Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

1 Exercise 1

a) During *supervised* learning we have a set of inputs $X = \{x_1, \ldots, x_n\}$ with it's corresponding outputs $Y = \{y_1, \ldots, y_n\}$ for all observations. In this case our goal is to approximate the coefficients of a a function $\hat{f}$ such that $\hat{f}(x) \approx y$ for all $x \in X$.
In the case of *unsupervised* learning, we only have the set of inputs X and try to find relationships and patterns in the dataset.

b) *Prediction* refers to the estimate of the correct output given an (unseen) observation. (In this case we are mostly interested in reducing the error of our prediction).
Inference on the other hands has the goal to find relationships between the inputs to compute the output.

c) *Classification* is used to map a datapoint to a given group or cluster that is known beforehand to exist in the dataset. This is often used when the output of our observations are of categorical nature.

Regression on the other hand fits a line (in dimensions > 2 a (hyper)plane) to model the output given the input. This results in a continuous output.

d) *Training data* refers to the subset of our observations used to train our model. Following the training the *test data*, consists of the rest of our data, which has not been seen by our model yet. Test data can be used to evaluate the performance/accuracy of our model, as it "simulates" the model in the real world.

e) When choosing a *parametric* model we make an assumption about the form of the function f. This makes it easier to estimate the (known) model parameters, but also may not capture the true form of f.
Non parametric models don't make an assumption about the functional form, this may produce a more accurate model for f, but also requires more data.

(277 of 300 words used)

Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

2 Exercise 2

I order to prove that $E[(y_0 - \hat{f}(x_0)^2] = Bias(\hat{f}(x_0))^2 + Var(\hat{f}(x_0)) + Var(\epsilon)$, we establish some side proofs. We will use these in to proof the equation.

Proof (I).

$$E[(x - E[x])^2] = Var[x] \tag{1}$$

$$\begin{aligned} E[(x - E[x])^2] &= E[x^2 - 2xE[x] + E[x]^2] \\ &= E[x^2 - 2xE[x]] + E[x]^2 \\ &= E[X^2] - 2E[x]E[x] + E[x]^2 \\ &= E[x^2] - E[x]^2 \\ &= Var[x] \end{aligned}$$

■

Proof (II).

$$E[\hat{f}(x_0) - f(x_0)]^2 = Bias(\hat{f}(x_0))^2 \tag{2}$$

$$\begin{aligned} E[\hat{f}(x_0) - f(x_0)]^2 &= E[\hat{f}(x_0)]^2 - 2E[\hat{f}(x_0)]E[f(x_0)] + E[f(x_0)]^2 \\ &= E[\hat{f}(x_0)]^2 - 2E[\hat{f}(x_0)](E[f(x_0)] + E[\epsilon]) + (E[f(x_0)] + E[\epsilon])^2 \\ &= E[\hat{f}(x_0)]^2 - 2E[\hat{f}(x_0)]E[f(x_0) + \epsilon] + E[f(x_0) + \epsilon]^2 \\ &= E[\hat{f}(x_0)]^2 - 2E[\hat{f}(x_0)]E[y_0] + E[y_0]^2 \\ &= E[\hat{f}(x_0) - y_0]^2 \\ &= Bias(\hat{f}(x_0))^2 \end{aligned}$$

■

Proof (III). This is a proof for equation (2.3) in ISLR.

$$E[(y - \hat{y})^2] = (f(x) - \hat{f}(x))^2 + Var(\epsilon) \tag{3}$$

Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

$$
\begin{aligned}
E[(y-\hat{y})^2] &= E[(f(x)+\epsilon-\hat{f}(x))^2] \\
&= E[f(x)^2+\epsilon^2+2f(x)\epsilon-2\hat{f}(x)\epsilon+2f(x)\hat{f}(x)] \\
&= E[f(x)^2]+E[\epsilon^2]+E[\hat{f}(x)^2]+\underbrace{2E[f(x)]E[x]}_{0}-\underbrace{2E[\hat{f}(x)]E[x]}_{0}-2E[f(x)]E[\hat{f}(x)] \\
&= \underbrace{E[f(x)^2]}_{\text{const.}}-\underbrace{E[2f(x)\hat{f}(x)]}_{\text{const.}}+\underbrace{E[f(x)^2]}_{\text{const.}}+E[\epsilon^2] \\
&= f(x)^2-2f(x)\hat{f}(x)+\hat{f}(x)^2+E[\epsilon^2]+\underbrace{E[\epsilon]^2}_{0} \\
&= (f(x)-\hat{f}(x))^2+Var(\epsilon)
\end{aligned}
$$

