TMA4315: Compulsory exercise 1 (title)

Group 0: Henrik Syversveen Lie, Mikal Stapnes

23.09.2018

Contents

To get a pdf file, make comments of the lines with the "html_document" information, and make the lines with the "pdf_document" information regular, and vice versa.

Part 1 (Explanatory analysis of the dataset)

Bold

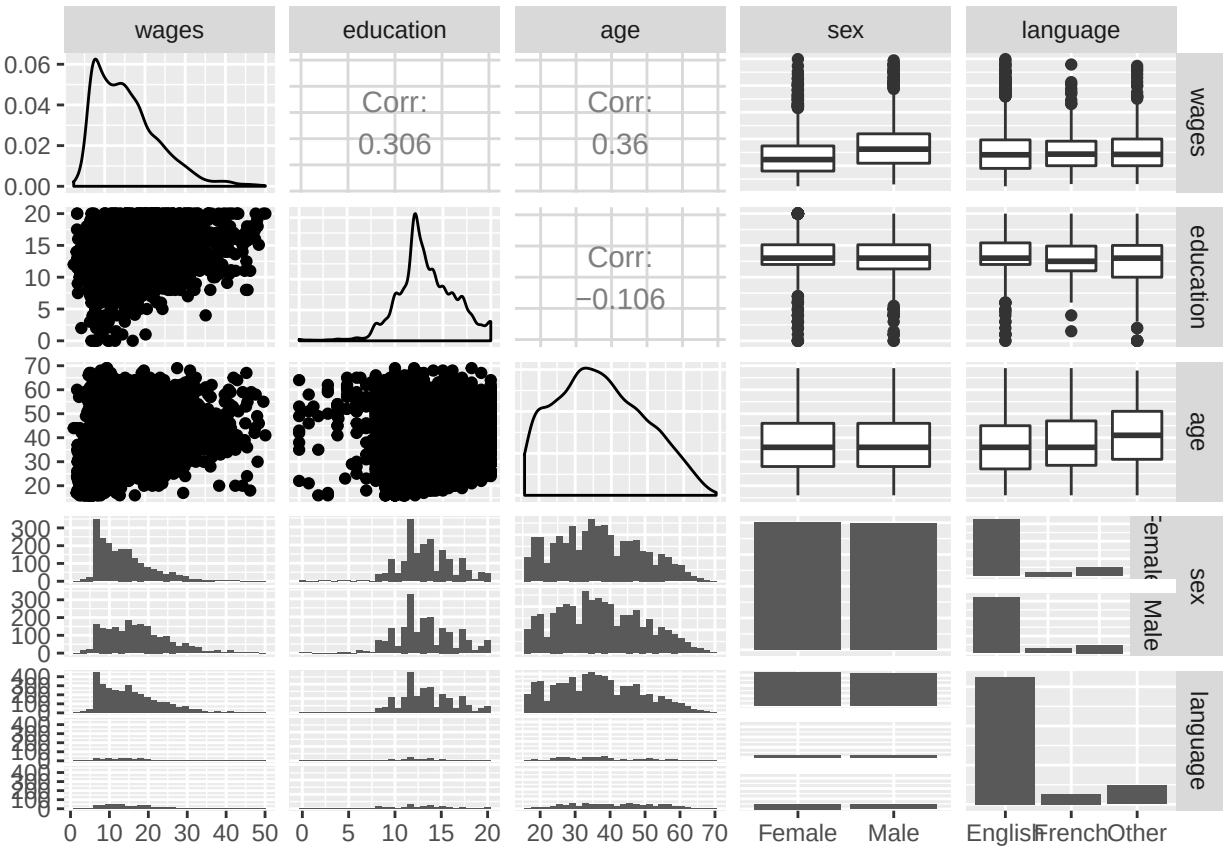
italic

a)

Task: * Draw a matrix of diagnostic plots of all the variables and comment briefly on the relationship between some of the variables. * Which assumptions do we need make about the data if we want to perform a multiple linear regression analysis?

Answer:

```
##      Package LibPath Version Priority Depends Imports LinkingTo Suggests
##      Enhances License License_is_FOSS License_restricts_use OS_type Archs
##      MD5sum NeedsCompilation Built
```



`Wage` has a positive correlation with the quantitative variables `education` and `age`. In addition we see that `sex`="Male" has a positive effect on `Wage`. Increasing `education`, `age` or switching `sex` to "Male" will, on average, result in an increase in `Wage` in our dataset.

