TMA4315: Compulsory exercise 1 (title)

Group 0: Name1, Name2 (subtitle)

25.09.2018

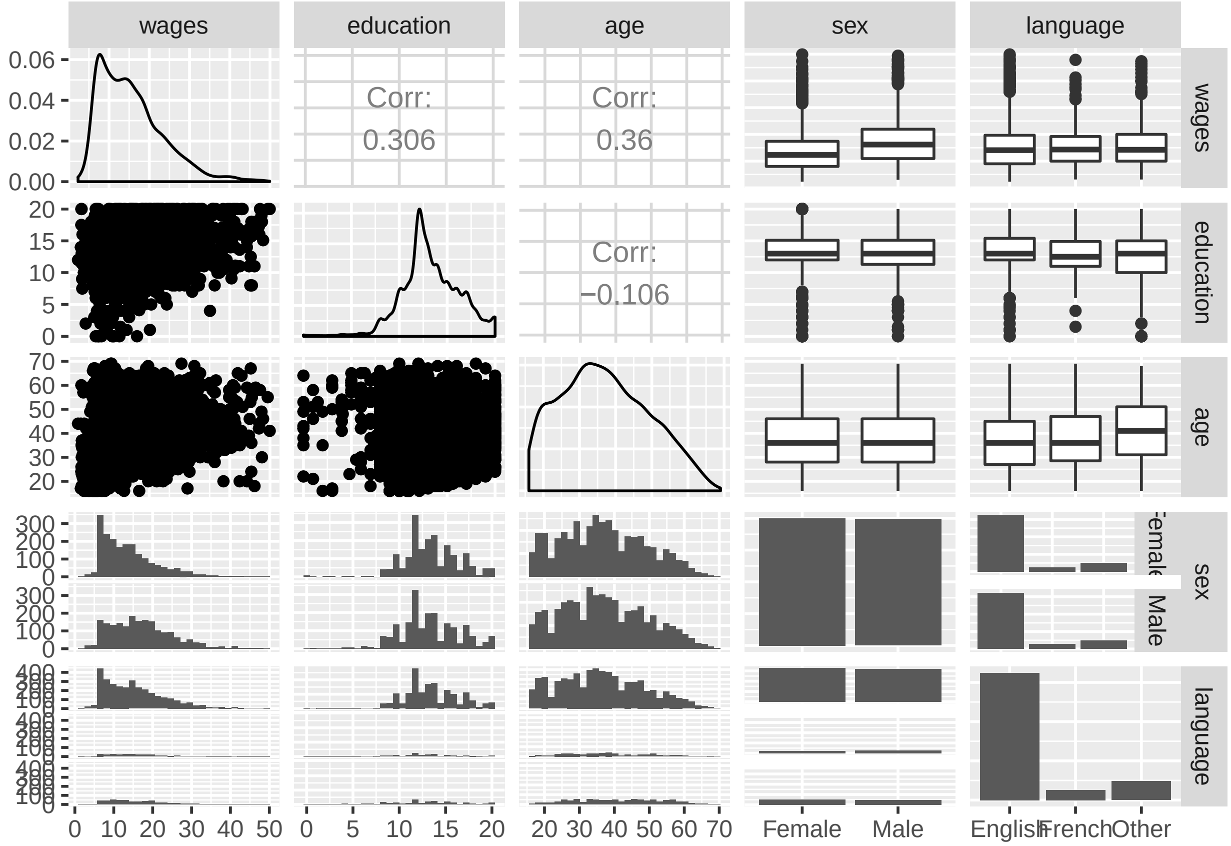
Introduction

In this compulsary exercise we are going to make an r package, called "mylm", with some of the same functionality as found in the already existing r package `lm`. We are going to test our implementations on the

Part 1

a)

The aim of part 1 is to get an overview of the dataset before making the linear regressions. The following diagram shows some diagnostic polts based on the data set from Canada in the car-library, containing information about `wages`, `education`, `age`, `sex` and `language`.



In the later exercises we are going to perform regression using `wages` as response. The properties that seem to be most correlated with the `wages` is `age` and `education`, with numerical correlation value of 0.36 and 0.306 respectively. On the scatter plots of `wages` versus `age` and `wages` versus `education`, it can be seen that among those with highest wages there are few with low age and low education.

On the other hand, the wage level does not seem to be related to the `language` of the observed people. It also shows little correlation with `sex`, but the average wage is slightly higher for the males than for the females.

If we want to perform a multiple linear regression analysis to predict wages based on some of the other variables we have to assume that relationship between wages and the varięles is in fact linear. Also, the residuals are assumed to be normally distributed and homoscedastic, and the explanatory covariates are assumed to be uncorrelated.

Based on the diagnostic plots, none of the explanatory variables seem to be highly correlated. However, for instance `education` and `age` correlates slightly, with coefficient of correlation of -0.106.