■

Proof.

$$
\begin{aligned}
&E[(y_0-\hat{f}(x_0)^2] \\
&= [f(x_0)-\hat{f}(x_0)]^2+Var(\epsilon) && \text{| Side proof III} \\
&= \underbrace{f(x_0)^2}_{\text{const.}}-2\underbrace{f(x_0)}_{\text{const.}}\underbrace{\hat{f}(x_0)}_{\text{const.}}+\underbrace{\hat{f}(x_0)^2}_{\text{const.}}+Var(\epsilon) && \text{| with } c=E[c] \\
&= E[f(x_0)^2]-2E[f(x_0)]E[\hat{f}(x_0)]+E[\hat{f}(x_0)^2]+Var(\epsilon) && |E[c^2]=E[c]^2 \\
&= E[f(x_0)]^2-2E[f(x_0)]E[\hat{f}(x_0)]+E[\hat{f}(x_0)]^2+Var(\epsilon) && |+E[\hat{f}(x_0)]^2-E[\hat{f}(x_0)]^2 \\
&= \underbrace{E[f(x_0)]^2-2E[f(x_0)]E[\hat{f}(x_0)]+E[\hat{f}(x_0)]^2}_{E[f(x_0)-\hat{f}(x_0)]^2}+\underbrace{E[\hat{f}(x_0)]^2-E[\hat{f}(x_0)]^2}_{Var(\hat{f}(x_0))}+Var(\epsilon) && \\
&= E[f(x_0)-\hat{f}(x_0)]^2+Var(\hat{f}(x_0))+Var(\epsilon) && |Var(x)=E[x^2]-E[x]^2 \\
&= E[\hat{f}(x_0)-f(x_0)]^2+E[(\hat{f}(x_0)-E[\hat{f}(x_0)]^2)]+Var(\epsilon) && |\text{Side proof I} \\
&= Bias(\hat{f}(x_0))^2+Var(\hat{f}(x_0))+Var(\epsilon) && |\text{Side proof II}
\end{aligned}
$$

■

Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

3 Exercise 3

a) The Gauss Markov Theorem states, that the linear model fitted using the least squares estimates has the lowest variance of all linear estimates, as long as we assume that the estimate is unbiased. This is important because we can than simply use the least square estimate to get the (unbiased) estimators with the lowest MSE without regard of our data.

b) The Gauss-Markov theorem assumes, that the irreducible error ε of an observation has a mean of zero, which also implies a constant variance and that the error terms of distinct observations are uncorrelated.
This can be expressed algebraically by the following terms:
Mean of zero:

$$\mathbb{E}(\varepsilon_j) = 0$$

Constant variance:

$$\mathrm{Var}(\varepsilon_j) = c$$

Uncorrelated error terms for $j \neq k$:

$$\mathrm{Cov}(\varepsilon_j, \varepsilon_k) = 0$$

c) Because the exptest test MSE is composed of only the bias of the estimation, the variance of the estimation and the variance of the irreducible error, it will be the best linear unbiased estimate because the irreducible error is the same for all estimates and we already now the variance of the estimation is the lowest for the least square estimates.

Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

4 Exercise 4

1) If we have very few observations compared to the number of dimensions, e.g. (the example from the book) 100 observations and 20 dimensions, it can happen that most of the datapoints lie very far apart. In return this means that the "nearest neighbors" of these points are not near at all and may not even carry any meaningful information about those points, thus non parametric models may perform worse than usual. This is the phenomenon described by "curse of dimensionality", the more dimensions the more data is needed.

2a) As our observations are uniform distributed we are looking to calculate the probability for a point p that is falls inside the 10 % range

$$P(x - 0.05 \leq p \leq x + 0.05) = \frac{x + 0.05 - (x - 0.05)}{1 - 0} = 0.1$$

On average 10% of our observations fall into this range.

2b) In this case we generalize into two dimensions and instead taking the lengths we are using the area of the ranges to calculate the probability:

$$P(p \in [x - 0.05, x + 0.05] \times [x - 0.05, x + 0.05]) = \frac{0.1^2}{1^2} = 0.01$$

2c) As we have seen in the two exercises above if we have m features the fraction of our observations in the 10% range can be calculated by:

$$P(p \in I) = 0.1^m$$

Where I denotes the spaces that contains 10% of the observations around x.