If we want to perform a MLR analysis, several assumptions need to be made. We assume a linear relationship between the response `Wage` and the covariates

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i$$

The errors ϵ_i are assumed to be independent, identically distributed with mean 0 and constant variance σ^2. If we wish to construct confidence intervals and compute test statistics we must also assume normality in the errors $\epsilon_i \sim N(0, \sigma^2)$ SPØR METTE OM ANTAGELSENE KOMMER I RIKTIG REKKEFØLGE HER

Part 2: Simple lienar regression with the `mylm` package

a)

Tasks: * Fill in the missing parts of the `mylm` function ... * Fit a simple linear regression model to the data ...

Answer:

We develop our `mylm` to estimate the coefficients using the least squares estimator

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

. We confirm that this results in the same coefficient estimates as the included `lm` function

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

```
## Call:
## mylm(formula = wages ~ education, data = SLID)
##
## Coefficients:
##      (Intercept) education
## [1,]    4.971691 0.7923091
```

```r
model2 = lm(wages ~ education, data = SLID)
print(model2)
```

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

b)

Tasks: * Develop the `mylm` function further, so it can calculate the estimated covariance matrix of the parameters... * What are the estimates and the SE of the intercept and the regression coefficients for this model? * Fill in the missing parts in summary s.t. it gives a similare table of significance-tests.

Answer:

As $\hat{\beta} = (X^T X)^{-1} X^T Y$ is a linear transformation of Y and it is assumed that $Cov(Y) = \sigma^2 I$,

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

As we have also assumed that $Y \sim N(\beta X, \sigma^2)$, we get that

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)$$

We can then test the significance of the coefficients by computing the test statistics

$$z_j = \frac{(\beta_j - 0)}{(c_{jj} \sigma^2)}, \quad c_{jj} = (X^T X)^{-1}_{jj}$$

And the corresponding p-values

$$p_j = 2 \Pr(Z \geq |z_j|)$$

Note that we in `mylm` assume asymptotic results ($n - p$ very large) and use the approximation $\hat{\sigma}^2 = \sigma^2$. If we did not assume asymptotic results we would have to use the t-distribution.

```
## Summary of object
## Call:
## mylm(formula = wages ~ education, data = SLID)
##
## Residuals:
##         0%        25%        50%        75%       100%
## -17.687873  -5.822482  -1.039400   4.147527  34.189836
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.9717    0.53429   9.305    2e-16 ***
## education     0.7923    0.03906  20.284    2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual Standard Error: 7.5 on 3985 degrees of freedom
## Multiple R-squared: 0.09359    Adjusted R-squared: 0.093359
## Chisq-statisic: 411 on 1 degrees of freedom, p-value: 2e-16
```

In our model we get the coefficient values $\hat{\beta}_0 = 4.97$ and $\hat{\beta}_1 = 0.79$ with the corresponding estimated standard errors $\widehat{SE(\beta_0)} = 0.53$ and $\widehat{SE(\beta_1)} = 0.039$. This gives us test statistics $z_0 = 9.30$ and $z_1 = 20.28$, which are both highly significant under model assumptions.

