TMA4315: Compulsory exercise 1

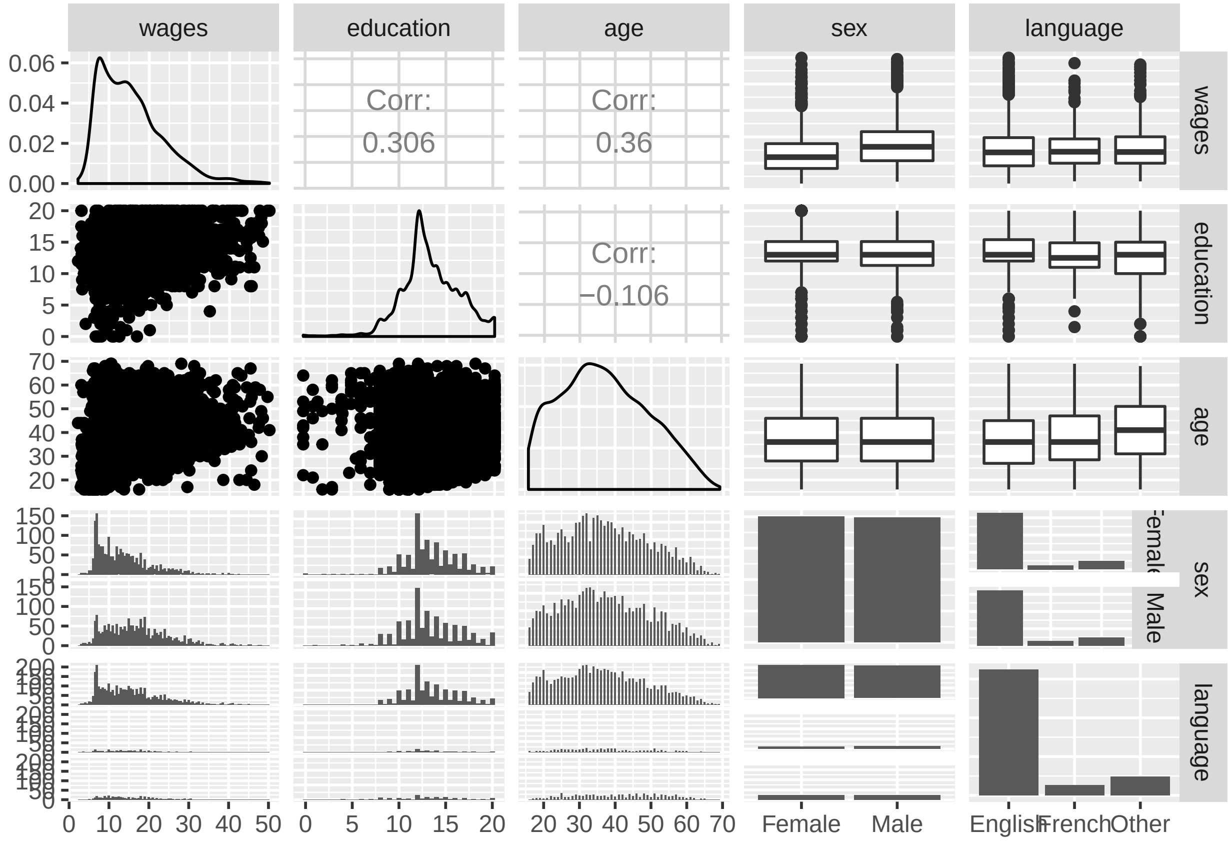
Group 4: Adrian Bruland og Mathias Opland

28.09.2018

Part 1

a)

```
library(GGally)
ggpairs(SLID, lower = list(combo = wrap(ggally_facethist, binwidth = 0.5)))
```



From the top row we can see that the wage variable shows a noticeable correlation with education, age and sex. Language on the other hand doesn't seem to have a large impact on the wages. One can see that people with high level (20) of education are distributed over the whole span of wages, but low educated people are centred around low wages, with very few or none at high level of wages.