2d) Because the hypercube is a generalization of a cube into more dimesnions we can calculate the length of it as we would in the 1, 2 or 3 dimensional case:

$$l = \sqrt[p]{0.1}$$

Where m is the number of dimensions. Thus we get the following values for $p = 1, 2, 100$:

$$\begin{aligned} p = 1 &\implies l = 0.1 \\ p = 2 &\implies l \approx 0.316 \\ p = 100 &\implies l \approx 0.977 \end{aligned}$$

Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

5 Exercise 6

1.) Looking at the plots we see a positive correlation between horsepower and displacement. On the other hand mpg and displacement seem to be negatively correlated. While horsepower, displacement, weight and mpg all seem to be correlated positively, acceleration is anti-correlated with all of those except mpg, with which it does not seem to be related at all. From the plots one can also conclude, that year has not relationship to any of the other values.

2.) Looking at the plots horsepower and weight seem to be correlated the most, while weight and mpg seem to be highly anti correlated. Looking at the correlation matrix confirms our guess about the correlation, however cylinders and displacement have an even higher correlation than horsepower and weight which one cannot easily tell from the plot.

3.) According to the p-values all predictors are statistically significant. The fitted models give the following R^2 values:

cylinders	0.604688988944125
displacement	0.648229400319304
horsepower	0.605948257889435
year	0.337027813309623

4a) As the model fitted using all predictors except "name" has a R^2 value of 0.8215 it performs better than all models above.

4b) Even tough above all the predictors seemed to be statistically significant this is not the case anymore, as for example cylinders and horsepower have p-values over 0.1, making them insignificant. If the sign of a coefficient is negative the predictor and the result are anti-correlated. If we for example predict the mpg using 100 as weight and then 1000 as weight we see that our estimate decreases ($42.08 \rightarrow 36.25$).

5) The residual plot is shown in the upper left corner of the plots. As the residual plot does display a slight U-Shape, which implies non linearity in our data. Looking at the plot we see some potential outliers (323, 327) etc. which can be confirmed when plotting the studentized residuals as there are some values above 3.

Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

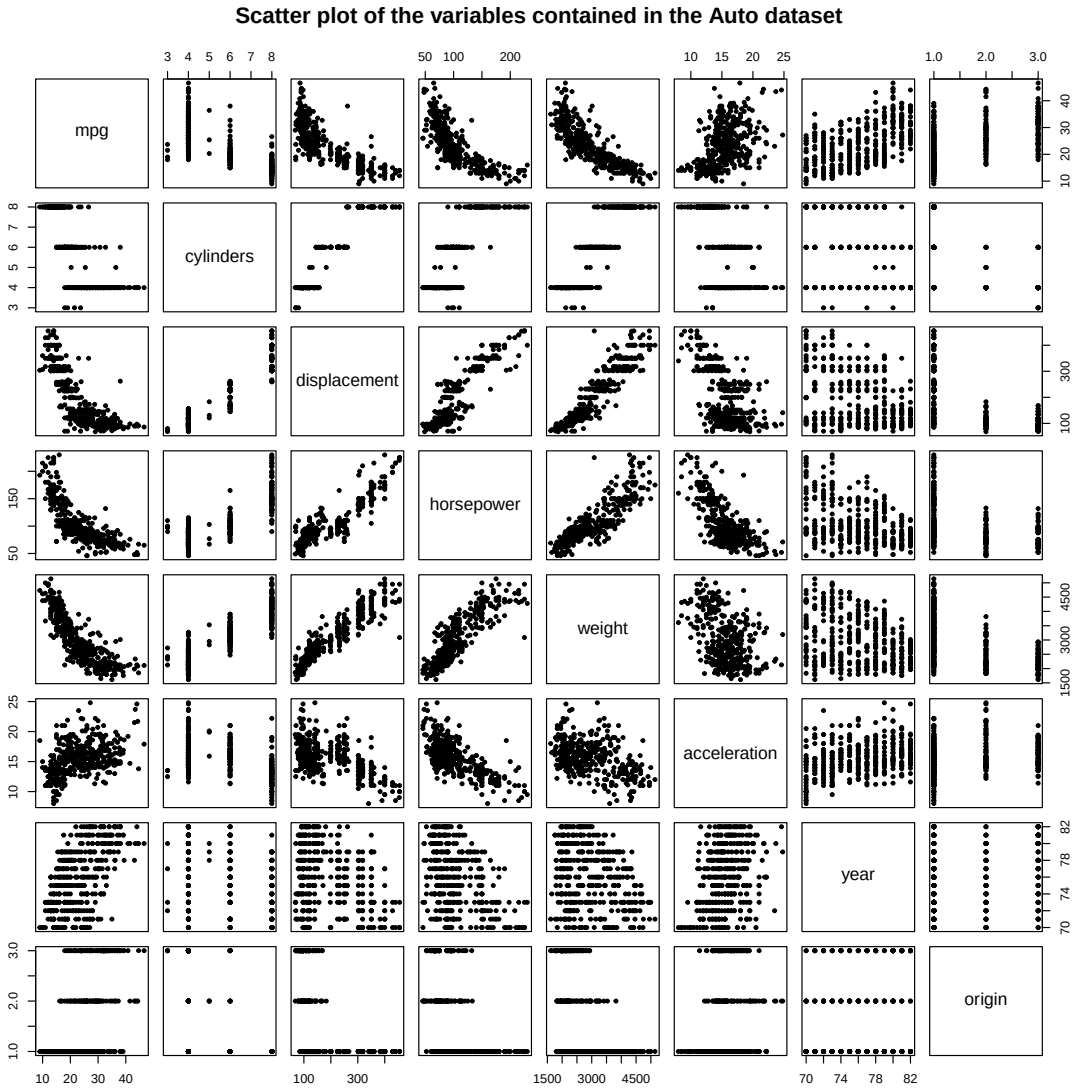

Figure 1: Scatter plots of all variables in the auto dataset

Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