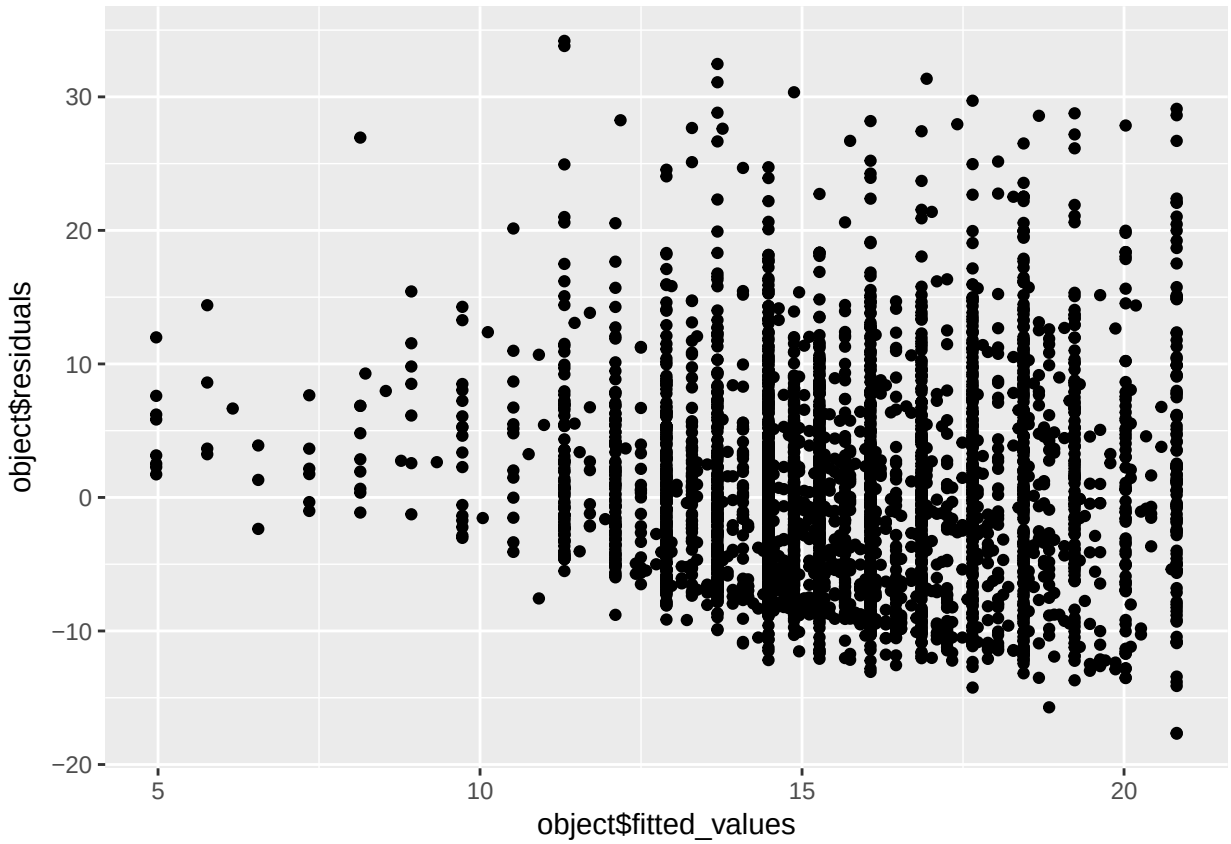
The coefficients can be interpreted as following: increasing `education` by one unit will increase the model response $\hat{Y}$ by 0.79 units. The `(Intercept)` is the mean of the responses, and is the model response in the case that `education = 0`. If our model assumptions are correct and our data is a sufficiently good representation of the remaining population we would also, on average, expect the same effects of the coefficients on the real response Y, `wages`.

c)

Tasks: * Implement a plot function for the `mylm` class that makes a scatter plot with fitted values on the x-axis and the residuals on the y-axis. * Comment on the plot

Answers: In `mylm` we implement the plot function with the fitted values on the x-axis and the residuals on the y-axis.

```
plot(model1)
```



From the plot we see that for lower values the residuels are mostly positive, whereas for higher values the are centered at 0. This means that there is some violation of the homoscedasticity of the errors. This is expected, as Y cannot be lower than 0 (negative wages does not make sense). We also observe that at higher fitted values, the density function is has a larger left tail (towards higher values) than left tail. This is also expected, as we again have a floor for the residuals (cannot have negative wages) but no roof. In summary, there is some violation of both the homoscedasticity and normality of the errors.

d)

The residual sum of squares `SSE` is the sum of squares $\sum^n (y_i - \hat{y}_i)^2$. The degrees of freedom for this model is the number of observations minus the number of fitted parameters, $n - p$. The total sum of squares `SST` is similarly defined as $\sum^n (y_i - \bar{y})^2$. Under H_0, i.e. the response is uncorrelated with all covariates,

$$\frac{(SST - SSE)/(p-1)}{SSE/(n-p)} \sim F_{p-1,n-p}$$

Again we assume asymptotic results, i.e. that $(n - p) \to \infty$, and get

$$\frac{(SST - SSE)}{SSE/(n-p)} \sim \chi^2_{p-1}$$

Which we can use to test the significance of the total regression.

```
cat("Chi-square test statistic: ", model1$F_statistic, "\n")
```

```
## Chi-square test statistic:  411.4471
```

```
cat("p-val: ", model1$F_p_val)
```

```
## p-val:  2e-16
```

