TMA4315: Compulsory exercise 1

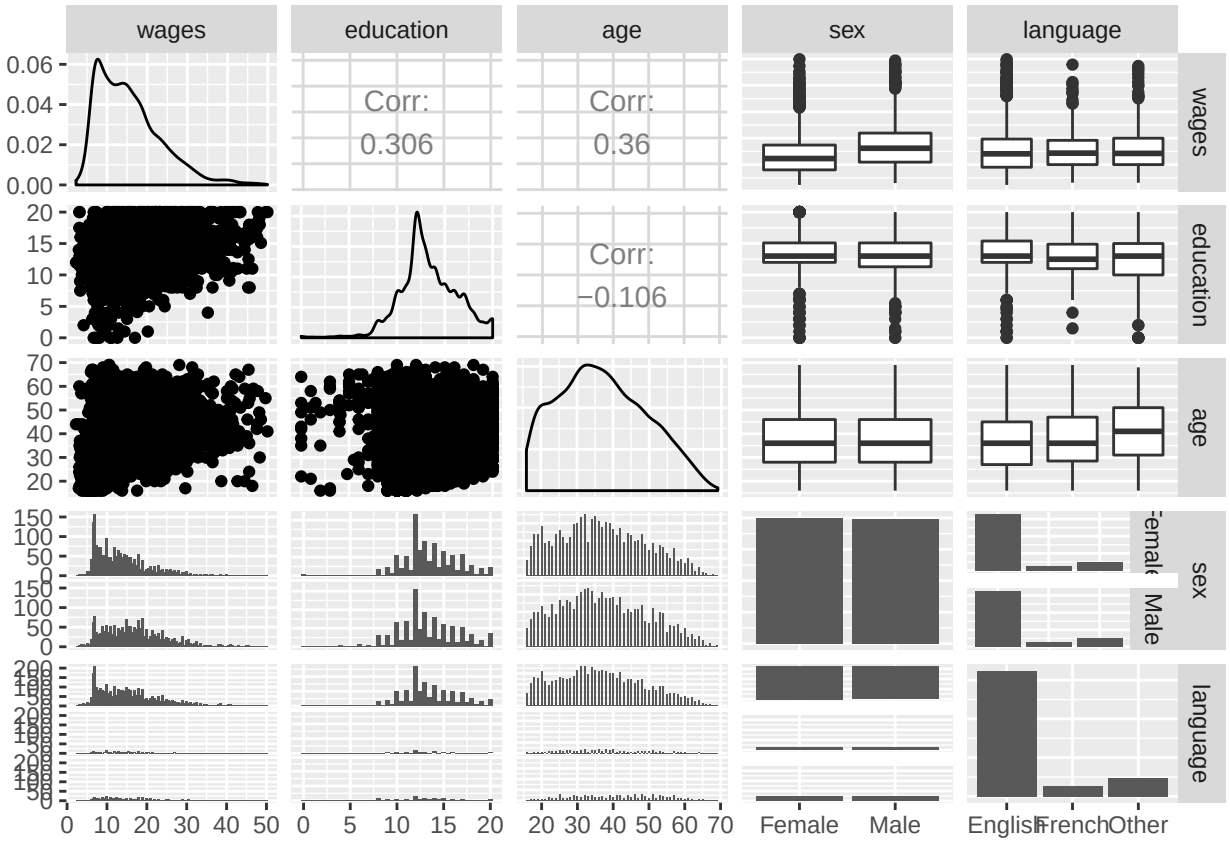
Group 4: Adrian Bruland og Mathias Opland

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. $Cov(\epsilon) = \sigma^2\mathbf{I}$, $Cov(\epsilon_i, \epsilon_j) = 0$. 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: $\varepsilon \sim \mathbf{N_n}(\mathbf{0}, \sigma^2\mathbf{I})$.

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.

