Assignment 5

Jenni Simon
09-116-005

Exercise 1 Figure 1 and listing 1 show results for the multiple linear regression model with target variable "Wage".

As can be expected we observe that ID is not significant (high value for $P(>|t|)$) while the other variables are clearly significant (very small $P(>|t|)$). Looking at the t-values, we see that "Education" is the most significant predictor (highest value for $|t|$). The residual plot in figure 1 indicates that the linear model fits the data quite well, given that the approximate mean (red line) is close to zero and the variance of the residuals appears close to constant. We observe that "Wage" increases with "Education" as well as being "Male".

Listing 1: Summary of multiple regression fit for exercise 1

```
t = 2.0074, df = 9, p-value = 0.07564
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -342.6602 5743.2296
sample estimates:
mean of the differences
               2700.285
```

Exercise 2 Figure 2 and listing 2 show results for the multiple linear regression model with target variable 'PRP' after performing forward selection for the variables to be included.

Forward selection was done by starting with an initial model using only an intersection term and then adding variables one by one by finding the variable which decreases the RSS the most. This is done for as long as the new model

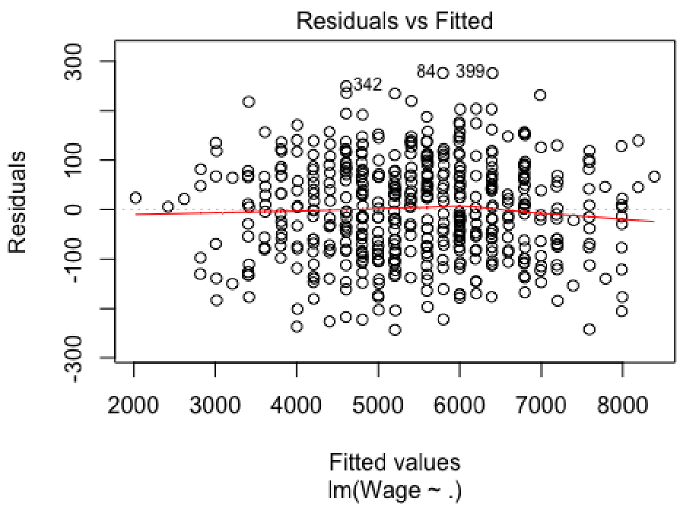

Fig. 1: Plot showing the residuals against the fit of the multiple regression model of exercise 1.

significantly reduces the RSS measured using F-statistics.

By construction and looking at the included variables in listing 2 we observe that given "MMAX", "CACH", "MMIN", "CHMAX" and "MYCT" the variables "MMIN" and "CGMIN" are not significant in predicting "PRP". Also, "PRP" increases with all the included variables.

Figure 2 also uncovers some quadratic behavior in the residual-plot which is not captured by the purely linear model. Extending the model with some quadratic terms would most probably lead to better results here.

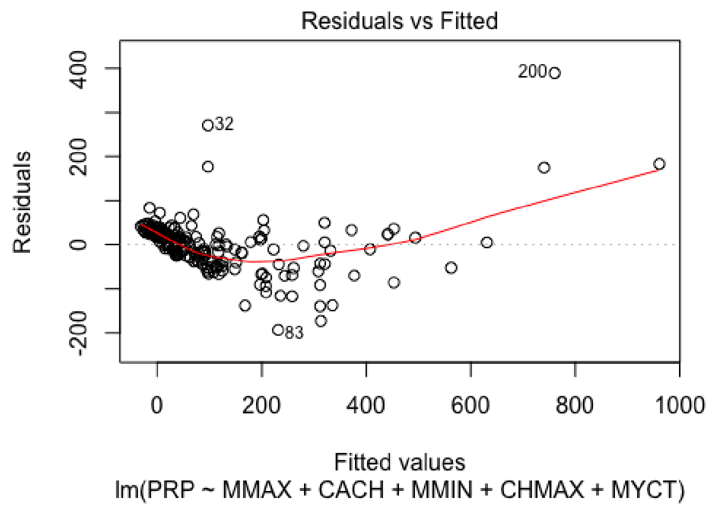

Fig. 2: Plot showing the residuals against the fit of the multiple regression model of exercise 2.

Listing 2: Summary of multiple regression fit for exercise 2

```
Residuals:
    Min      1Q  Median      3Q     Max
-193.37  -24.95    5.76   26.64  389.66

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.608e+01  8.007e+00  -7.003 3.59e-11 ***
MMAX         5.562e-03  6.396e-04   8.695 1.18e-15 ***
CACH         6.298e-01  1.344e-01   4.687 5.07e-06 ***
MMIN         1.518e-02  1.788e-03   8.490 4.34e-15 ***
CHMAX        1.460e+00  2.076e-01   7.031 3.06e-11 ***
MYCT         4.911e-02  1.746e-02   2.813   0.0054 **
---

Residual standard error: 59.86 on 203 degrees of freedom
Multiple R-squared:  0.8648,    Adjusted R-squared:  0.8615
F-statistic: 259.7 on 5 and 203 DF,  p-value: < 2.2e-16
```