In simple linear regression, $p = 2$ and the null hypothesis for the tests tested by the z-statistic and χ^2-statistic become equivalent. In addition, the z- and χ^2-statistics become equivalent measure. This comes simply as a result of the definition of a χ^2 variable, which is that if $Z_1, Z_2, \ldots, Z_k \sim N(0, 1)$, we have that the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2,$$

is distributed according to the χ^2 distribution with k degrees of freedom, denoted as $Q \sim \chi_k^2$. Seeing as the z-statistic is standard normal distributed, its square Z^2 will be χ^2 distributed with 1 degree or freedom, $Z^2 \sim \chi_1^2$. This means that the z- and χ^2 statistics will reject the null hypothesis at the same levels of significance.

We confirm this by listing the critical Z-values of some some quantiles along with square of the χ^2-values for the same quantiles. Note that the normal distributed quantiles are two-sided wheres the χ^2 quantiles are one-sided.

```
interval = c(0.7, 0.85, 0.9, 0.95)
interval2 = c(0.4, 0.6, 0.8, 0.9)
cat(abs(qnorm(interval)), "\n")
```

```
## 0.5244005 1.036433 1.281552 1.644854
```

```
cat(sqrt(qchisq(interval2, 1)))
```

```
## 0.5244005 0.8416212 1.281552 1.644854
```

e)

The coefficient of determination, R^2, is the square of the sample correlation between y and $\hat{Y}$. It is a measure of the proportion of variance explained by the regression.

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

We observe from the summary that in our model we get $R^2 = 0.09$, which indicates that we are not able to explain a large amount of the variance in the data using our model. This might indicate that the response we are trying to model, `wages`, has high variability, or that our model is too simple capture the true probability density function. In this example we suspect it to be a combination of both.

$ Part 3: Multiple linear regression

a)

So far we have implemented the our `mylm`to handle both simple and multivariate linear regression. For the details of implementation, see Part 1.

```
model3 = mylm(wages ~ education + age, data = SLID)
```

b)

```
summary(model3)
```

```
## Summary of object
## Call:
```

```
## mylm(formula = wages ~ education + age, data = SLID)
##
## Residuals:
##          0%         25%         50%         75%        100%
## -24.3030226  -4.4950673  -0.8070213   3.6736376  37.6279284
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.0217   0.618924  -9.729    2e-16 ***
## education     0.9015   0.035760  25.209    2e-16 ***
## age           0.2571   0.008951  28.721    2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual Standard Error: 6.8 on 3984 degrees of freedom
## Multiple R-squared: 0.2491    Adjusted R-squared: 0.24869
## Chisq-statisic: 661 on 2 degrees of freedom, p-value: 2e-16
```

In our model we get the coefficient values $\hat{\beta}_0 = -6.02$, $\hat{\beta}_{edu} = 0.902$, $\hat{\beta}_{age} = 0.257$ with the corresponding estimated standard errors $\widehat{SE}(\beta_0) = 0.619$, $\widehat{SE}(\beta_{edu}) = 0.036$ and $\widehat{SE}(\beta_{age}) = 0.257$. This gives us test statistics $z_0 = -9.73$, $z_1 = 25.2$ and $z_2 = 28.7$, which are all highly significant under model assumptions.

The coefficients can be interpreted as following: increasing only `education` by one unit will increase the model response $\hat{Y}$ by 0.902 units. Increasing only `age` by one unit will increase the model response $\hat{Y}$ by 0.257 units. The `(Intercept)` is the mean of the responses, and is the model response in the case that $'education' = 0$ and $'age' = 0$. If our model assumptions are correct and our data is a sufficiently good representation of the remaining population we would also, on average, expect the same effects of the coefficients on the real response Y, `wages`.