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

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

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

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

```
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
```

```
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. 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. The value of the coefficient estimate for edjucation 0.79231 and the estimated standard error is 0.03906. Thus the regression propose that for each year of edjucation, the composite hourly wage rate will increase by approxomatly 0.79 dollars (The documentation of the SLID data set doesn't say anything about currency for the data, so based on the numbers we get and the fact that the publisher of the data set is canadian, we guess that the data is given in Canadian or US dollar).

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 afterwards. In our case, the H_0 hypotheses is that μ is zeros, and thus we get $Z = x/\sigma$:

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

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

Computing the p-values are done by the integral

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

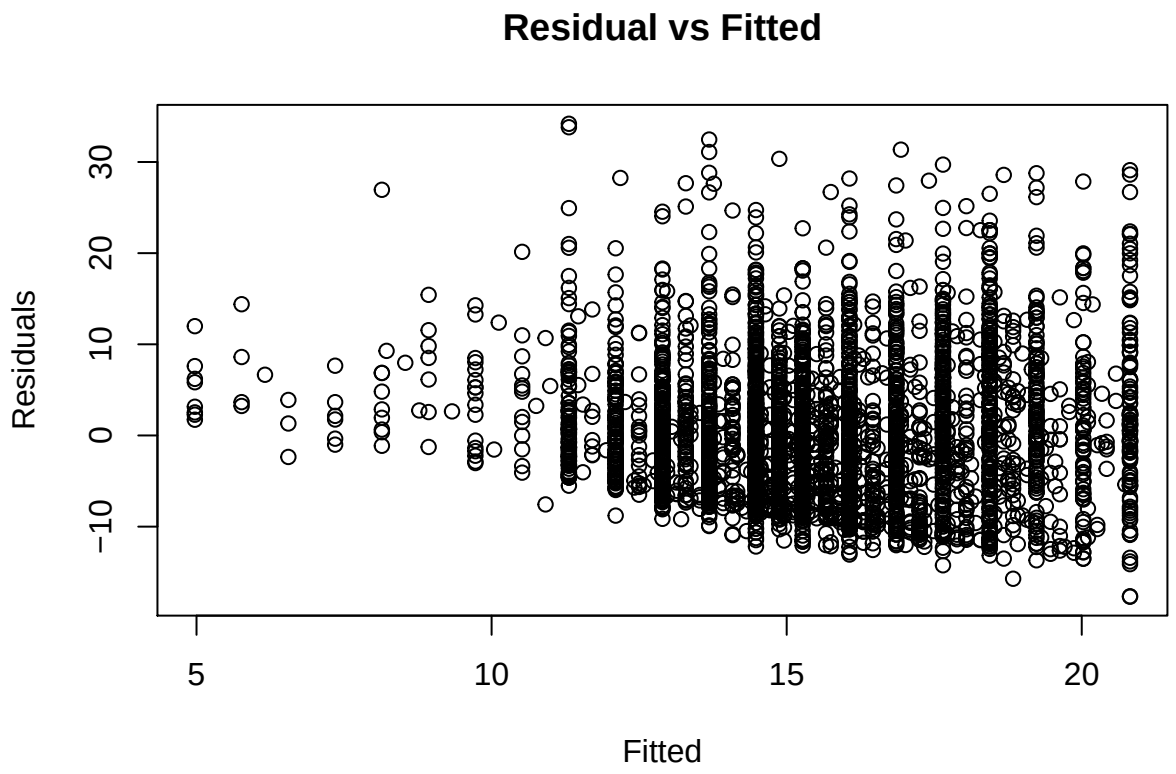
but as it is not analytic solvable for a unknown z, we use the built-in function pnorm, which gives the values:

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

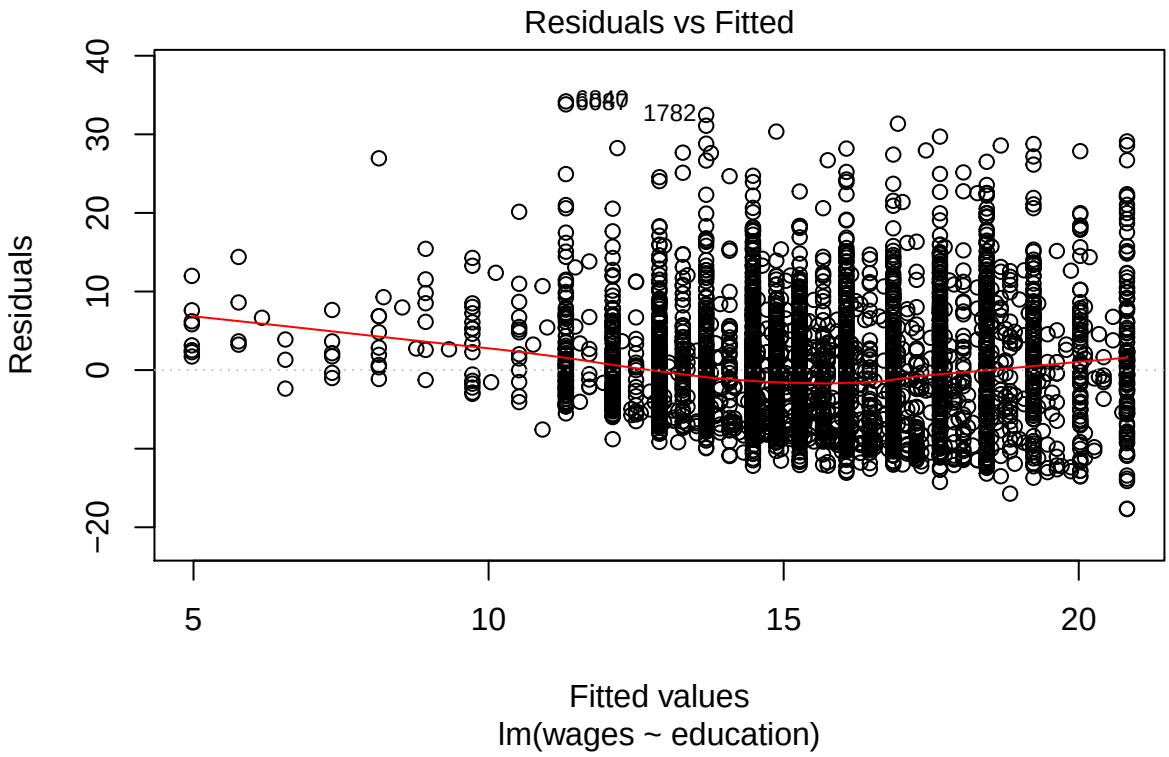
```
##                       [,1]
## (Intercept) 1.337833e-20
## education   1.774739e-91
```

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 $\sum_{i=1}^{n}(Y - \bar{(Y)})^2$:

```
## SST:
```

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

The χ^2-statistics is computed by $\chi^2_r = r \cdot \frac{SST-SSE}{SSE}$, 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-statistics p-value:

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

The F-distribution is asymptotic, and will close in on the χ^2 as the number of observations gets higher: $r \cdot F_{r,n} \xrightarrow{n\to\infty} \chi^2_r$. Here we can see that we get a answer close to the lm F-statistics, which is expected with a so high number of data. We are using r = 1 in this computation, since that is the number of regression coefficients.

- In a simple linear regression, the χ^2-test and the z-test wil describe the same thing, as the χ^2-test checks the significance of all the covariates together. In other words, the χ^2-test checks the significance of the regression. The z-test on the other hand checks the significance of each of the covariates. Thus they will in a simple linear regression check the significancy of the covariate, and ultimately test the same thing, but if there is more than one, they will differ. However they have different methods of finding the p-value, as the z-test uses the normal distribution, and χ^2-test uses the chi^2 distribution. This gives us in a 95% confidence interval the critical z-value:

```
## Critical z-value:

## [1] 1.959964

## Critical chi-squared value:

## [1] 3.841459
```

We are dividing 0.05 by 2 in the computation of the critical z-value as the distribution is two tailed, with each tail being 0.025 each. The χ^2 distribution on the other hand is not two tailed, and thus we are using 0.05.
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
library(mylm)
model2 <- mylm(wages ~ education + age, data = SLID)
model2b <- lm(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
```