Part 2

a)

In order to estimate the regression coefficients, $\hat{\beta}$ we have used the matrix formula

$\hat{\beta} = (X^T X)^{-1} X^T Y$

The following printouts gives the estimated intercept and regression coefficient from the linear regression with `wages` as response and `education` as predictor using `mylm` and `lm` respectively.

```
## Coefficients:
## (Intercept): 4.971691
## education: 0.7923091
```

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

This shows that the two functions calculates equal values, indicating that our function works as it was ment to in this case.

b)

We have estimated the variance as

$\hat{\sigma}^2 = \frac{1}{n-p}(Y - X\hat{\beta})^T(Y - X\hat{\beta})$

and further found the covariance matrix as

$Cov = \hat{\sigma}^2 X^T X$

The standard deviation for the estimated coefficients are set to the corresponding diagonal element of the covariance matrix.

The z-statistic for the z-test has the value

$z = \frac{\hat{\beta}-\beta_0}{SD} = \frac{\hat{\beta}}{SD}$

where β_0 is the beta-value of the the null hypothesis, in this case $\beta_0 = 0$.

The p-value is calculated as $2 \times pnorm(-|z|)$, where "pnorm" is a function in r which returns the p-value corresponding to the value of its input statistic. We are interested in the p-value corresponding to the two-sided test, so therefore we are multiplying the value by two.

R^2 is calculated as

$R^2 = 1 - \frac{SSE}{SST}$

where SSE and SST is

$SST = \mathbf{Y}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y}$

$SSE = (\mathbf{Y} - \mathbf{X}\hat{\beta})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})$

These values are presented in a "summary-table" in a similar way as the summary corresponding to `lm`. The following shows the summary of the regression using `mylm` and `lm` respectively.

```
## R squared: 0.09358627
## Coefficients:
```

```
##             Estimates      SD  ztest    pvalue
## (Intercept)    4.9717 0.53429  9.305 1.338e-20
## education      0.7923 0.03906 20.284 1.775e-91
```

```
##
## 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
```

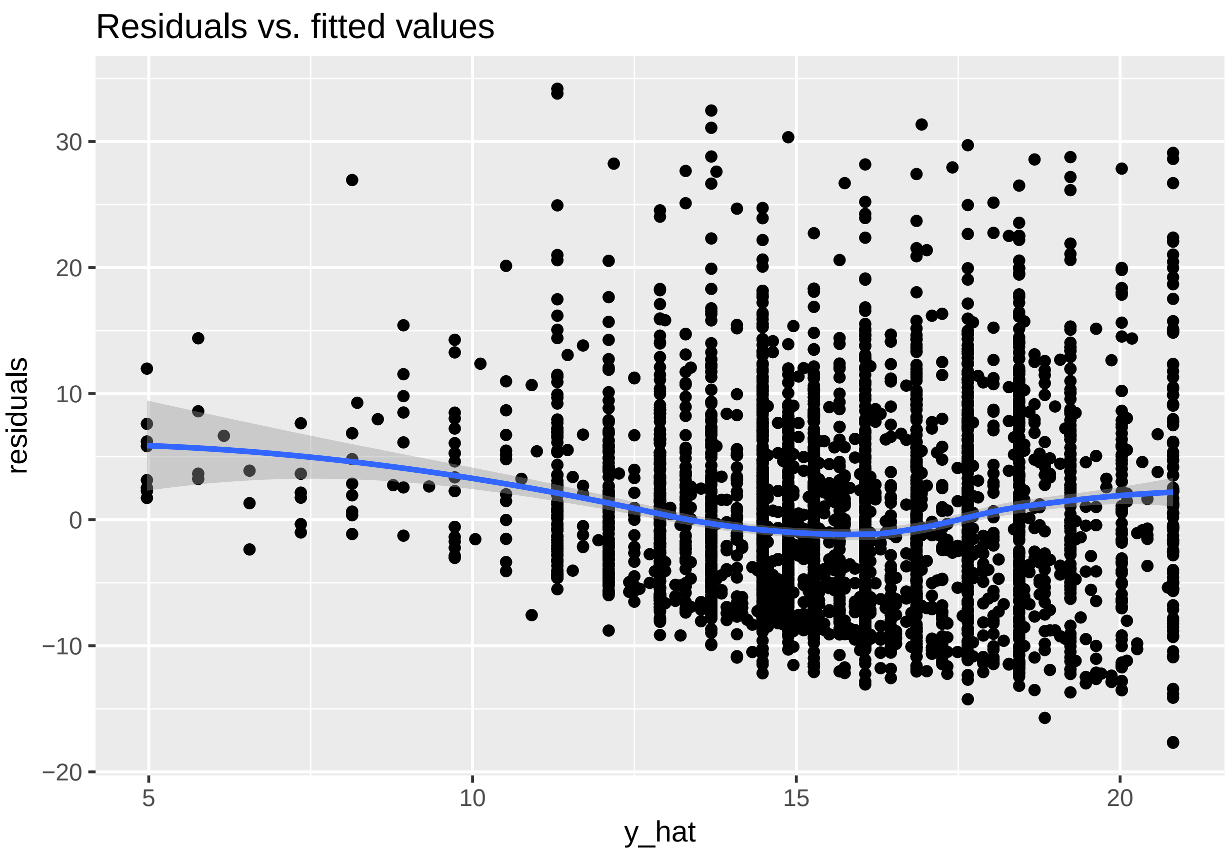
c)