c)

```
model30 = mylm(wages ~ education, data = SLID)
model31 = mylm(wages ~ age, data = SLID)
summary(model30)
```

```
## Summary of object
## Call:
## mylm(formula = wages ~ education, data = SLID)
##
## Residuals:
##         0%        25%        50%        75%       100%
## -17.687873  -5.822482  -1.039400   4.147527  34.189836
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.9717    0.53429   9.305    2e-16 ***
## education     0.7923    0.03906  20.284    2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual Standard Error: 7.5 on 3985 degrees of freedom
## Multiple R-squared: 0.09359    Adjusted R-squared: 0.093359
## Chisq-statisic: 411 on 1 degrees of freedom, p-value: 2e-16
```

```
summary(model31)
```

```
## Summary of object
## Call:
## mylm(formula = wages ~ age, data = SLID)
##
## Residuals:
##         0%        25%        50%        75%       100%
## -17.747375  -4.847109  -1.506569   3.913970  35.063157
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.8909   0.374047   18.42    2e-16 ***
## age           0.2331   0.009583   24.33    2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual Standard Error: 7.3 on 3985 degrees of freedom
## Multiple R-squared: 0.1293    Adjusted R-squared: 0.12907
## Chisq-statisic: 592 on 1 degrees of freedom, p-value: 2e-16
```

ANSWER THIS: Why and when does the parameter estimates found (using two simple and one multiple) differ?

The parameter estimates for the simple and multiple regression differ if the covariates are correlated. Then, in the case of the simple regression, the single explanatory variable will be able to explain some of the variance that would otherwise be explained by the additional explanatory variable.

