TMA4315: Compulsory exercise 1 Linear models for Gaussian data

Group 7: Henrik Syversveen Lie, Mikal Stapnes, Oliver Byhring

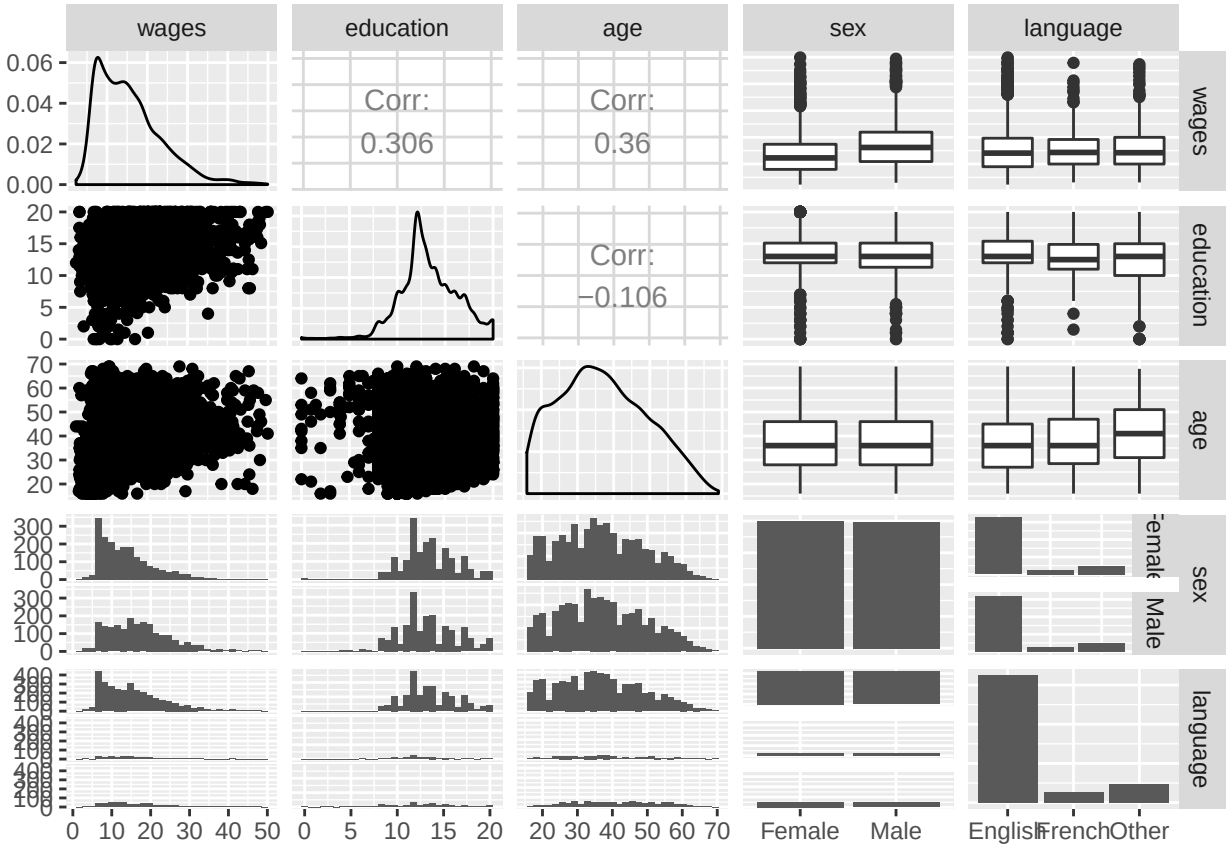
28.09.2018

Contents

Part 1: Explanatory analysis of the dataset

The following is a matrix of diagnostic plots of all the variables in the dataset from the `car` library.

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



From the diagnostic plots, we see that `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 $\epsilon_i \sim N(0, \sigma^2)$.