Here we get a linear regression with wages as respones, and edjucation and age as covariates. The intercept are at -6.0217, which doesn't make any sense (paying for working), but the covariates at the intercept will be a zero year old with no edjucation. They will of course not work, and there is no data at those values. As mention earlier, we have to look at the range of data which the regression is based on. The coefficient for education suggests that the composit hourly wage increases with approximately 0.9 dollars per year of edjucation, and the coefficient for age suggests that for each year older one gets, the wages will increase by approximately 0.26 dollars per hour. **b)**

```r
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
```

The estimated standard error for the intercept is 0.618924, and for the regression coefficients edjucation and age it is respectivly 0.035760 and 0.008951. As we can see there is a much larger estimated standard error for the intercept, but the estimated value is also much larger. The z-value gives a better representation of the standard error with respect to the size of the estimate. Here we can see that the intercept has the lowest absolute z-value, which means that the standard error is largest with respect to the estimate. If we then look at the p-value for the estimates, we can see that both the coeffisients an the intercept is significant (with 5%-level), but intercept has the highest p-value. **c)**

```
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
```

```
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
```

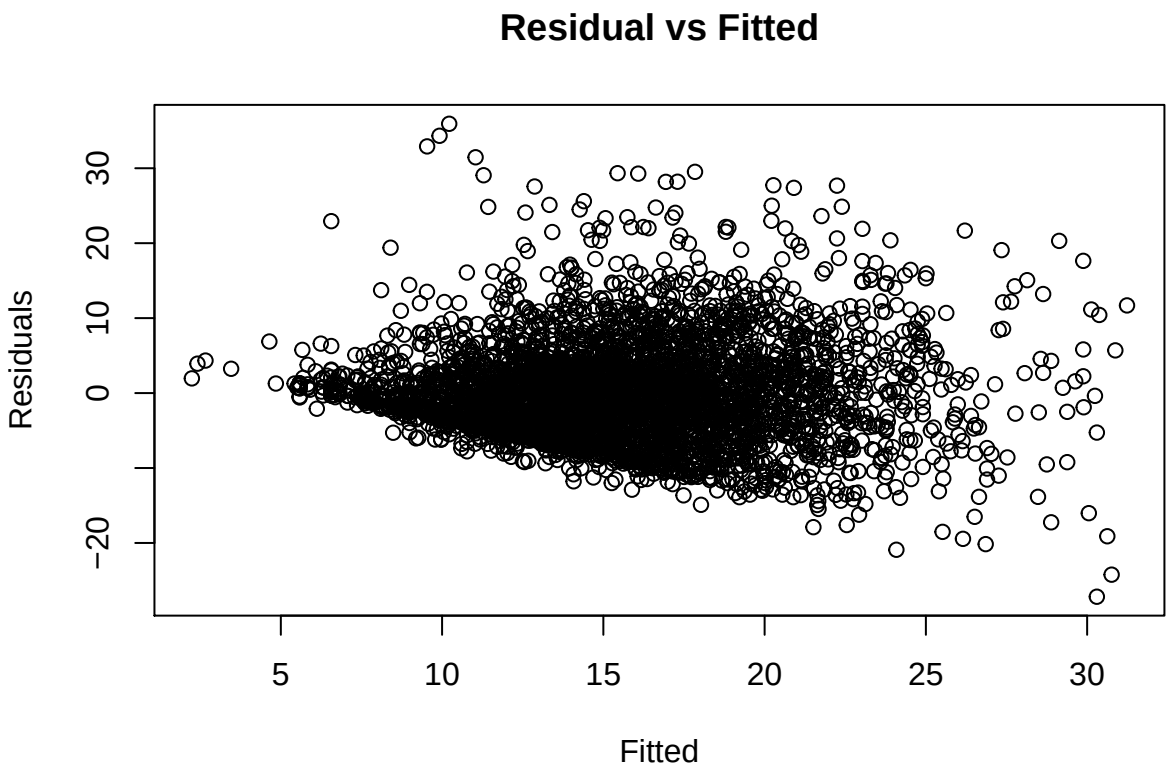
The two simple linear regressions both tries to explain the wages based on only one covariate. If the two covariates was uncorrelated, the multiple regression would be the same as two simple regression, but as shown in the plot in task 1, there is a negative correlation between age and edjucation, and thus both the regression coefficients are higher in the multiple linear regression. The simple regression with age as a covariate will try to explain the difference in wages only based on age, but as one gets older, there one is less likley to have much edjucation (due to the negative correlation), so it also takes that in to account. Thus true effect of the age will not be shown, and the regression coefficient is lower than in the multiple regression. The same argument can be made for the simple regression using edjucation as the only covariate, and thus both the simple regression has smaller estimates with higher standard error than the multiple regression. We can also see that the R^2 value for the multiple linear regression is more than the sum of the R^2 values for the simple regression, suggesting that the multiple linear regression explains more of the varianace in the wage than the simple linear regressions combined. # Part 4