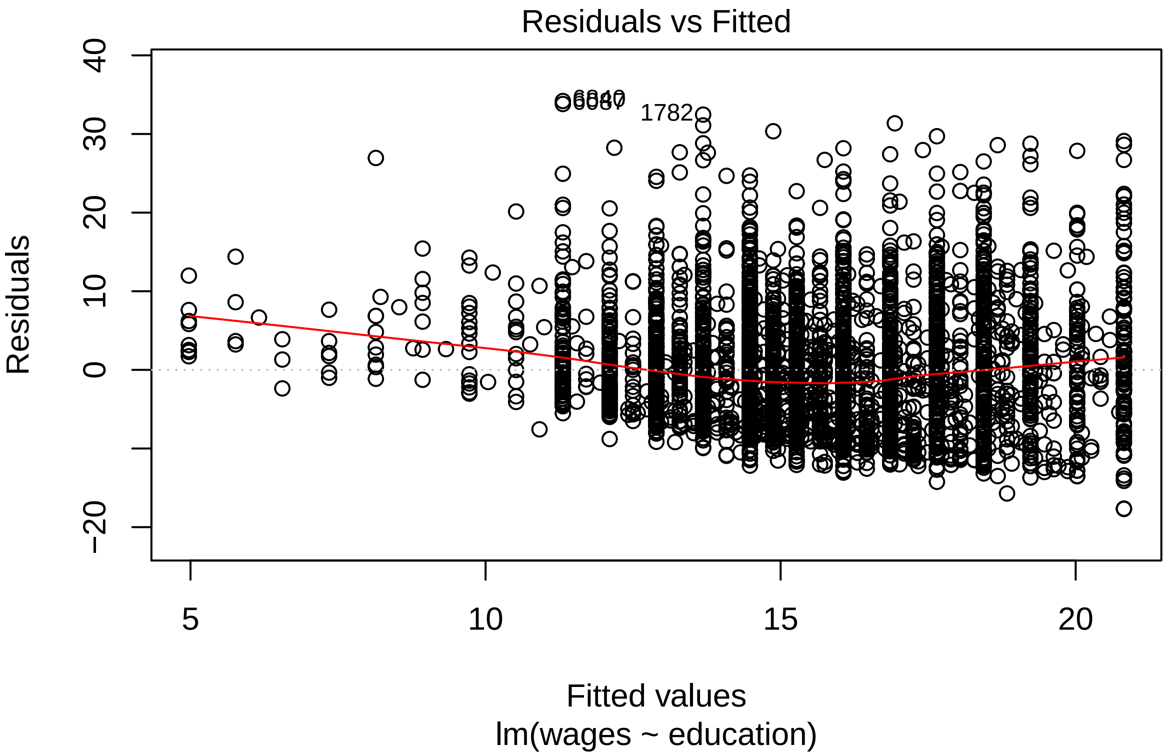
In the code chunk below, the code for the residual plot is shown along with the plot for `model1`. The raw residuals is the difference between the observed responses Y and the estimated response $\hat{Y}$. For this specific dataset, these values is respectively the observed and estimated values for `wages`. The formula for the the residuals is