The age correlates to wages in that there are people from all age categories that have low wages, but that people are more likely to have a higher wage around age 40. Visually, it seems that the average wage decreases with every age bracket above 40, such that those aged 60 have a lower average wage than 50-year-olds, and so forth. The correlation is posivite and substantial, so the overall trend is that higher age correlates to higher wages. There is a correlation between age and education, in that the education level decreases as the age increases. It is a known fact that the average education level have increased over the last 50 years, which correlates with the data set.

As for the sex variable, males have a somewhat higher median wage, and the first and third wage quartiles in males are respectively higher than those in females. The wage outliers among males also tend to earn more than those in females, suggesting that that the upper few percents of earners will tend to be male.

First we assume that there is a linear relationship between the covariates, ie. the relationship can be expressed as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

In order to perform a multiple linear regression analysis, we must make the following assumptions: First, the response, which is wage, is a linear combination of the covariates, and errors are additive onto the linear combination, i.e. $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. Second, that we have homoscedastic error variance and uncorrelated errors, i.e. Cov( )=$^2\mathbf{I}$, Cov($_i$, $_j$)=0. For this model, this means that along age, sex, language and education, the variance in wage is the same for any observation, or set of observations.

In addition, the model must be a "normal model" (cf. Module 2 in the course) in order for us to perform linear regression. In a normal linear model, we assume that the errors are independent and normally distributed, with the same variance for all errors: $\varepsilon \sim \mathbf{N_n}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$.

Part 2

a) Here is a print of the mylm-package and the built-in lm-package.

```
library(mylm)
model1 <- mylm(wages ~ education, data = SLID)
print(model1)
```

```
## Call:
## mylm(formula = wages ~ education, data = SLID)
##
## Coefficients:
##      (Intercept) education
## [1,]      4.9717   0.79231
```

```
model1b <- lm(wages ~ education, data = SLID)
print(model1b)
```

```
##
## Call:
## lm(formula = wages ~ education, data = SLID)
##
## Coefficients:
## (Intercept)    education
##      4.9717       0.7923
```

As shown, the two packages produces the same information when using the print-function. The result is a linear regression with wage as response, and education as the only covariate. The intercept can be interpreted as the expected wage when one have no edjucation, but as shown in the plot in task 1, there is none or very few that has under 5 years of edjucation, and few that has under 10 years of edjucation. Thus the regression will not be very informative when under 5 years of edjucation, as the regression is based on very little or none data from under this value.

b) Here is a print out of tha covariance matrix defined as:

$$\Sigma = E\Big[(X - E[X])(X - E[X])^T\Big] = \frac{1}{n}\Big(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)\Big)(X^T X)^{-1}$$

```r
print.default(model1$covariance_matrix)
```

```
##              (Intercept)    education
## (Intercept)  0.28532651 -0.020338410
## education   -0.02033841  0.001524956
```

```r
summary(model1)
```

```
## Call:
## mylm(formula = wages ~ education, data = SLID)
##
## Residuals:
## Min: [1] -17.688
## 1Q:    25%
## -5.822
## Median: [1] -1.039
## 3Q:   75%
## 4.148
## Max: [1] 34.19
##
## Coefficients:
## Estimate               [,1]
## (Intercept) 4.97169
## education   0.79231
##
## Std. Error               [,1]
## (Intercept) 0.53429
## education   0.03906
##
## z-value              [,1]
## (Intercept)  9.305
## education   20.284
##
## Pr(>|z|)                    [,1]
## (Intercept) 1.337833e-20
## education   1.774739e-91
##
## R-squared:[1] 0.09358627
## Adjusted R-squared:[1] 0.09335881
```

```r
summary(model1b)
```

```
##
## Call:
## lm(formula = wages ~ education, data = SLID)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.688  -5.822  -1.039   4.148  34.190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.97169    0.53429   9.305   <2e-16 ***
## education    0.79231    0.03906  20.284   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.492 on 3985 degrees of freedom
## Multiple R-squared:  0.09359,    Adjusted R-squared:  0.09336
## F-statistic: 411.4 on 1 and 3985 DF,  p-value: < 2.2e-16
```