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

```
## Call:
## mylm(formula = wages ~ sex + age + languageRELEVEL + I(education^2),
##     data = SLID)
##
## Residuals:
## Min: [1] -27.171
## 1Q:    25%
## -4.276
## Median: [1] -0.7631
## 3Q:   75%
## 3.218
```

```
## Max: [1] 35.929
##
## Coefficients:
## Estimate                              [,1]
## (Intercept)            -2.010072
## sexMale                 3.408700
## age                     0.248625
## languageRELEVELEnglish  0.134540
## languageRELEVELFrench   0.059008
## I(education^2)          0.034815
##
## Std. Error                             [,1]
## (Intercept)            0.534762
## sexMale                0.208420
## age                    0.008663
## languageRELEVELEnglish 0.323153
## languageRELEVELFrench  0.507173
## I(education^2)         0.001290
##
## z-value                           [,1]
## (Intercept)            -3.7588
## sexMale                16.3550
## age                    28.7009
## languageRELEVELEnglish  0.4163
## languageRELEVELFrench   0.1163
## I(education^2)         26.9907
##
## Pr(>|z|)                                [,1]
## (Intercept)             1.707210e-04
## sexMale                 4.008036e-60
## age                    3.719546e-181
## languageRELEVELEnglish  6.771638e-01
## languageRELEVELFrench   9.073773e-01
## I(education^2)         1.901746e-160
##
## R-squared:[1] 0.3022198
## Adjusted R-squared:[1] 0.3013434
```

```
plot(model3)
```