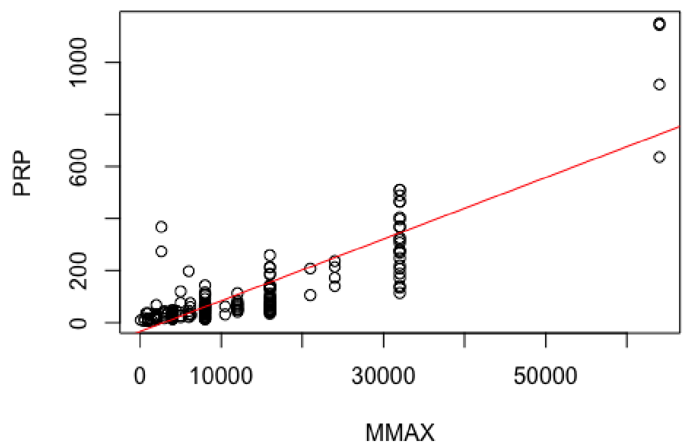

Fig. 3: Plot showing the linear regression model of exercise 3.

Exercise 3 The most important variable of the linear model of exercise 2 and also the first variable added during forward selection is "MMAX". Figure 3 shows a plot of the linear model using "MMAX" only against "PRP". We observe that the linear model clearly underestimates at $MMAX = 0$ and around $MMAX = 6500$. This suggest that probably a quadratic term $MMAX^2$ model would fit better.

Exercise 4 Figure 4 and listing 3 show results for the multiple linear regression model with target variable 'mpg' after performing forward selection for the variables to be included.

By construction and looking at the included variables in listing 2 we observe that given "weight", "year" and "origin" the variables "acceleration", "horsepower", "displacement" and "cylinders" are not significant in predicting "mpg". Also, "mpg" increases with "year" and decreases with "weight". Supposing that "origin" is a categorical variable with vales USA (1), Asia (3) or Europe (2), USA has the lowest value of mpg followed by Europe and Asia.

Figure 4 shows some non-linear behavior in the residual-plot which is not captured by the linear model. Extending the model with some higher order or

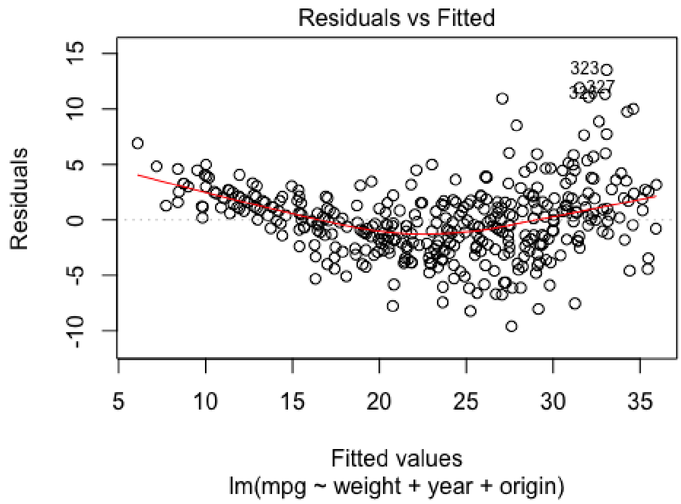

Fig. 4: Plot showing the residuals against the fit of the multiple regression model of exercise 4.

interaction terms would most probably lead to better results. We also observe some non-constant variance (i.e. heteroscedasticity?) in the data.

Listing 3: Summary of multiple regression fit for exercise 4

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.6025 -2.1132 -0.0206  1.7617 13.5261

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.831e+01  4.017e+00  -4.557 6.96e-06 ***
weight      -5.887e-03  2.599e-04 -22.647  < 2e-16 ***
year         7.698e-01  4.867e-02  15.818  < 2e-16 ***
origin2      1.976e+00  5.180e-01   3.815 0.000158 ***
origin3      2.215e+00  5.188e-01   4.268 2.48e-05 ***
---

Residual standard error: 3.337 on 387 degrees of freedom
Multiple R-squared:  0.819,     Adjusted R-squared:  0.8172
F-statistic: 437.9 on 4 and 387 DF,  p-value: < 2.2e-16
```

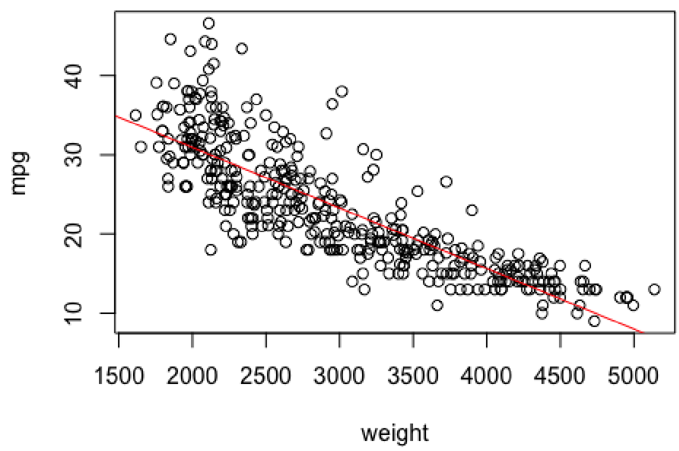

Fig. 5: Plot showing the linear regression model of exercise 4.

The most important variable of the linear model and also the first variable added during forward selection is "weight". Figure 5 shows a plot of the linear model using "weight" only against "mpg". We again observe that the model tends to underestimate at the borders and overestimate in the middle. Probably a quadratic term $weight^2$ would give a better fit.