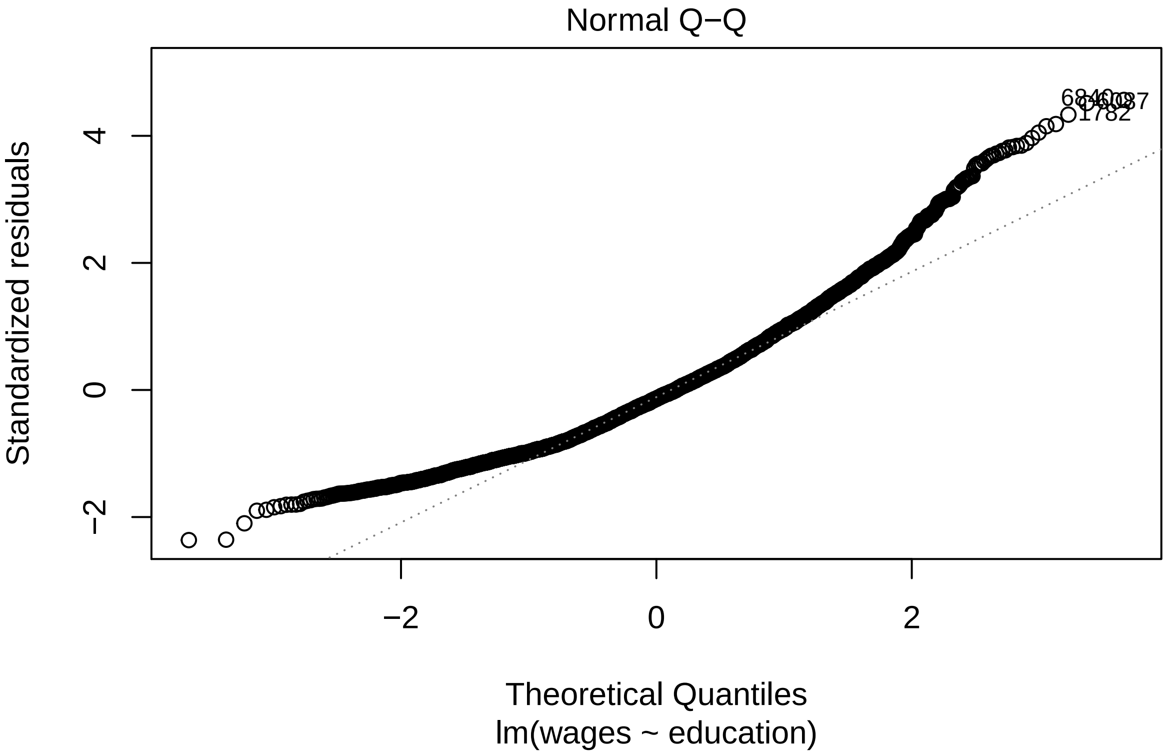
$$\hat{\epsilon} = Y - \hat{Y}$$

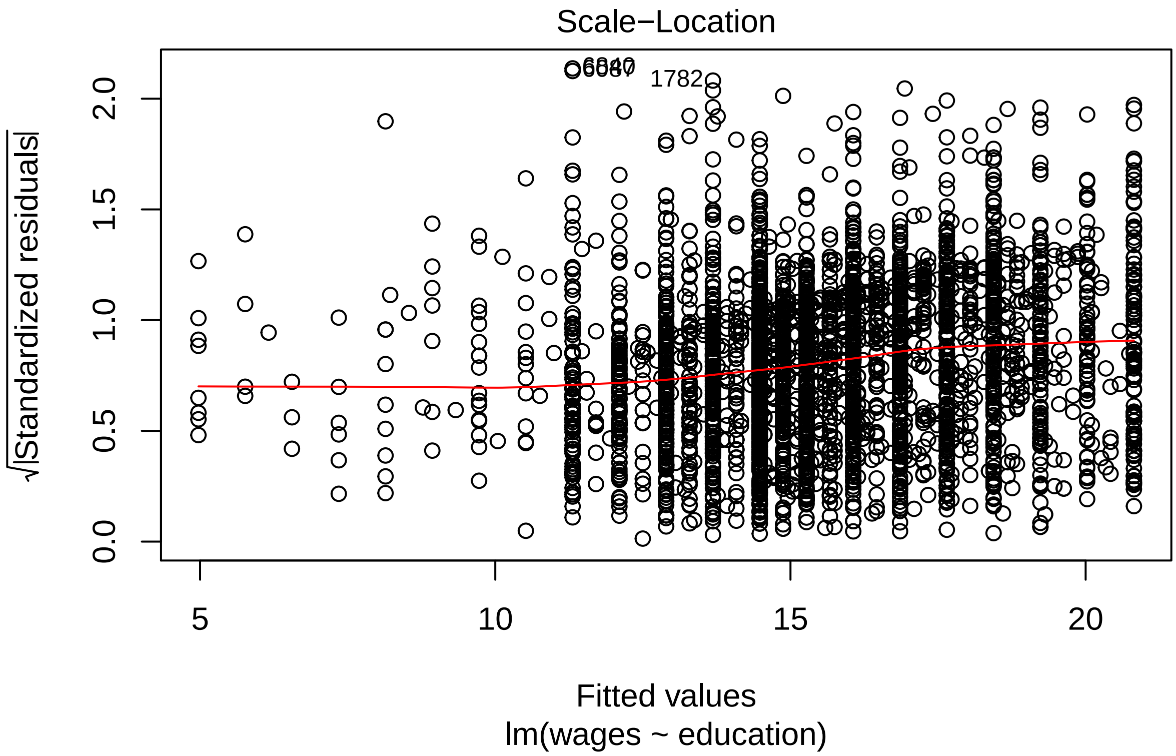
, where

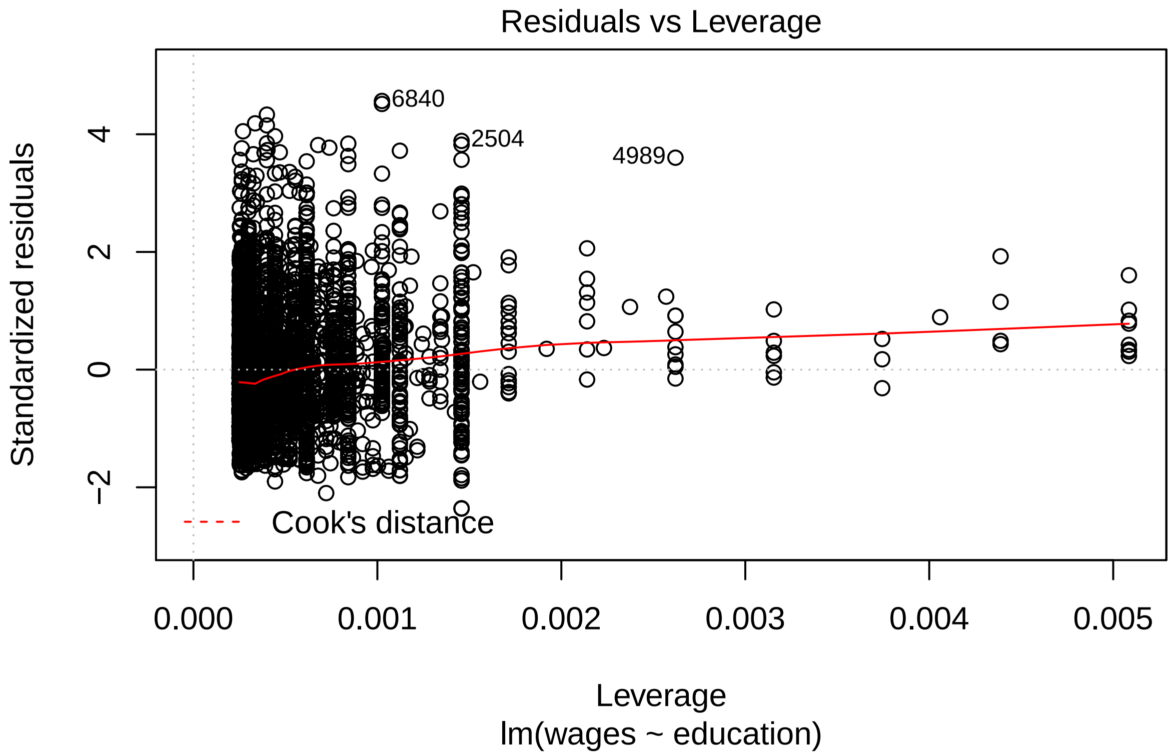
$$\hat{Y} = X\hat{\beta}$$


Residuals vs. fitted values
residuals
30
20
10
0
−10
−20
5
10
15
20
y_hat


Residuals vs Fitted
Residuals
40
30
20
10
0
−20
6840
6087
1782
5
10
15
20
Fitted values
lm(wages ~ education)


Normal Q−Q
Standardized residuals
4
2
0
−2
6840
4087
1782
−2
0
2
Theoretical Quantiles
lm(wages ~ education)


Scale−Location
6840
6067
1782
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5
10
15
20
Fitted values
lm(wages ~ education)



In the mylm-residual plot for the SLID dataset, it seems that the residuals are not homoscedastic, meaning that the variances seem to increase with higher values on the x-axis. However, there are fewer data points for lower values of the fitted values, so this tendency might be due to randomness. We also want the residuals to be symmetrically distributed around zero, which is not the case. This implies that the residuals might be correlated and we might want to try other regression models as the model assumptions is not preserved.

To ensure the function was coded correctly, the residual plots was compared to the plot from the plotting function for lm-objects.

d)

The residual sum of squares, SSE, for a model can be found by

$$SSE = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta),$$

and for the model at hand, $SSE = 223694.3$. For a model defined by a design matrix of dimension $n \times p$ the degrees of freedom, df, is given by $df = n - p$. In this case $df = 3987 - 2 = 3985$.

The total sum of squares, SST, is defined as

$$SST = \mathbf{Y}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y}.$$