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

a)

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.

```
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
```

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

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

We also note that in the case of MLR, the maximum likelihood estimator is the same as the least squares estimator.

b)

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

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

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

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

We can then conduct a hypothesis test with hypotheses

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

We test the hypothesis using the test statistics

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

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), which gives

$$z_j \sim N(0, 1).$$

If we did not assume asymptotic results we would have that

$$z_j \sim T_{n-p}$$

```
## 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
```

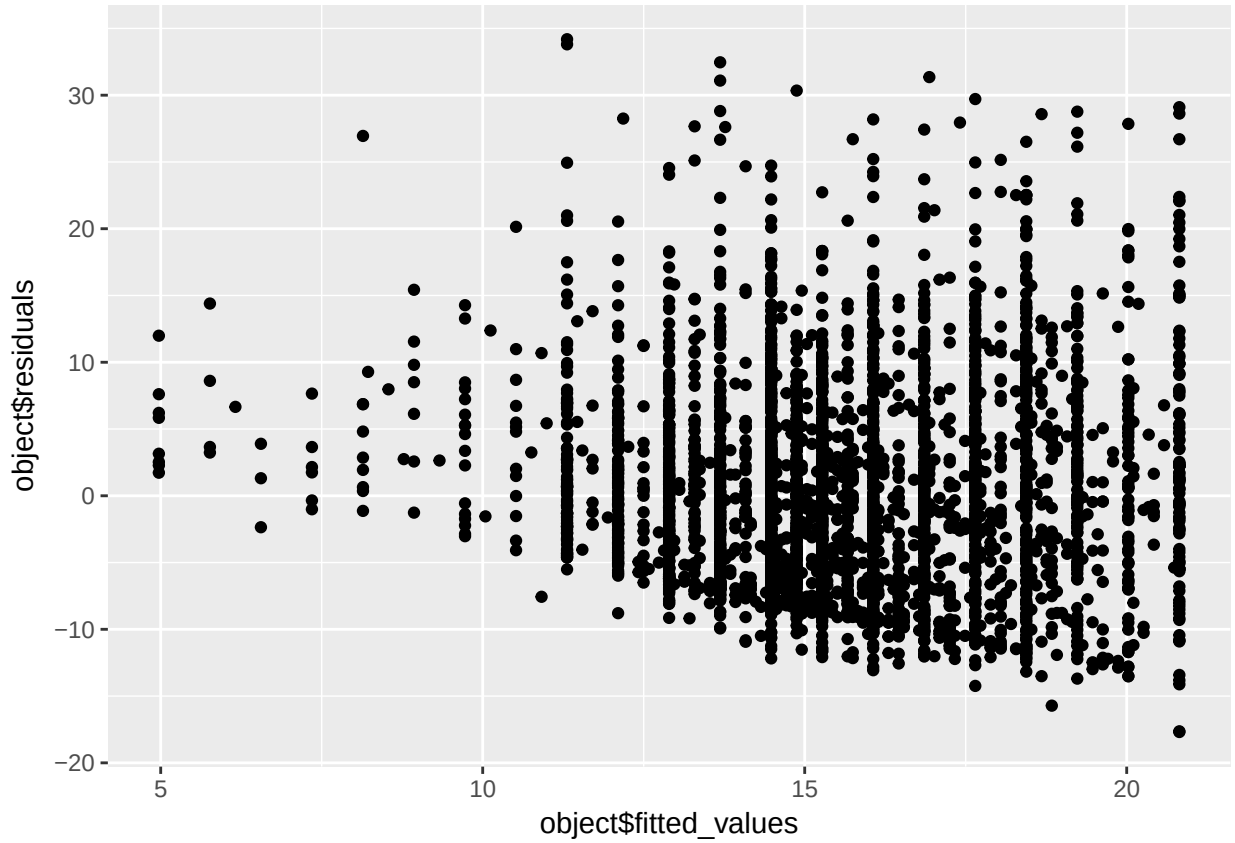
In our model we get the coefficient values $\hat{\beta}_0 = 4.97$ and $\hat{\beta}_1 = 0.79$ with the corresponding estimated standard errors $\widehat{SE}(\beta_0) = 0.53$ and $\widehat{SE}(\beta_1) = 0.039$. This gives us test statistics $z_0 = 9.30$ and $z_1 = 20.28$, which are both highly significant under our 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 expected mean value of the response, when `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 covariates on the real response Y, `wages`.

c)

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 residuals are mostly positive, whereas for higher values they are centered at 0. This is expected, as Y cannot be lower than 0 (negative wages does not make sense). The variance also seems to increase with higher fitted values, meaning that there is some violation of the homoscedasticity of the errors. At higher fitted values, the density has a larger left tail (towards higher values) than right tail, which is in violation with the normality assumption of the errors. These violations indicate that our model assumptions are not entirely reasonable and we should proceed with caution when conducting inference.

d)

The residual sum of squares `SSE` is the sum of squares $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The degrees of freedom for the `SSE` 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 $\sum_{i=1}^{n}(y_i - \bar{y})^2$. The regression sum of squares is the difference $SST - SSE$, with $p - 1$ degrees of freedom. Under H_0 from b) we have that

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

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

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

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

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

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

Because the p-value is below any reasonable significance level, we say that the regression is significant.