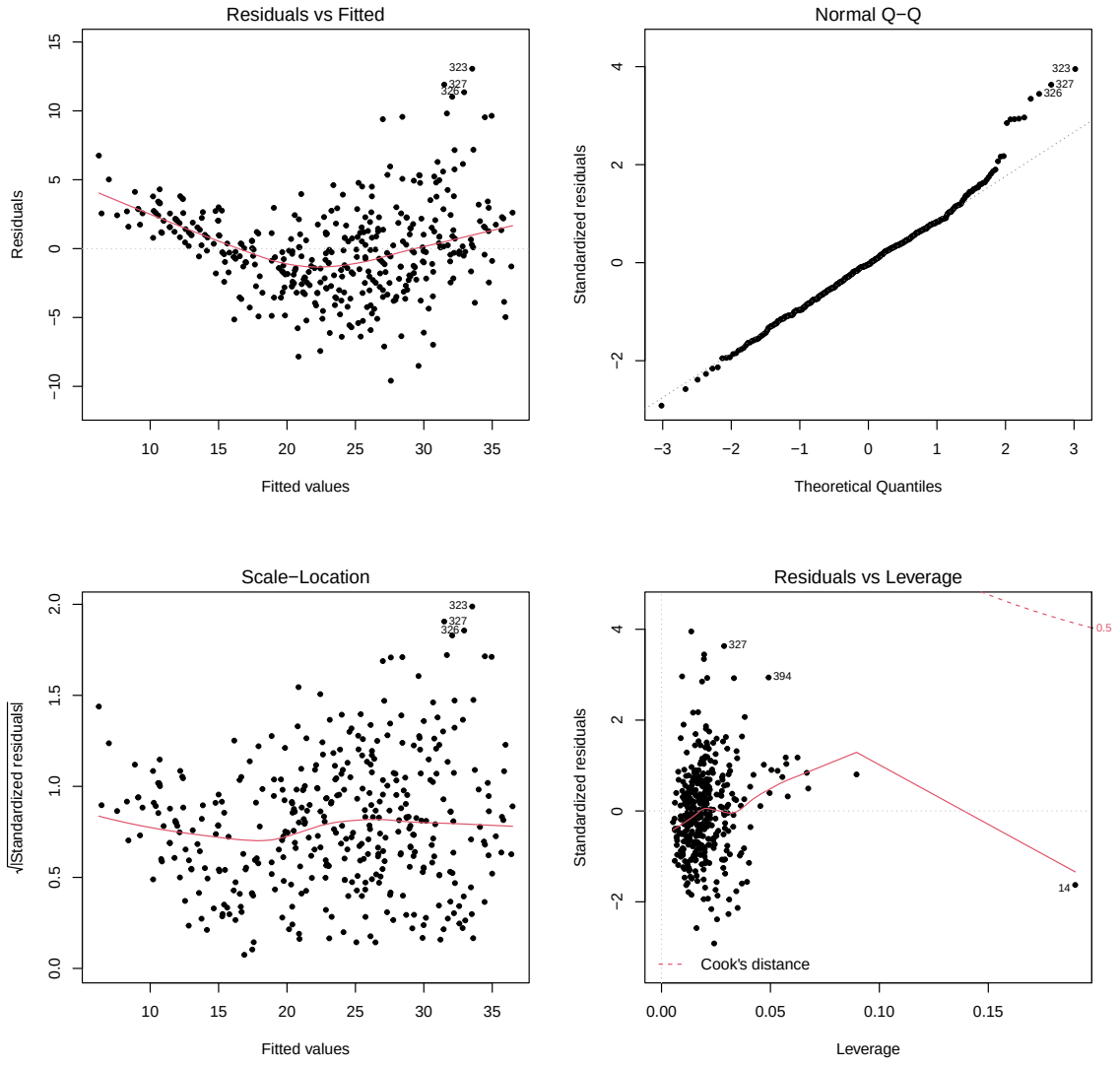

Figure 2: Diagnostic plots ouf our fitted model

Max Rausch-Dupont (7006185)
Felix Ceard-Falkenberg (7007490)

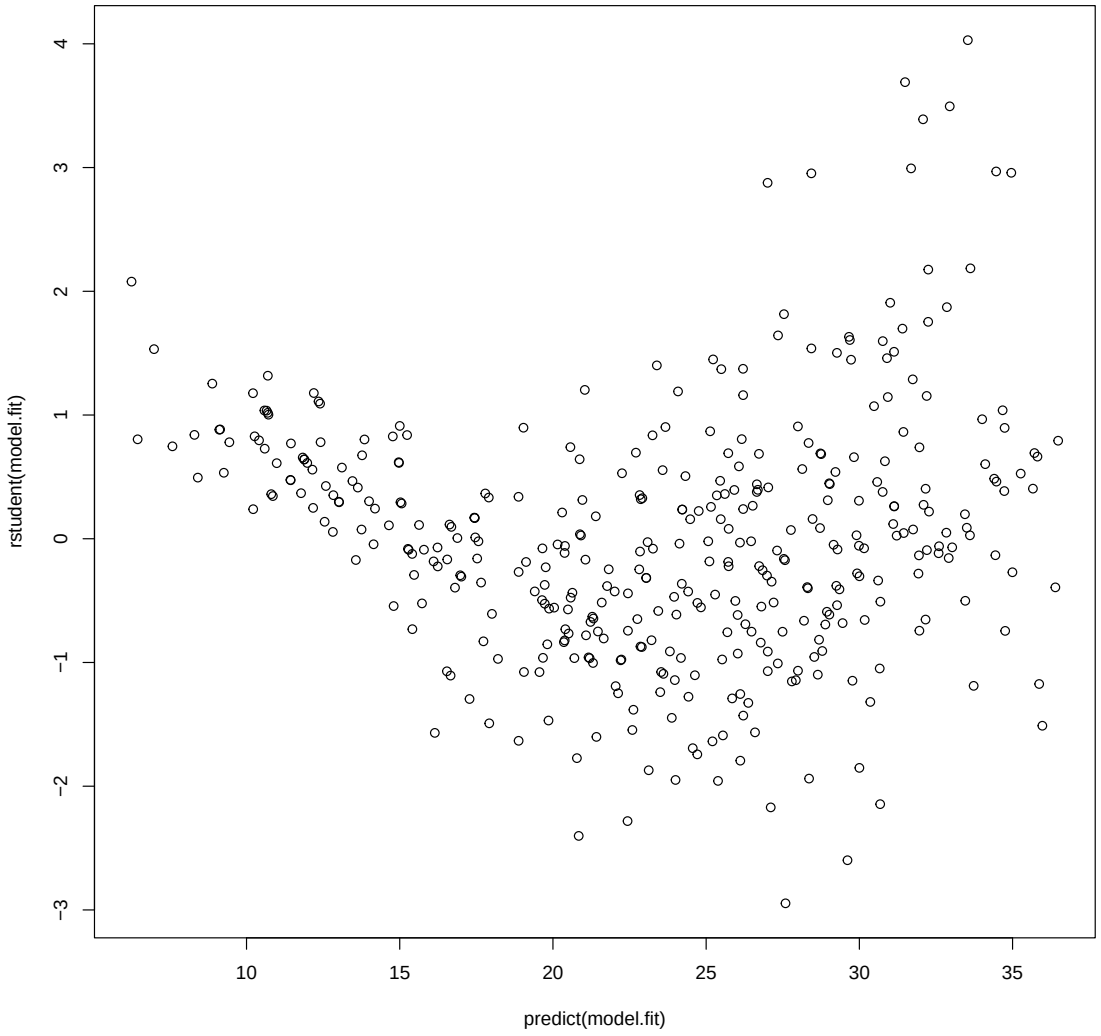

Figure 3: Studentized residual