ------------------------------------------------------------------------------------------------
m4 <- lm(wage ~ gender_1 + education, dat)
sum = summary(m4)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the model containing education variable", digits = 2)


## ----out.width="40%", fig.align="center", fig.cap="F distribution"-------------------------------
x = seq(0, 5, length = 100)
distribution = df(x = x, df1 = 1, df2 = 1715)
plot(x, distribution,type="l")


## ------------------------------------------------------------------------------------------------
m5 <- lm(wage ~ gender_1 + education + experience, dat)
sum = summary(m5)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the model containing variable education and experienc


## ----out.width="40%", fig.align="center", fig.cap="F distribution"-------------------------------
x = seq(0, 5, length = 100)
distribution = df(x = x, df1 = 1, df2 = 1715)
plot(x, distribution,type="l")


## ------------------------------------------------------------------------------------------------
dat_m = subset(dat, gender == "1")
dat_f = subset(dat, gender == "2")

model_m = lm(log_wage ~ education + experience, dat_m)
model_f = lm(log_wage ~ education + experience, dat_f)

difference = ((model_m$coefficients[1]-model_f$coefficients[1])+(model_m$coefficients[2]-model_f$coef


## ------------------------------------------------------------------------------------------------
fit1 <- lm(education ~ mother_schooling + father_schooling + live + father_occupation + mother_occupa
sum = summary(fit1)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the model: education ~ IV", digits = 3)


## ------------------------------------------------------------------------------------------------
fit2 <- lm(m2$residuals ~ mother_schooling + father_schooling + live + mother_occupation + lived_with
sum = summary(fit2)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the model: residuals ~ IV", digits = 3)
```

```r
## ------------------------------------------------------------------------------------------------
sls <- lm(education ~ mother_schooling + father_schooling + lived_with, dat)
sum = summary(sls)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the model: education ~ IV", digits = 3)
IV = fitted.values(sls)
dat = cbind(dat, IV)


## ------------------------------------------------------------------------------------------------
tsls2 = lm(wage ~ gender_1 + IV, dat)
sum = summary(tsls2)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the Two-Stage least squares model containing educatio


## ----out.width="40%", fig.align="center", fig.cap="F distribution"-------------------------------
x = seq(0, 5, length = 100)
plot(x, df(x = x, df1 = 1, df2 = 1715),type="l")


## ------------------------------------------------------------------------------------------------
tsls3 = lm(wage ~ gender_1 + IV + experience, dat)
sum = summary(tsls3)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the Two-Stage least squares model containing educatio


## ----out.width="40%", fig.align="center", fig.cap="F distribution"-------------------------------
x = seq(0, 5, length = 100)
plot(x, df(x = x, df1 = 1, df2 = 1715),type="l")


## ------------------------------------------------------------------------------------------------
sls_m = lm(education ~ mother_schooling + father_schooling + lived_with, dat_m)
sls_f = lm(education ~ mother_schooling + father_schooling + lived_with, dat_f)

IV_m = fitted.values(sls_m)
IV_f = fitted.values(sls_f)

dat_m = cbind(dat_m, IV_m)
dat_f = cbind(dat_f, IV_f)

model_m = lm(log_wage_BC ~ IV_m + experience + experience_sq, dat_m)
model_f = lm(log_wage_BC ~ IV_f + experience + experience_sq, dat_f)

difference = ((model_m$coefficients[1]-model_f$coefficients[1])+(model_m$coefficients[2]-model_f$coef


## ------------------------------------------------------------------------------------------------
```

```r
table_a = data.frame(Model = c("OLS","2SLS", "Difference"),
                     Gender = c("8.95", "10.01", "1.06"),
                     Education = c("1.84", "1.01", "-0.83"))

table_b = data.frame(Model = c("OLS","2SLS", "Difference"),
                     Gender = c("5.61", "7.96", "2.35"),
                     Education = c("2.39", "1.94", "-0.45"),
                     Experience = c("0.53", "0.36", "-0.16"))

table_c = data.frame(Model = c("OLS","2SLS", "Difference"),
                     Log_wage = c("0.44", "0.47", "0.03"),
                     Wage = c("1.19", "1.27", "0.08"))


## ------------------------------------------------------------------------------------------------
kable(table_a, caption = "Table for comparison, task 4a and 6a")


## ------------------------------------------------------------------------------------------------
kable(table_b, caption = "Table for comparison, task 4b and 6b")


## ------------------------------------------------------------------------------------------------
kable(table_c, caption = "Table for comparison, task 4c and 6c")


## ------------------------------------------------------------------------------------------------
mx = lm(log_wage ~ IV:gender + experience:gender, data = dat_gender)
sum = summary(mx)
sum$coefficients %>%
   kable(caption = "Regression Coefficients for the Two-Stage least squares model grouped by gender",


## ------------------------------------------------------------------------------------------------
test = linearHypothesis(mx, "IV:genderFemale=IV:genderMale")
as.data.frame(test) %>%
   kable(caption = "Linear hypothesis test for education", digits = 2)


## ------------------------------------------------------------------------------------------------
test = linearHypothesis(mx, "genderFemale:experience=genderMale:experience")
as.data.frame(test) %>%
   kable(caption = "Linear hypothesis test for experience", digits = 2)


## ------------------------------------------------------------------------------------------------
summary(m1)


## ------------------------------------------------------------------------------------------------
summary(m2)
```

```r
## ------------------------------------------------------------------------------------------------
summary(m3)


## ------------------------------------------------------------------------------------------------
summary(m4)


## ------------------------------------------------------------------------------------------------
summary(m5)


## ------------------------------------------------------------------------------------------------
linearHypothesis(mx, "IV:genderFemale=IV:genderMale")


## ------------------------------------------------------------------------------------------------
linearHypothesis(mx, "genderFemale:experience=genderMale:experience")
```