The intercept estimate as shown in the print out is 4.97169 and the estimated standard error is 0.53429. For the regression coefficient the estimate is 0.79231 and the estimated standard error is 0.03906. Using a Z-test, we get:

$$P(Z \leq |z|) = 2 \cdot \Phi(-|Z|), \qquad Z = \frac{x-\mu}{\sigma}$$

Here we use that the normal distribution is two-sided, but we also normalize the data with mean $\mu = 0$ and variance $\sigma = 1$. Thus we know that we only have to look at one side of the distribution, and multiply by two afterwords. In our case, the H_0 hypotheses is that μ is zeros, and thus we get $Z = x/\sigma$:

```
cat("Z-values for the regression coefficients: ")
```

```
## Z-values for the regression coefficients:
```

```
print.default(model1$z_value)
```

```
##                  [,1]
## (Intercept)  9.305166
## education   20.284158
```

Computing the p-values are done by the integral

$$P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2} dt,$$

but as it is not analytic solvable for a unknown z, we use the built-in function pnorm, which gives the values:

```
cat("P-values for the regression coefficients: ")
```

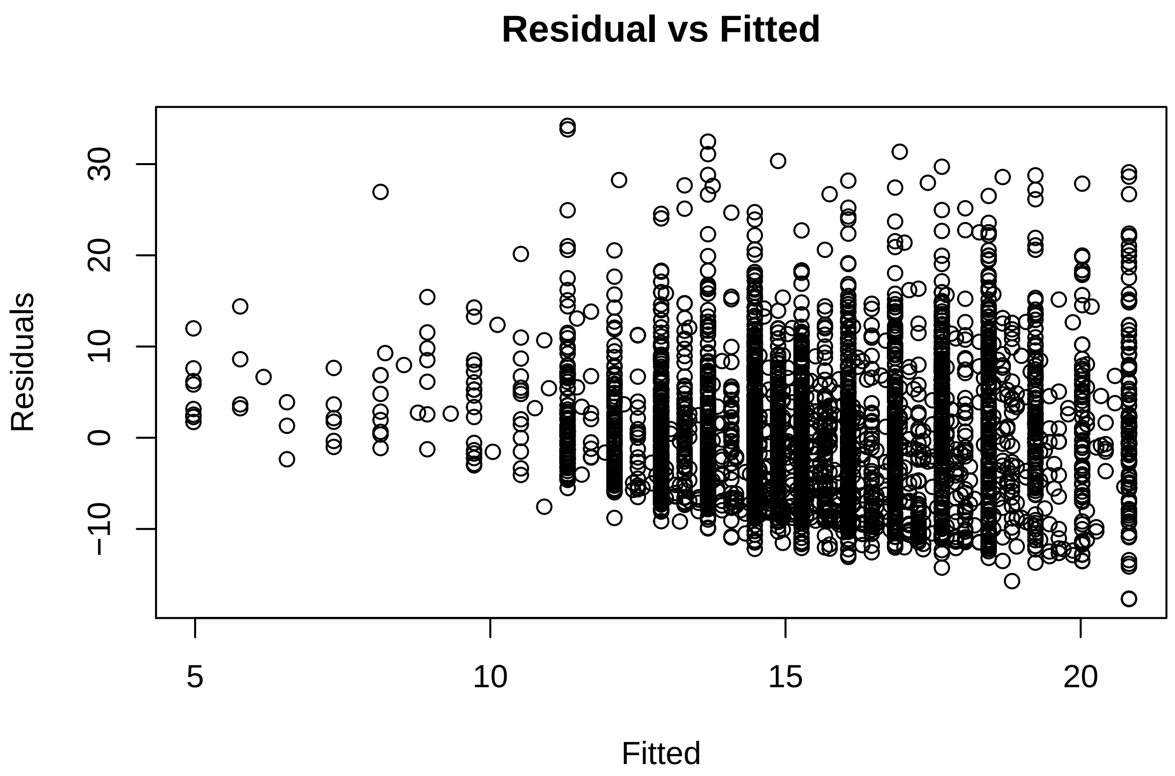
```
## P-values for the regression coefficients:
```

```
print.default(model1$p_value)
```