In simple linear regression, we have $p = 2$. The definition of a χ^2 variable 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, $z^2 \sim \chi_1^2$. But the χ^2-test statistic will also have $p - 1 = 2 - 1 = 1$ degree of freedom. So the square of the z-test statistic and the χ^2-test statistic will have the same distribution. This means that the square of the z-statistic and the χ^2 statistics will reject the null hypothesis at the same critical value.

We first confirm that the square of our z_1-statistic is equal to the χ^2-statistic:

```
cat(model1$coeff_z[2]^2, "\n")
```

```
## 411.4471
```

```
cat(model1$F_statistic)
```

```
## 411.4471
```

We then confirm that the critical values coincide by listing the critical Z-values of some some quantiles along with the square root 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, 0.8, 0.9, 0.95)
interval2 = c(0.4, 0.6, 0.8, 0.9)
cat(abs(qnorm(interval)), "\n")
```

```
## 0.5244005 0.8416212 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 distribution. In this example we suspect it to be a combination of both.

Part 3: Multiple linear regression

a)

So far we have implemented the `mylm` package to handle both simple and multivariate linear regression. For implementation details, see Part 2.

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

b)

```r
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 $\widehat{SE}(\beta_{age}) = 0.009$. This gives us test statistics $z_0 = -9.73$, $z_1 = 25.2$ and $z_2 = 28.7$, with all p-values $< 2 \cdot 10^{-16}$, meaning that all are highly significant under our model assumptions. We get a χ^2-statistic of 661 on 2 degrees of freedom and a p-value of $< 2 \cdot 10^{-16}$, which indicates that the total regression is also highly significant.

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 expected mean value of the response, when `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 covariates on the real response Y, `wages`.

c)

```r
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
```

The parameter estimates for the simple and multiple regression differ if the covariates are correlated (not orthogonal). Then, in the case of the simple regression, the single explanatory variable pick up some of the variance that would otherwise be explained by the additional explanatory variable. The (p, p) matrix $(X^T X)$ is

$$\begin{pmatrix} x_0^T x_0 & x_0^T x_1 & x_0^T x_2 \\ x_1^T x_0 & x_1^T x_1 & x_1^T x_2 \\ x_2^T x_0 & x_2^T x_1 & x_2^T x_2 \end{pmatrix},$$

If we have uncorrelated (orthogonal) covariates, we get that $x_i^T x_j = 0 \quad \forall i \neq j$, such that $(X^T X)$ becomes

$$\begin{pmatrix} x_0^T x_0 & 0 & 0 \\ 0 & x_1^T x_1 & 0 \\ 0 & 0 & x_2^T x_2 \end{pmatrix},$$

which is diagonal and thus each coeffecient estimate $\hat{\beta}_j = (x_j^T x_j)^{-1} x_j^T y_j$ is independent of the other covariates.

In our case we get different estimates in the simple and mupltiple case, which is reasonable as `education` and `age` have a weak positive correlation.