For the model in this task $SST = 246790.5$. In order to test the significance of the regression, a χ^2-test can be used. The test statistic is then given by $\chi^2 = (\text{numerator degrees of freedom})F$, where F is the test statistic for a F-test and the numerator degrees of freedom is equal to $p - 1 = k$. Hence, the test statistic for the χ^2-test is

$$\chi^2 = \frac{(SST - SSE)}{\frac{SSE}{n-p}}.$$

The value of the test statistic is $\chi^2 = 411.4471$.To test for the significance of the regression one can find the p-value associated with this test statistic. The p-value in this case is $1.77 \cdot 10^{-91}$. With a p-value this small, it is reasonable to conclude that the regression is significant.

When testing one regression parameter the F-statistic is equal to the squared of the T-statistic. This is the case for single linear regression. The χ^2-distribution is the asymptotic distribution of the F-distribution, and the normal distribution is the asymptotic distribution of the T-distribution. Therefore the χ^2-statistic should be equal to the square of the z-statistic for a simple linear regression. The z-statistic for the regression coefficient in the model at hand is $z = 20.284$, and $z^2 = 411.4407$. The difference between the test statistics is probably due to rounding errors. The p-values for the two test statistics is also equal. The critical value for the χ^2 test is found to be 3.841 for significance level $\alpha = 0.05$ and the critical value for the z-test is 1.645 for $\alpha = 0.05$.

Part 3

a)

The matemathical formulas presented in part 2 works for multiple linear regression as well. The printout of the regression of wages with predictors education and wage using the mylm function and the lm function are

```
## Coefficients:
## (Intercept): -6.021653
## education: 0.9014644
## age: 0.2570898
```

```
##
## Call:
## lm(formula = wages ~ education + age, data = SLID)
##
## Coefficients:
## (Intercept)    education          age
##     -6.0217       0.9015       0.2571
```

b)