We confirm this by observing that `education` and `age` has a weak positive correlation.

```
model3$corr_coeff
```

```
##             (Intercept)  education        age
## (Intercept)   1.0000000 -0.8276025 -0.6184303
## education    -0.8276025  1.0000000  0.1062790
## age          -0.6184303  0.1062790  1.0000000
```

Part 4: Testing the `mylm`package

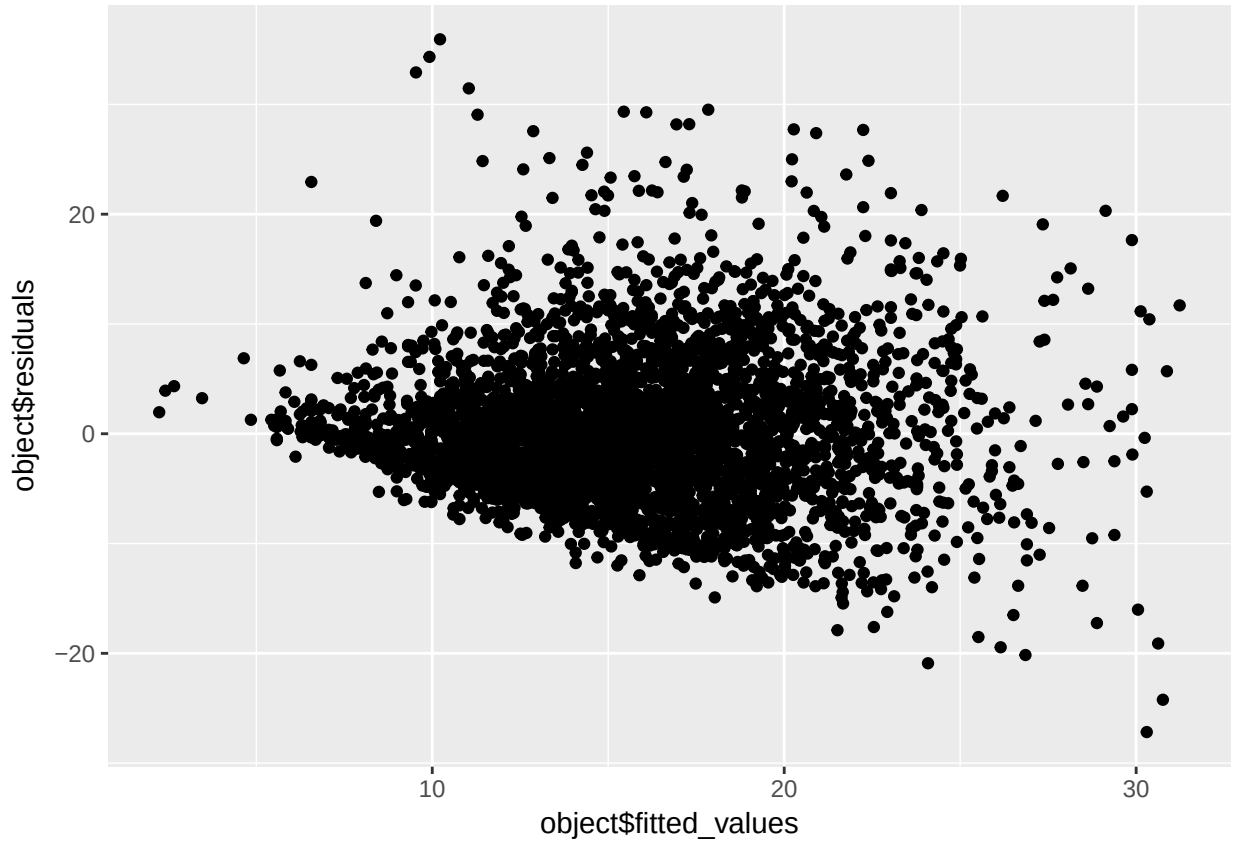
Tasks: * Write a few sentences about the interpretation and significance of the parameters, and mention one small change that could make the model better.

```
model40 = mylm(wages ~ sex + age + language + I(education^2), data = SLID)
summary(model40)
```

```
## Summary of object
## Call:
## mylm(formula = wages ~ sex + age + language + I(education^2),
##     data = SLID)
##
## Residuals:
##          0%         25%         50%         75%        100%
## -27.1711960  -4.2761844  -0.7631001   3.2176183  35.9289483
##
## Coefficients:
##                Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    -1.87553   0.440345 -4.2592 2.051e-05 ***
## sexMale         3.40870   0.208420 16.3550 2.000e-16 ***
## age             0.24862   0.008663 28.7009 2.000e-16 ***
```

```
## 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 2.000e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual Standard Error: 6.6 on 3981 degrees of freedom
## Multiple R-squared: 0.3022    Adjusted R-squared: 0.30134
## Chisq-statisic: 345 on 5 degrees of freedom, p-value: 2e-16
```

```
plot(model40)
```