```
model3$corr_coeff
```

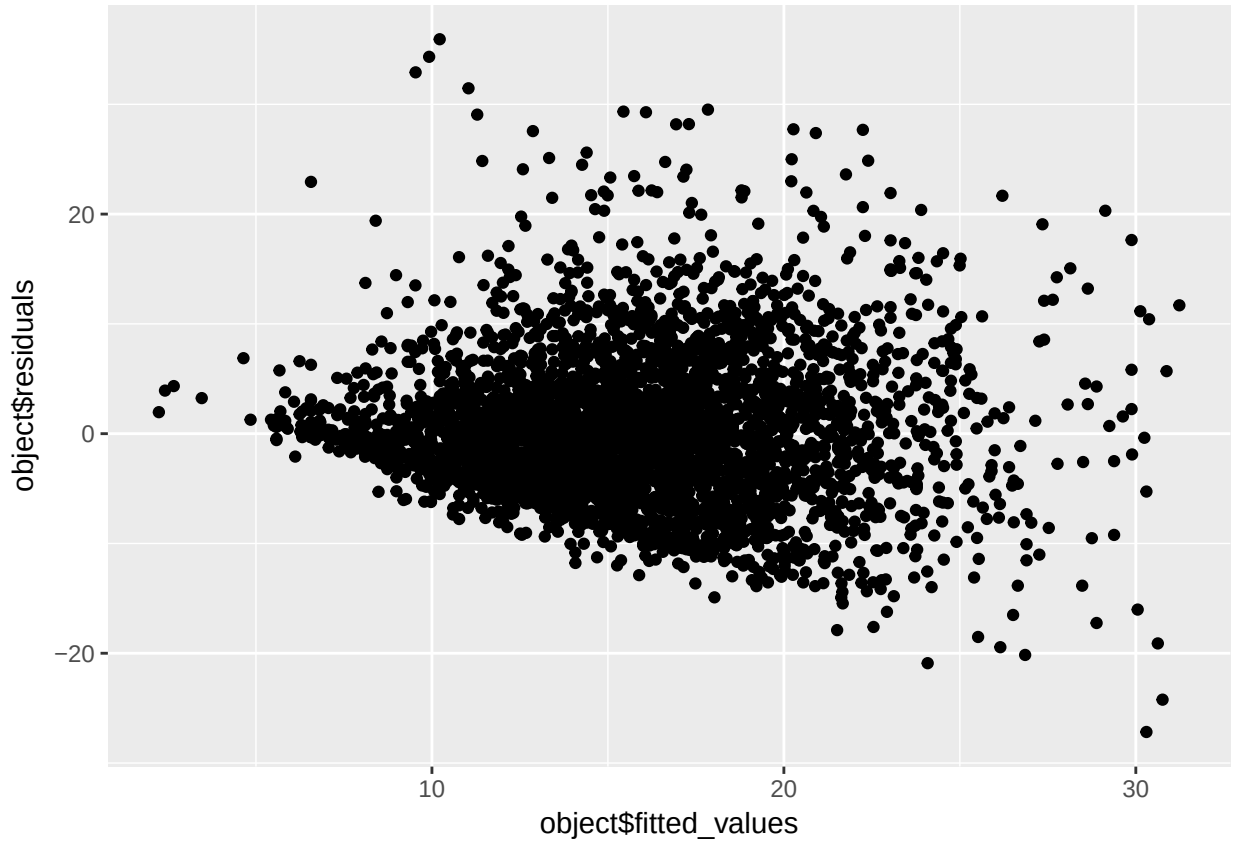
```
##             (Intercept)  education        age
## (Intercept)   1.0000000 -0.8276025 -0.6184303
## education    -0.8276025  1.0000000  0.1062790
## age          -0.6184303  0.1062790  1.0000000
```

Part 4: Testing the `mylm`package

```
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 adding a squared term, `education`^{2}. Our linear model assumption is now

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

Fitting our model, we see positive terms for the covariates `sex`, `age` and `education$^2$` and negative for both levels of `language`.

The coefficient $\hat{\beta}_{edu^2}$ can be interpreted as the following: increasing the square of `education` by one unit will increase the model response, $\hat{Y}$, by 0.035 units. We also note that covariates that are fitted to a negative coefficient will produce a decrease in the model response $\hat{Y}$. All remaining coefficients can be interpreted as in 3b).

We notice that all variables attain a high level of significance with the exception of language. Thus, to improve the model, language should be removed, as even under model assumptions it is probable that it is uncorrelated with the response. The χ^2-statistic of 345 on 5 degrees of freedom with p-value $< 2 \cdot 10^{-16}$ indicates that the total regression is highly significant.