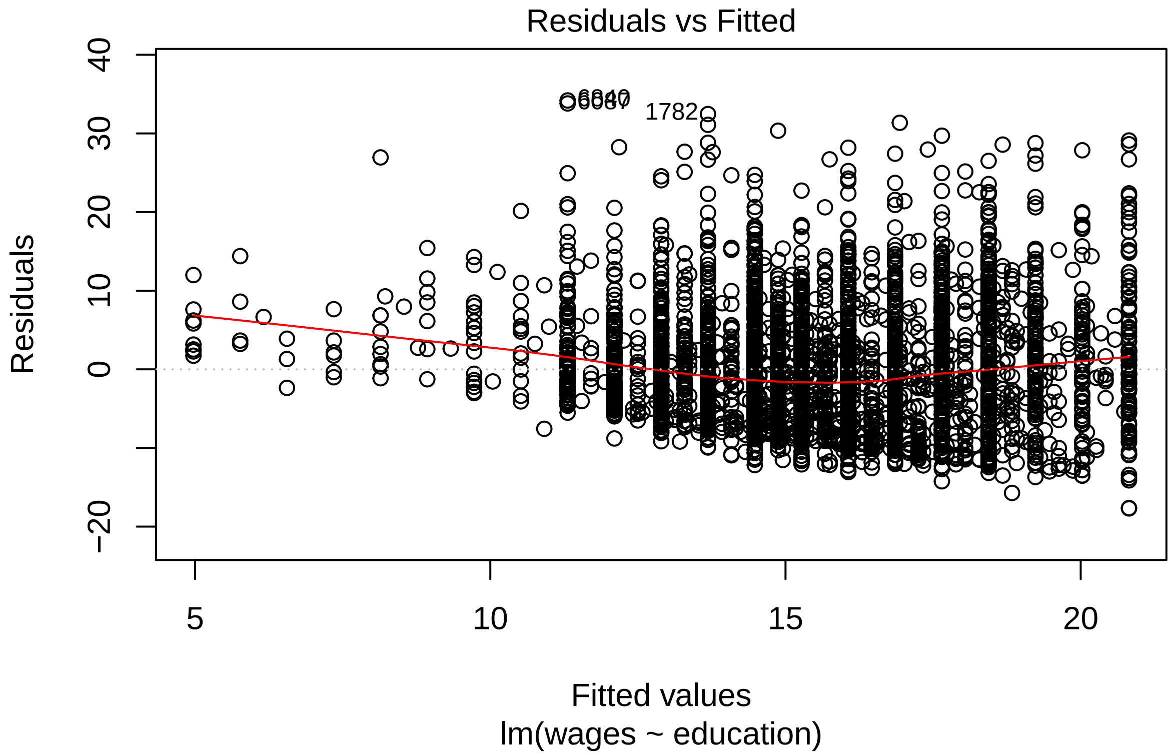
```
##                     [,1]
## (Intercept) 1.337833e-20
## education   1.774739e-91
```

This suggests that both the intercept and the regression coefficients are significant. The usual level to determen if a coefficient is siginficant or not is a 95%-confidence interval where $P(Z \leq z) < 0.05$. If true, the parameter in question is significant at a 5%-level, which both our parameters are in this case. **c)**

```
library(ggplot2)
plot(model1)
```



```
plot(model1b, which = c(1))
```