Thu summaries for this regression using lm and mylm including standard deviation and z-test are shown below.

```
## R squared: 0.2490697
## Coefficients:
```

```
##             Estimates       SD  ztest     pvalue
## (Intercept)   -6.0217 0.618924 -9.729  2.263e-22
## education      0.9015 0.035760 25.209 3.203e-140
## age            0.2571 0.008951 28.721 2.077e-181
```

```
##
## Call:
## lm(formula = wages ~ education + age, data = SLID)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -24.303  -4.495  -0.807   3.674  37.628
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.021653   0.618924  -9.729   <2e-16 ***
## education    0.901464   0.035760  25.209   <2e-16 ***
## age          0.257090   0.008951  28.721   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.82 on 3984 degrees of freedom
## Multiple R-squared:  0.2491, Adjusted R-squared:  0.2487
## F-statistic: 660.7 on 2 and 3984 DF,  p-value: < 2.2e-16
```

The estimated coefficient of predictor i can be interpreted as the value by which the wages will increase when predictor i increases by one and the other predictors are kept constant.

c)

Below, three different models are fitted using the `mylm`-package with `wages` as response. The first is with `age` as the only covariate, the second is with `language` as the only covariate, and the third is with both `age` and `language` in a multiple regression.

```
## R squared: 0.1292891
## Coefficients:
```

```
##             Estimates       SD ztest     pvalue
## (Intercept)    6.8909 0.374047 18.42  8.662e-76
## age            0.2331 0.009583 24.33 1.059e-130
```

```
## R squared: 0.09358627
## Coefficients:
```

```
##             Estimates      SD  ztest    pvalue
## (Intercept)    4.9717 0.53429  9.305 1.338e-20
## education      0.7923 0.03906 20.284 1.775e-91
```

```
## R squared: 0.2490697
## Coefficients:
```

```
##             Estimates       SD  ztest     pvalue
## (Intercept)   -6.0217 0.618924 -9.729  2.263e-22
## education      0.9015 0.035760 25.209 3.203e-140
## age            0.2571 0.008951 28.721 2.077e-181
```

We observe that the coefficients for `age` in the simple and multiple model is respectively 0.23 and 0.26. For education, the values are 0.79 and 0.90. The values differ slightly. This is because the matrix $\mathbf{X}^\mathbf{T}\mathbf{X}$ is not diagonal and the design matrix is not othogonal. This means that in our case, the estimated coefficients are not independent of each other. The estimated coefficient for `age` will affect the value for the estimated coefficient for `education`. This is also what was found in problem 1, where we found a small correlation between `age` and `education`.

Part 4

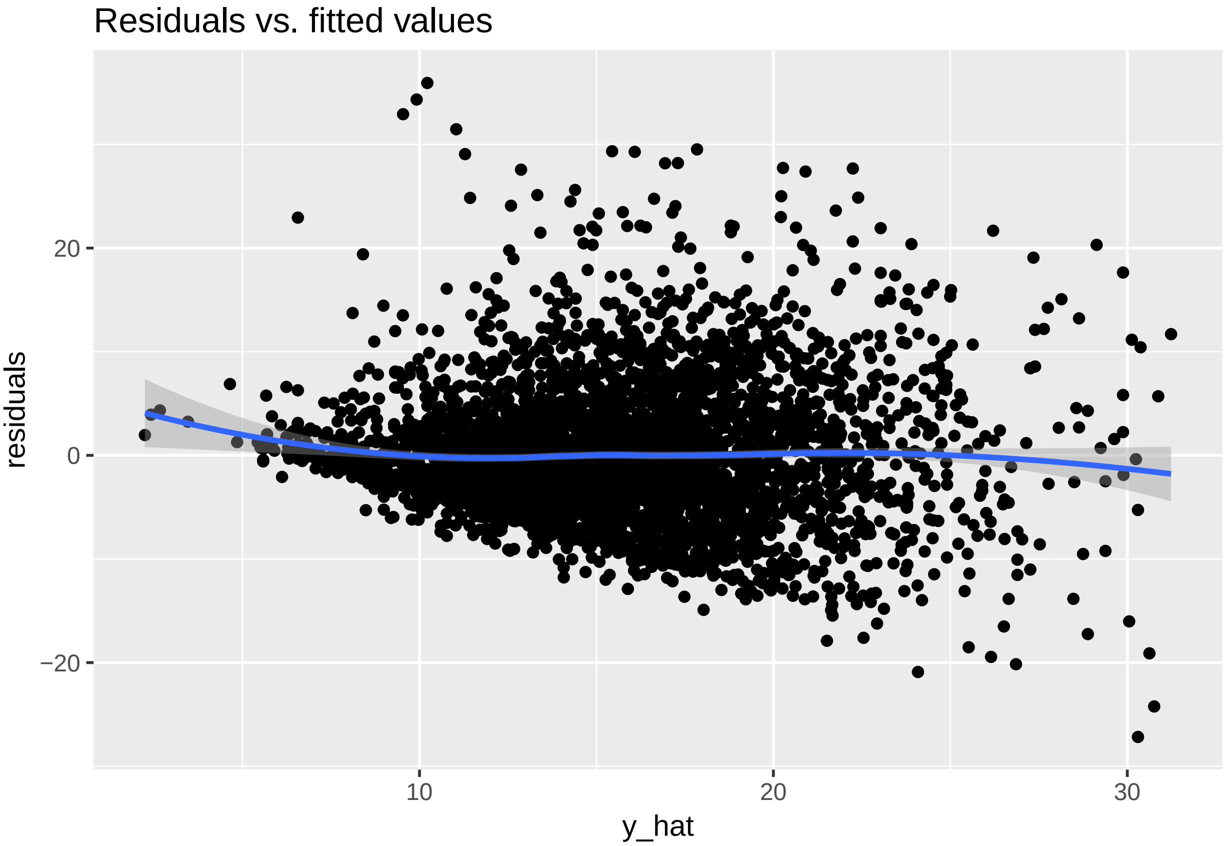
In this part we test our `mylm`-function on three different models. To ensure our package is correct, all values are compared to values generated by the `lm`-function.