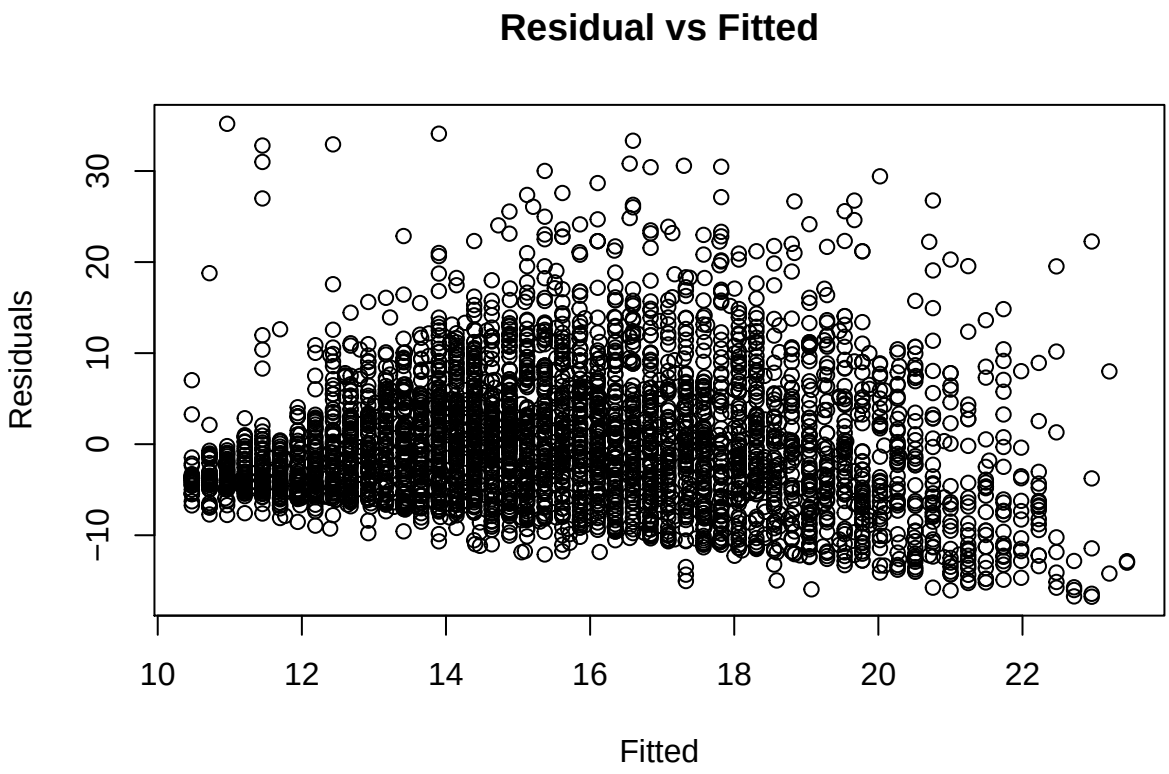
The mylm package produces the same values as the lm function up to 3 digits, so mylm passes the three test cases. In the first case, it's hard to know why we would like to use the square of the education variable, as it has no obvious real-world interpretation. From the matrix of plots from Part 1, we cannot see any visible quadratic relationship between education and wages, which also suggests that including education squared might have no practical use in a model. The easiest fix would be to simply use education with no exponent, as the linear education-wage relationship is a lot stronger in the models seen above. Assuming one still wants to include education squared, one could remove the language variable, as it has substantially smaller regression parameters than sex or age.

```
library(mylm)
model4 <- mylm(wages ~ languageRELEVEL + age + languageRELEVEL * age,
    data = SLID)
summary(model4)
```

```
## Call:
## mylm(formula = wages ~ languageRELEVEL + age + languageRELEVEL *
##     age, data = SLID)
##
## Residuals:
## Min: [1] -16.751
## 1Q:    25%
## -4.832
## Median: [1] -1.412
## 3Q:   75%
## 3.938
## Max: [1] 35.187
##
```

```
## Coefficients:
## Estimate                                [,1]
## (Intercept)                  7.404415
## languageRELEVELEnglish      -0.848621
## languageRELEVELFrench        2.012004
## age                          0.207838
## languageRELEVELEnglish:age   0.037014
## languageRELEVELFrench:age   -0.046914
##
## Std. Error                              [,1]
## (Intercept)                 1.16491
## languageRELEVELEnglish      1.23518
## languageRELEVELFrench       1.93282
## age                         0.02732
## languageRELEVELEnglish:age 0.02934
## languageRELEVELFrench:age   0.04763
##
## z-value                           [,1]
## (Intercept)                   6.3562
## languageRELEVELEnglish       -0.6870
## languageRELEVELFrench         1.0410
## age                           7.6063
## languageRELEVELEnglish:age    1.2615
## languageRELEVELFrench:age    -0.9849
##
## Pr(>|z|)                                [,1]
## (Intercept)                 2.067908e-10
## languageRELEVELEnglish      4.920565e-01
## languageRELEVELFrench       2.978910e-01
## age                         2.820625e-14
## languageRELEVELEnglish:age 2.071149e-01
## languageRELEVELFrench:age   3.246937e-01
##
## R-squared:[1] 0.1311705
## Adjusted R-squared:[1] 0.1300792
```

```
plot(model4)
```