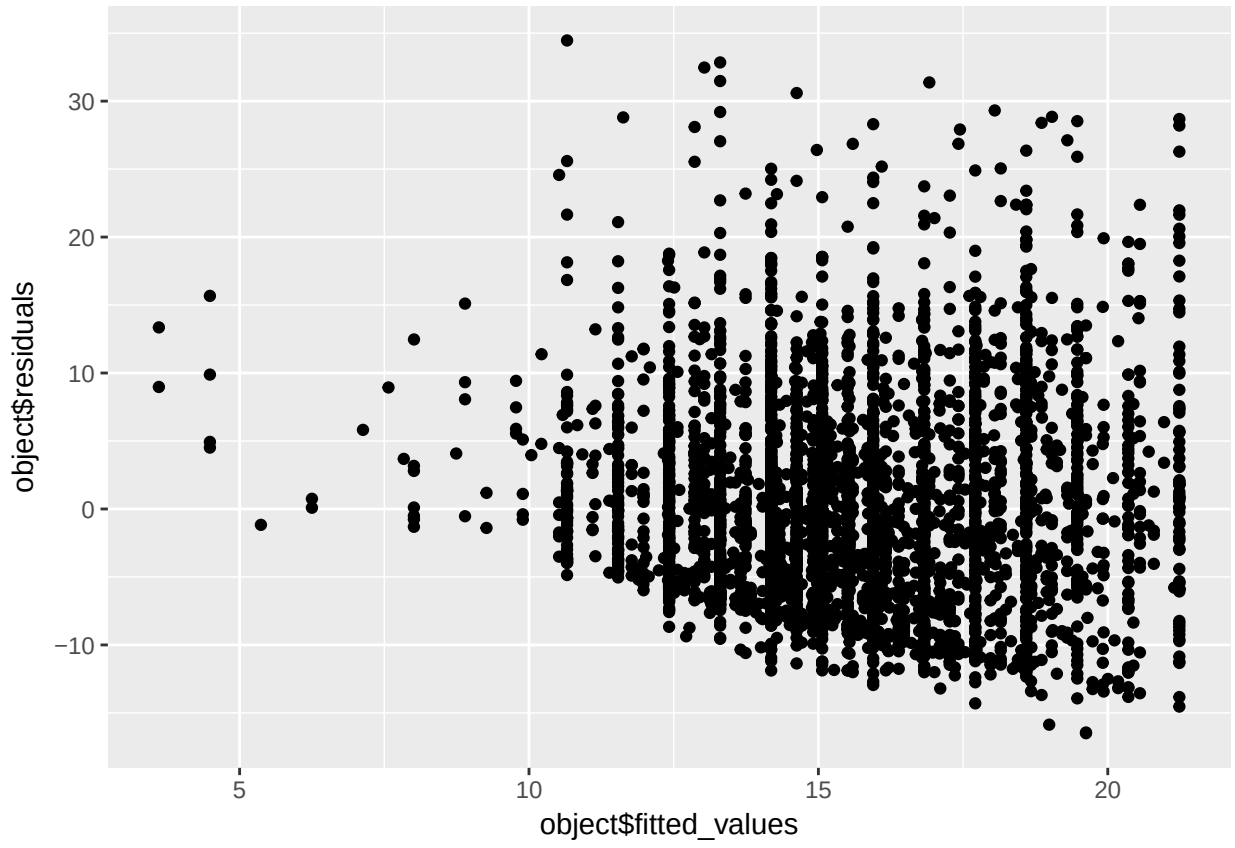
The error plot shows a lower bound of the residuals in the left corner. Again the explanation is that we cannot have negative `wages`. Beyond this, we observe only some level of homoscedasticity and normality in our errors, which indicates that our model assumptions are not entirely reasonable.

```
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,

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

We interpret the interaction term between `language` and `education` in the following way. When the level of language increases from 0 to 1 (from `English` to `French`), the response is increased by the coefficient estimate

for `French` plus an interaction term, which is the coefficient estimate for the interaction multiplied by the `education` covariate,

$$Y_{increase} = \beta_{languageFrench} + \beta_{edu\&lan} x_{edu}.$$

A χ^2-statistic with a corresponding p-value $< 2 \cdot 10^{-16}$ signifies that the total regression is significant. The intercept, `languageOther` and `education` are all significant on a 0.001 level. The interaction between `languageOther`and `education` is significant on a 0.01 level, `languageFrench` on a 0.05 level and the interaction between `languageFrench` and `education` on a 0.1 level.

Seeing as `languageFrench` and `languageOther` have roughly the same coefficient estimate, one could use the natural assumption that they have the same effect and merge these two levels into one, giving `language` only two levels.