Model 1

In this model, we estimate `wages` as a function of `sex`, `age`, `language` and `education^2`. The summary and residual plot is included below.

```
## R squared: 0.3022198
## Coefficients:
```

```
##                 Estimates       SD   ztest     pvalue
## (Intercept)      -1.87553 0.440345 -4.2592  2.051e-05
## sexMale           3.40870 0.208420 16.3550  4.008e-60
## age               0.24862 0.008663 28.7009 3.720e-181
## languageFrench   -0.07553 0.425136 -0.1777  8.590e-01
## languageOther    -0.13454 0.323153 -0.4163  6.772e-01
## I(education^2)    0.03482 0.001290 26.9907 1.902e-160
```



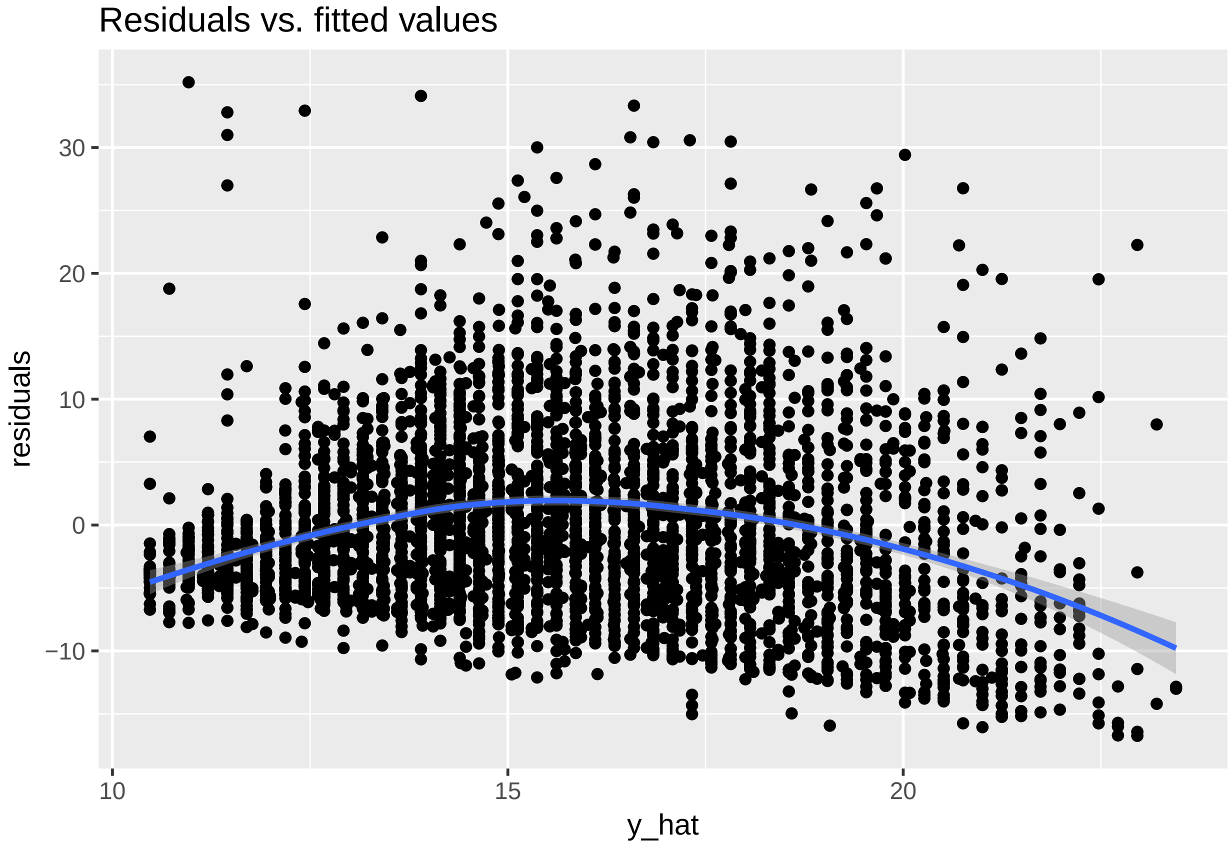
The R^2 value is the fraction of variance that is explained by the regression. For this regression, we see from the summary above that that $R^2 = 0.30$. From the p-values, we see that all covariates exececpt `language` are significant. A change to make this model better might be to exclude `language`.