The difference between this model and earlier MLR is that we are now a squared term, $'education'^2$. Our linear model assumption is now,

$$Y_i = \beta_0 + \beta_{sex}x_{i,sex} + \beta_{age}x_{i,age} + \beta_{lan}x_{i,lan} + \beta_{edu^2}x_{i,edu}^2 + \epsilon_i$$

The coefficient $\hat{\beta}_{edu^2}$ can be interpreted as following: increasing the square of `education` by one unit will increase the model response, $\hat{Y}$, by 0.035 units. All remaining coefficients can be interpreted as in Part 3b. We notice that all variables attain a high level of significance with the exception of language. Thus language should be removed from the model, as even under model assumptions it is probable that it is uncorrelated with the response.

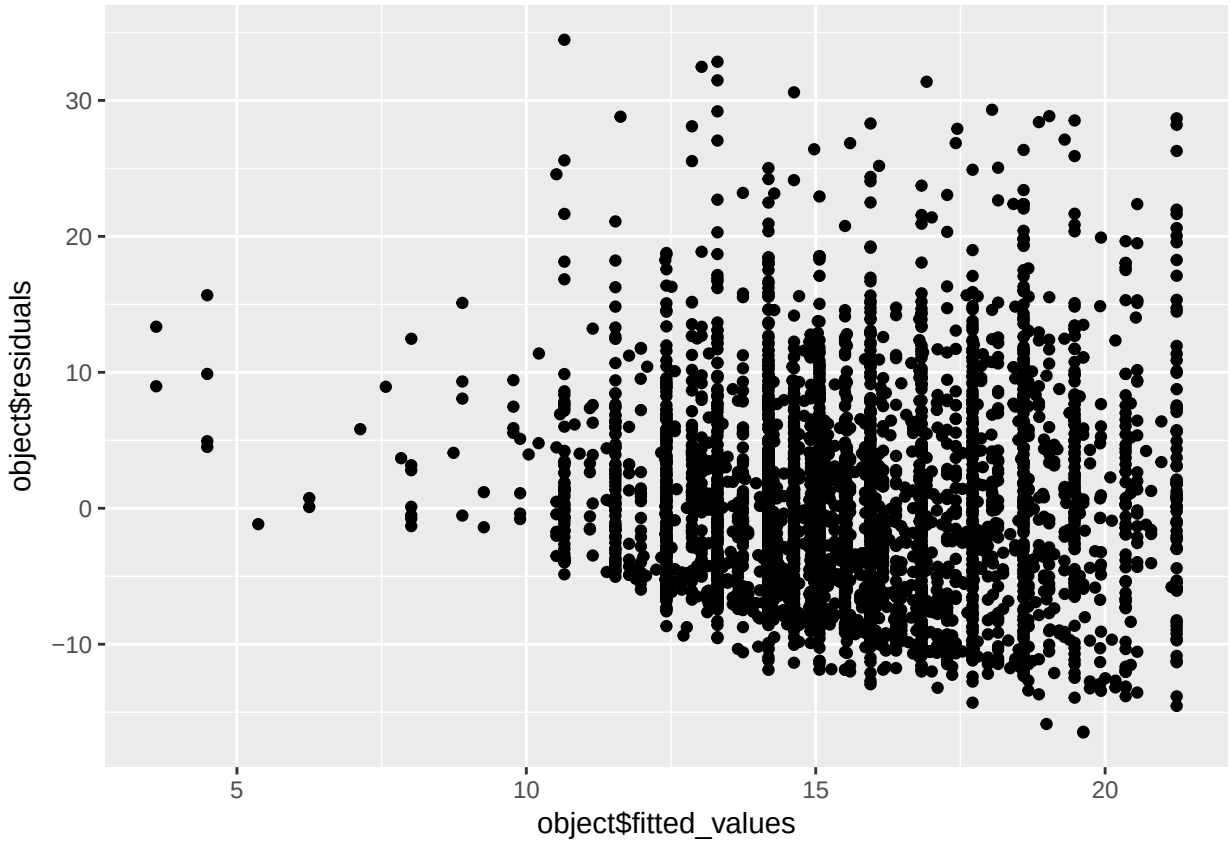
The error plot shows a lower bound of the residuals in the left corner. Again the explanation is that we cannot have negative `wages`. Beyond this, we observe a high level of homoscedasticity and normality in our errors, which indicates that our model assumptions are not unreasonable.

```
model41 = mylm(wages ~ language + education + language * education, data = SLID)
summary(model41)
```

```
## Summary of object
```

```
## Call:
## mylm(formula = wages ~ language + education + language * education,
##     data = SLID)
##
## Residuals:
##          0%         25%         50%         75%        100%
## -16.4938410  -5.6930927  -0.9909205   4.1249151  34.4622671
##
## Coefficients:
##                            Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                  3.6050    0.64087   5.625 1.854e-08 ***
## languageFrench               4.2548    2.05958   2.066 3.884e-02   *
## languageOther                4.4063    1.27818   3.447 5.661e-04 ***
## education                    0.8816    0.04653  18.946 2.000e-16 ***
## languageFrench:education    -0.2934    0.15373  -1.908 5.633e-02   .
## languageOther:education     -0.2544    0.09580  -2.655 7.928e-03  **
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual Standard Error: 7.5 on 3981 degrees of freedom
## Multiple R-squared: 0.09798    Adjusted R-squared: 0.096844
## Chisq-statisic: 86 on 5 degrees of freedom, p-value: 2e-16
```

```
plot(model41)
```



Now we have included an interaction term. Our linear model assumption is now,

$$Y_i = \beta_0 + \beta_{edu} x_{i,edu} + \beta_{lan} x_{i,lan} + \beta_{edu\&lan} x_{i,edu} x_{i,lan} + \epsilon_i$$

```r
model32 = mylm(wages ~ education - 1, data = SLID)
summary(model32)
```

```
## Summary of object
## Call:
## mylm(formula = wages ~ education - 1, data = SLID)
##
## Residuals:
##          0%         25%         50%         75%        100%
## -19.8039419  -5.3420851  -0.6623925   4.4646348  36.3264232
##
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## education   1.1467   0.008767  130.79    2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual Standard Error: 7.6 on 3986 degrees of freedom
## Multiple R-squared: 0.1970    Adjusted R-squared: 0.073892
## Chisq-statisic: Inf on 0 degrees of freedom, p-value: NaN
```

```r
plot(model32)
```