The plot shows the residuals plotted against the fitted values. A residual plot shows if the linear regression is appropriate for the data. A random distribution around the horisontal-axis suggest that there is no systematical error in the regression, and that a linear regression is appropriate. On the other side, if the residuals follow a systematix distribution around the horisontal-axis, there is likely that the relationship between the covariates and the response is non-linear. In this plot we can see that the points lies closer to zero at at the start of the x-axis, and as we move along, the points spread out more. However there is also a lot more points as we moce along the x-axis, which explains why some of the points also lies further from zeros. On the other hand, we can see that the points seem to lie much closer under the zero line, and that there is much more spread out on above zero. This is not a random pattern, and lead us to the conclusion that the regression is not a very good fit for the datapoints. From this plot we think that there is easier to predict some sort of minimum wage given a certain education level, but above this wage-level there is more variance, and therefore not as defined maximum wage-level given a edjucation level.

d)

- The residual sum of squares (SSE) for this model is computed as: $SSE = \sum_{i=1}^{n} \epsilon_i^2$ where $\epsilon_i = (I - H)Y$.

```
## SSE:
```

```
## [1] 223694.3
```

In this case The degrees of freedom for this model is the number of dimesions that are free, which can be expressed by $df = n - p - 1$, where n is the number of datapoints (in this caes 3987) and p is the number of explanatory parameters (in this case 1). Thus $df = 3985$.

- The total sum of squares (SST) is $\sum_{i=1}^{n}(Y - mean(Y))^2$

```
## SST:
```

```
## [1] 246790.5
```

The χ^2-statistics is computed by $\chi^2 = df \cdot \frac{SST-SSE}{SSE}$, which gives:

```
## Chi-squared statistics from mylm-package:
```

```
## [1] 411.4471
```

```
## Chi-squared p-value:
```

```
## [1] 1.774739e-91
```

```
##
## F-statistics from lm-package:
```

```
##    value
## 411.4471
```

```
## F p-value:
```

```
## [1] 3.872986e-87
```

e) The coefficient of determination R^2 is computed $R^2 = SSR/SST = 1 - SSE/SST$, where SSR/SST can be interpreted as how much of the total variability in the data (SST) is described by the regression (SSR). We want the regression to describe the variability in the data, and thus a R^2-value as close to 1 as possible is desired $(0 \leq R^2 \leq 1)$. In a simple linear regression, R^2 is the squared correlation coefficient between the response and the predictor (in this case wage and education*$\hat{\beta}$), and for multiple linear regression R^2 is the squared correlation coefficient between the response and the predicted response. The value for our model is:

```
## R-squared:
```

```
## [1] 0.09358627
```

This value is not very good, and suggests that education alone is not a very good predictor for the wage, and that it doesn't desctibe the variability in the data.

Part 3

a)

```
# R code for part 3a)
library(mylm)
model2 <- mylm(wages ~ education + age, data = SLID)
print(model2)
```

```
## Call:
## mylm(formula = wages ~ education + age, data = SLID)
##
## Coefficients:
##      (Intercept) education     age
## [1,]     -6.0217   0.90146 0.25709
```

b)

```
# R code for part 3b)
summary(model2)
```

```
## Call:
## mylm(formula = wages ~ education + age, data = SLID)
##
## Residuals:
## Min: [1] -24.303
## 1Q:    25%
## -4.495
## Median: [1] -0.807
## 3Q:   75%
## 3.674
## Max: [1] 37.628
##
## Coefficients:
## Estimate                [,1]
## (Intercept) -6.02165
## education    0.90146
## age          0.25709
##
## Std. Error                [,1]
## (Intercept) 0.618924
## education   0.035760
## age         0.008951
##
## z-value              [,1]
## (Intercept) -9.729
## education   25.209
## age         28.721
##
## Pr(>|z|)                     [,1]
## (Intercept)  2.262949e-22
## education   3.203056e-140
## age         2.076930e-181
##
## R-squared:[1] 0.2490697
## Adjusted R-squared:[1] 0.2486927
```

