Assignment 4

Jenni Simon
09-116-005

Exercise 1 Figure 1 and table 1 show the linear regression model for the male examples and figure 2 and table 2 the results for females. We observe that the estimated intercept is significantly higher for the female set, while the proportionality (linear coefficient in the model) is only slightly greater compared to the male set.

The model therefore predicts women to have a higher wage compared to men given the same amount of education.

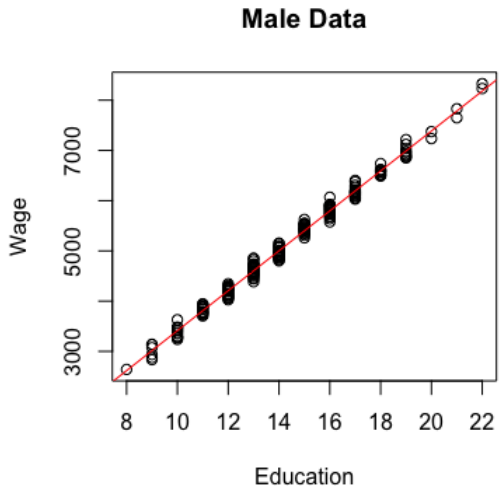

Fig. 1: Plot showing years of education vs. wage for men. Linear regression model given by the red line.

Tab. 1: Linear regression model for men.

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-563.61	37.50	-15.03	<2e-16
Education	397.54	2.58	154.10	<2e-16

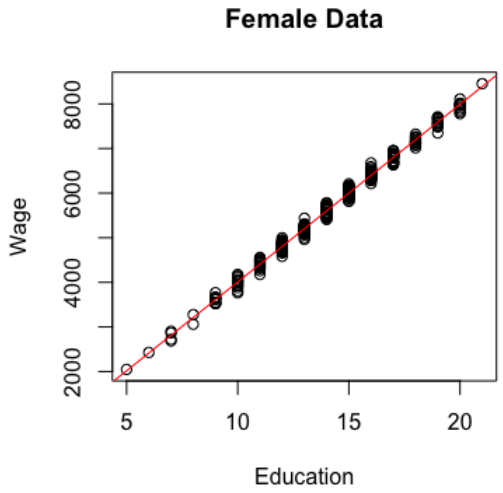

Fig. 2: Plot showing years of education vs. wage for women. Linear regression model given by the red line.

Tab. 2: Linear regression model for women.

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	24.199	27.937	0.866	0.387
Education	398.250	1.919	207.559	<2e-16

Exercise 2 Looking at the description of the data, the following attributes can not be used to predict the performance:

- Vendor and model: As they are non-descriptive and non-exhaustive (not all vendors/models). Therefore useless for prediction.
- PRP: The goal field.
- ERP: The linear regression guesses.

Exercise 3 Fitting a model using all possible predictors (excluding the ones from Ex2) we choose the most significant variable, i.e. the one with the highest value for $|t|$ or the lowest value for $P(>|t|)$ respectively. This turned out to be the variable **MMAX**.

Table 3 shows the confidence intervals for the estimated model parameters. We are confident in observing a linear increase of PRP with MMAX.

Tab. 3: Confidence interval for the linear regression model $PRP \sim MMAX$.

	2.5%	97.5%
Intercept	-49.77265640	-18.22582429
MMAX	0.01088673	0.01278561

Exercise 4 Figure 3 shows the found linear relationship.

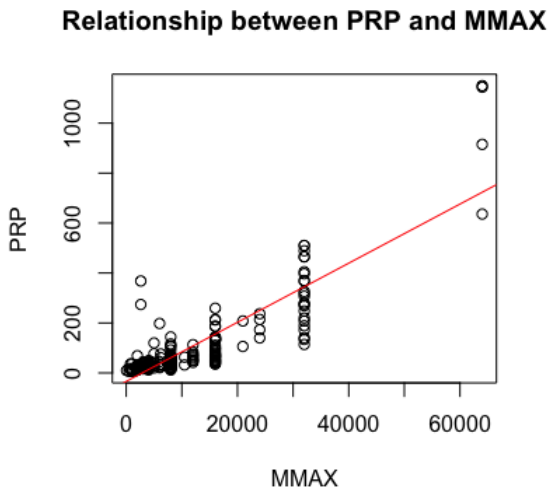

Fig. 3: Plot showing MMAX vs. PRP. Linear regression model given by the red line.

Exercise 5 The data contained examples with NA values for the variable horsepower. These examples have been removed from the dataset prior to further analysis.

The name of the car cannot be used because if we would use it, our model

would not be able to predict mpg for any previously unseen car-model. It is still interesting to observe what happens when we incorporate the variable "name" in a fit however: We observe that in cases where the name contains (diesel), the name is actually very significant. This suggests that a categorical variable indicating the fuel-type would be useful for the model.

It is a bit unclear how origin is defined, but it seems like it indicates whether the car was produced in the USA (1), Asia (3) or Europe (2). As long as we do not want to predict mpg for cars originating from outside of these zones (whatever that might be) and therefore assume that all cars can be classified as having origin 1, 2 or 3, it makes sense to use this attribute. Given the uncertainty of the definition however, it would be sensible to exclude this variable.

Exercise 5 Fitting a model using all possible predictors (excluding the ones from Ex4) we choose the most significant variable. This turned out to be the variable **weight**.

Table 4 shows the confidence intervals for the estimated model parameters. We are confident in observing a linear decrease of mpg with weight.

Tab. 4: Confidence interval for the linear regression model $mpg \sim weight$.

	2.5%	97.5%
Intercept	44.646282308	47.78676679
weight	-0.008154515	-0.00714017

Exercise 6 Finally, figure 4 shows the found linear relationship.

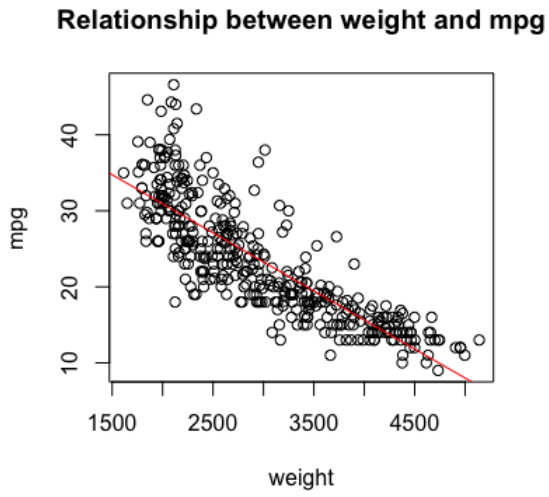

Fig. 4: Plot showing mpg vs. weight. Linear regression model given by the red line.