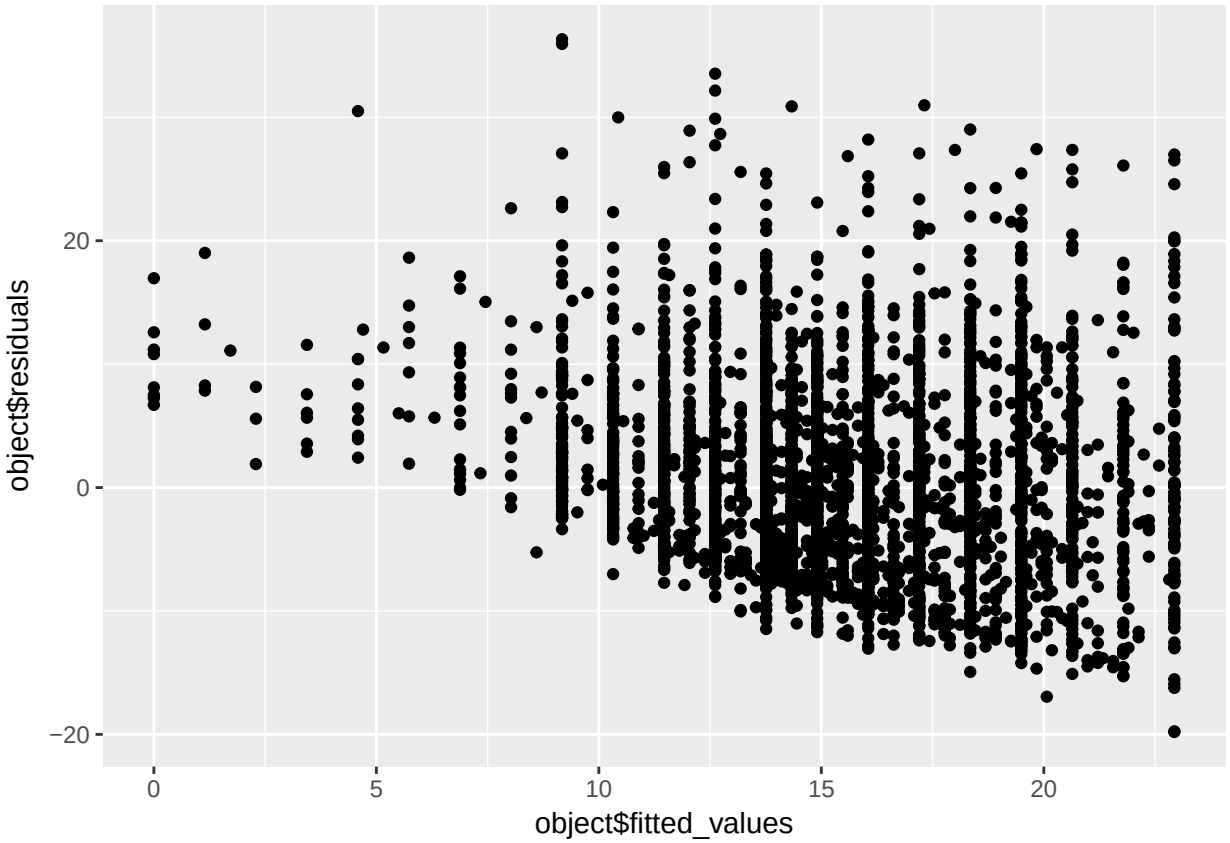
In addition, a multiple R-squared of 0.1 can be interpreted as the model explaining 10% of the variance, this is not a high number, and the model could be improved by for example adding another covariate, e.g. `sex` or `age`.

The error plot shows the same lower bound as previously discussed. We observe a trend of increasing variance with higher fitted values. For lower fitted values we see some off-centering but around a relatively small amount of points. Again we observe a larger left tail. There is significant violation of both the centering, homoscedasticity and normality of the errors.

```
model42 <- mylm(wages ~ education - 1, data = SLID)
summary(model42)
```

```
## 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
```

```
plot(model42)
```



Now we have a model without intercept. Our linear model assumption is now

$$Y_i = \beta_{edu} x_{i,edu} + \epsilon_i.$$

The interpretation of this model is that `wage` is only a scaling of `education`, with some individual differences ϵ_i. From a realistic point of view, this model does not make sense; one would not expect zero income with zero education, as some jobs does not require education at all. Thus, the model would be improved if we add intercept.

The coefficient estimate of `education` is positive and significant at a 0.001 level, which is as expected because higher education usually equates to higher income.

The error plot strongly resembles that of the previous example and induces the same conclusions.