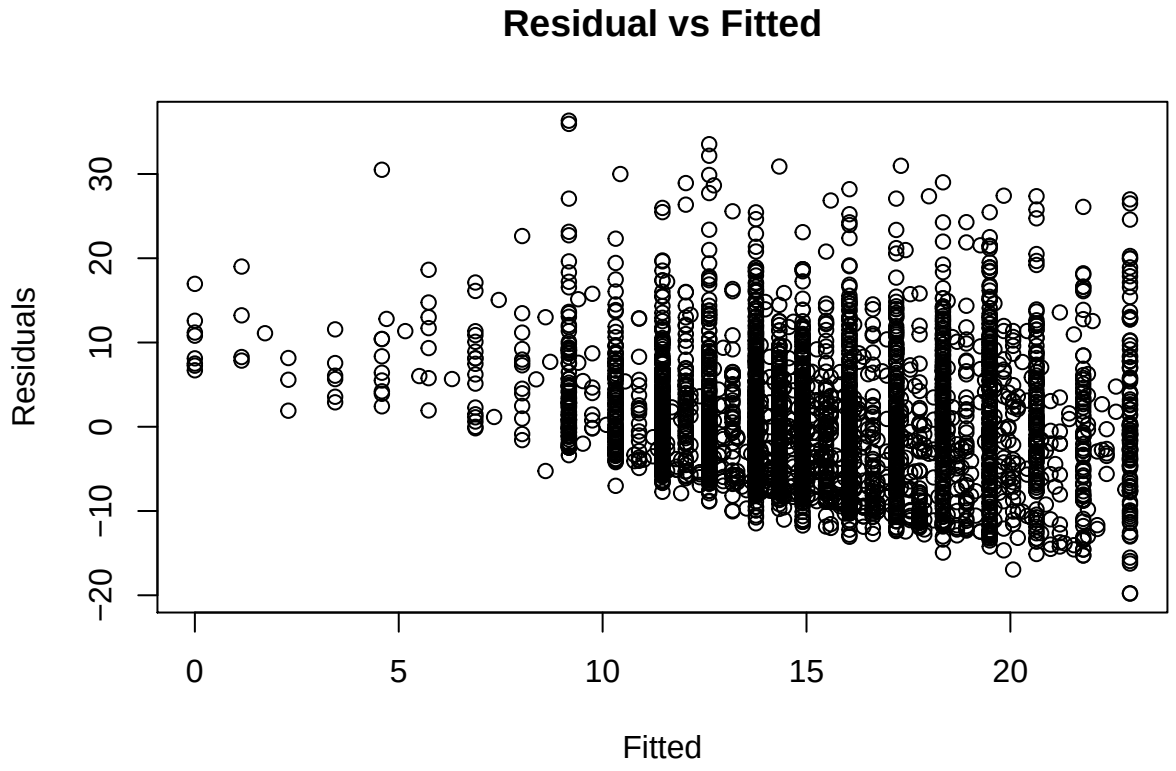
As for the second case, the interaction term adds the effect of how age affects wage differently for different language groups. We see that age has the greatest impact on wage amongst English speakers ($\beta_{age*English}$ = .245), followed by the speakers of other languages (.208) and finally French speakers (.245-.084=.161). So an increase in one year of age will, according to our model, increase wage by .245 dollars pr hour for English speakers. We see from the matrix of plots in Part 1 that sex seems to have a large impact in predicting wages than language, so one could add that to the model, plus the interaction, to better catch onto the factors that cause wage levels.

```
library(mylm)
model5 <- mylm(wages ~ education - 1, data = SLID)
summary(model5)
```

```
## Call:
## mylm(formula = wages ~ education - 1, data = SLID)
##
## Residuals:
## Min: [1] -19.804
## 1Q:    25%
## -5.342
## Median: [1] -0.6624
## 3Q:   75%
## 4.465
## Max: [1] 36.326
##
## Coefficients:
## Estimate            [,1]
## education 1.1467
```

```
##
## Std. Error              [,1]
## education 0.008767
##
## z-value           [,1]
## education 130.8
##
## Pr(>|z|)          [,1]
## education    0
##
## R-squared:[1] 0.07389171
## Adjusted R-squared:[1] 0.07389171
```

```
plot(model5)
```



On the third model, the intercept is taken away. This greatly restricts the sensitivity of the education parameter, as the model now only has one degree of freedom, so the model risks producing the same parameter even for highly different data sets that would produce different parameters had the intercept been included. The doesn't seem to be much point in making a model with no intercept. We see that the value of the parameter is very different from the one in 2a), where the intercept was included, so this data set exemplifies the change that a model instance can undergo just from removing the intercept.