c)

```
# R code for part 3c)
model2a <- mylm(wages ~ education, data = SLID)
model2b <- mylm(wages ~ age, data = SLID)
summary(model2a)
```

```
## Call:
## mylm(formula = wages ~ education, data = SLID)
##
## Residuals:
## Min: [1] -17.688
## 1Q:    25%
## -5.822
## Median: [1] -1.039
## 3Q:   75%
## 4.148
## Max: [1] 34.19
##
## Coefficients:
## Estimate               [,1]
## (Intercept) 4.97169
## education   0.79231
##
## Std. Error               [,1]
## (Intercept) 0.53429
## education   0.03906
##
## z-value              [,1]
## (Intercept)  9.305
## education   20.284
##
## Pr(>|z|)                    [,1]
## (Intercept) 1.337833e-20
## education   1.774739e-91
##
## R-squared:[1] 0.09358627
## Adjusted R-squared:[1] 0.09335881
```

```r
summary(model2b)
```

```
## Call:
## mylm(formula = wages ~ age, data = SLID)
##
## Residuals:
## Min: [1] -17.747
## 1Q:    25%
## -4.847
## Median: [1] -1.507
## 3Q:   75%
## 3.914
## Max: [1] 35.063
##
## Coefficients:
## Estimate                [,1]
## (Intercept) 6.89090
## age         0.23311
##
## Std. Error                 [,1]
## (Intercept) 0.374047
## age         0.009583
##
## z-value             [,1]
## (Intercept) 18.42
## age         24.33
##
## Pr(>|z|)                     [,1]
## (Intercept)  8.661939e-76
## age         1.058845e-130
##
## R-squared:[1] 0.1292891
## Adjusted R-squared:[1] 0.1290706
```

Part 4

```r
SLID$languageRELEVEL <- relevel(SLID$language, ref = "Other")
library(mylm)
model3 <- mylm(wages ~ sex + age + languageRELEVEL + I(education^2),
    data = SLID)
```

```r
model3b <- lm(wages ~ sex + age + languageRELEVEL + I(education^2), data = SLID)
print(model3)
```

```
## Call:
## mylm(formula = wages ~ sex + age + languageRELEVEL + I(education^2),
##     data = SLID)
##
## Coefficients:
##      (Intercept) sexMale     age languageRELEVELEnglish
## [1,]     -2.0101  3.4087 0.24862                0.13454
##      languageRELEVELFrench I(education^2)
## [1,]              0.059008       0.034815
```

```r
print(model3b)
```

```
##
## Call:
## lm(formula = wages ~ sex + age + languageRELEVEL + I(education^2),
##     data = SLID)
##
## Coefficients:
##            (Intercept)                  sexMale                      age
##               -2.01007                  3.40870                  0.24862
## languageRELEVELEnglish   languageRELEVELFrench          I(education^2)
##                0.13454                  0.05901                  0.03482
```

```r
library(mylm)
model4 <- mylm(wages ~ language + age + language * age, data = SLID)
model4b <- lm(wages ~ language + age + language * age, data = SLID)
print(model4)
```

```
## Call:
## mylm(formula = wages ~ language + age + language * age, data = SLID)
##
## Coefficients:
##      (Intercept) languageFrench languageOther     age languageFrench:age
## [1,]      6.5558         2.8606       0.84862 0.24485          -0.083928
##      languageOther:age
## [1,]         -0.037014
```

```r
print(model4b)
```

```
##
## Call:
## lm(formula = wages ~ language + age + language * age, data = SLID)
##
## Coefficients:
##        (Intercept)      languageFrench       languageOther
##            6.55579             2.86063             0.84862
##                age  languageFrench:age   languageOther:age
##            0.24485            -0.08393            -0.03701
```

```r
library(mylm)
model5 <- mylm(wages ~ education - 1, data = SLID)
model5b <- lm(wages ~ education - 1, data = SLID)
print(model5)
```

```
## Call:
## mylm(formula = wages ~ education - 1, data = SLID)
##
## Coefficients:
##      education
## [1,]    1.1467
```

```r
print(model5b)
```

```
##
## Call:
## lm(formula = wages ~ education - 1, data = SLID)
##
## Coefficients:
## education
##     1.147
```