Model 2

Here we estimate `wages` as a function of `language` and `age`, in addition to an interaction term between the covariates.

```
## R squared: 0.1311705
## Coefficients:
```

```
##                     Estimates      SD  ztest      pvalue
## (Intercept)           6.55579 0.41068 15.963  2.304e-57
## languageFrench        2.86063 1.59607  1.792  7.309e-02
## languageOther         0.84862 1.23518  0.687  4.921e-01
## age                   0.24485 0.01069 22.910 3.689e-116
## languageFrench:age  -0.08393 0.04046 -2.075  3.803e-02
## languageOther:age   -0.03701 0.02934 -1.262  2.071e-01
```



For this model, $R^2 = 0.13$ which is lower than for `model1`. Since the number of estimated covariates is the same for the two models, this means that a higher fraction of the variance is explained by factors we have not included in this model. Also, only intercept and `age` are significant on significance level 0.001. The interaction between `languageFrench` and `age`is significant on level 0.05. Since a model with interaction but no main effect is hard to interpret, we might want to consider a model without `language` but also include `education` which has turned out to be significant previously.

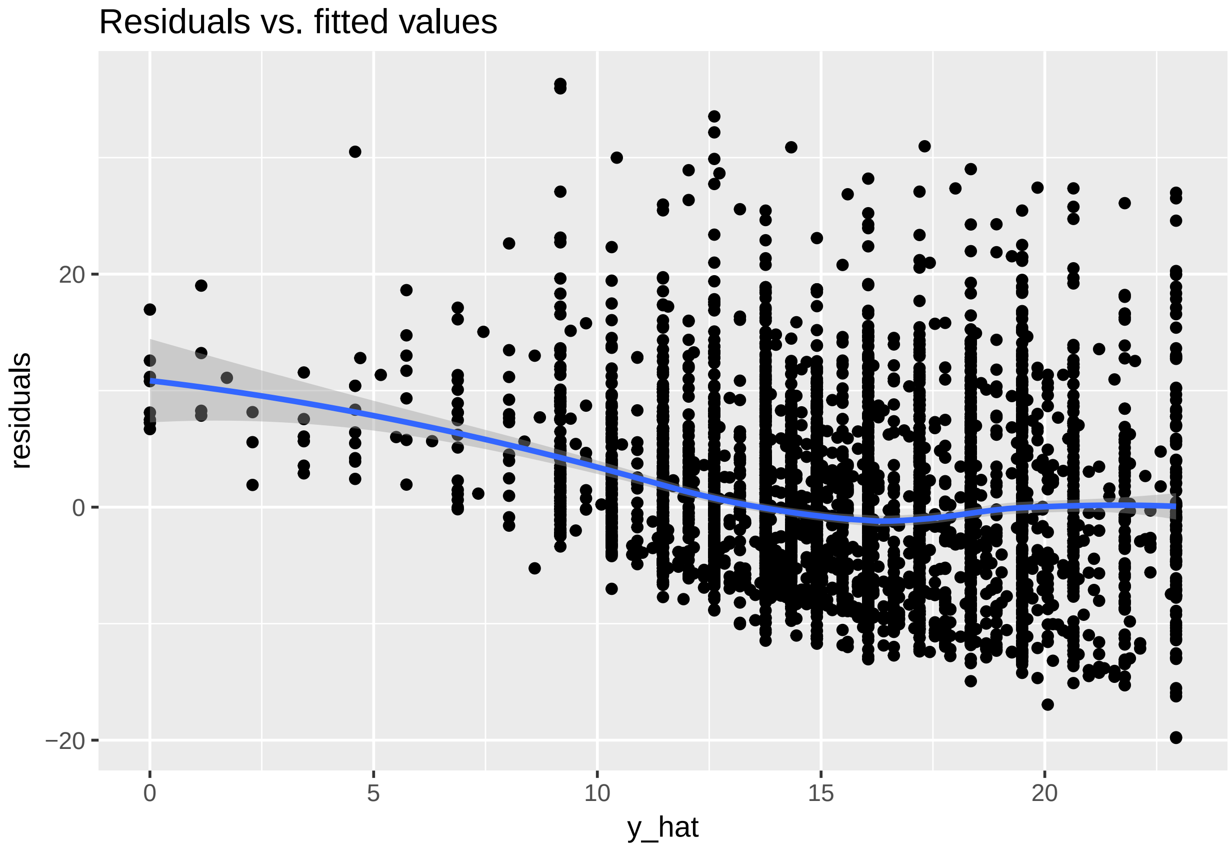
From the residual plot, we see that the residuals are not homoscedastic, indicating that we might want to try a different model.

Model 3

Now, we remove the intercept from our model and estimate `wages` as a function of education only.

```
## R squared: 0.07389171
## Coefficients:
```

```
##           Estimates       SD ztest pvalue
## education      1.147 0.008767 130.8      0
```



Now, the R^2-value do not make sense, but from the p.value in the summary, we see that the regression is significant. In the residual plot, we see that the variance is homoscedastic for values over 15 of estimated `wages`, but they are not quite symmetrically distributed around zero. For lower values, there seems to be an increasing variation, but we note that there are also fewer data points, so this might be due to randomness.

A small change that could make the model better is to include `age`, which have seemed to be significant in the earlier models.