Report Regression Models

Cocu23

In this excercise the mtcars dataset will be analyzed regarding the following two main questions:
- Is an automatic or manual transmission better for miles per gallon (MPG)?
- How different is the MPG between automatic and manual transmissions?

Statistical inference, regression models and exploratory data analyses were applied to mainly explore how automatic (am = 0) and manual (am = 1) transmissions features affect the MPG feature. The t-test shows that the performance differs between cars with automatic and manual transmission. Using simple linear regression analysis,the signficant difference between the mean MPG for automatic and manual transmission cars will be revelead and proved. As a result, manual transmissions achieve a higher value of MPG compared to automatic transmission. This increase is approximately 1.8 MPG when switching from an automatic transmission to a manual one, with all else held constant. The results of the regression analysis are not printed out due to reasons of simplicity. However, the interested reader can print it out by running the code easily.

1. Preprocessing

```
# Load required packages and load dataset
library(ggplot2)
data(mtcars)
# basic transformation of data types (change type of necessary variables for
m numeric to factor)
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

2. Basic exploratory data analysis

Some basic exploratory data analysis was conducted by inspecting different plots. The respective code junks and figures can be found in the appendix at the end of this document. According to the box plot, generally speaking manual transmission yields higher values of MPG. The pair graphs indicate higher correlations between variables like “wt”, “disp”, “cyl” and “hp”.

3. Inference

The null hypothesis states that the MPG of the automatic and manual transmissions are from the same population. Therfore a normal distribution of MPG is assumed. The two sample t-test lead to the following results:

```
result <- t.test(mpg ~ am, data=mtcars)
result$p.value
result$estimate
```

The t-test reveals that the p-value is 0.00137, which suggest rejecting the null hypothesis. Hence, the automatic and manual transmissions are from different populations. As it can be seen, the mean of the two transmissions differs significantly from each other.

4. Regression Analysis

Linear regression models using different variables in order to find the best fit will be applied in this section. They will be compared with the base model using ANOVA. First, let the full model fit as follows:

```
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel)
```

The Adjusted R-squared value is 0.779. Hence, the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level. Using backward selection for choosing some statistically significant variables.

```
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel)
```

This model is “mpg ~ wt + qsec + am”. The value of the Adjusted R-squared raises to 0.8336, i.e. the model can now explain about 83% of the variance for MPG. Additionally, all coefficients are significant at given 0.05 level. According to the scatter plot, it indicates that there appear to be an interaction term between “wt” variable and “am” variable. Thus, the following model shall include the interaction term:

```
amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(amIntWtModel)
```

Again, the model improved, now covering 88 % of the variance (adjusted R-squared =0.8804), by keeping the 0.05 significant level fullfilled for all coefficients. Next, the model will be fitted with MPG as the outcome variable and transmission as the predictor variable.

```
amModel<-lm(mpg ~ am, data=mtcars)
summary(amModel)
```

It shows that on average, a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased. However, the low Adjusted R-squared value of 0.338 indicates that additional variables are required for the model. Finally, the final model is:

```
anova(amModel, stepModel, fullModel, amIntWtModel)
confint(amIntWtModel)
```

Hence, selecting the model with the highest Adjusted R-squared value leads to “mpg ~ wt + qsec + am + wt:am”.

```
summary(amIntWtModel)$coef
```

Hence, if weight and qsec remain constant, cars with manual transmission add 14.079 + (-4.141)*wt

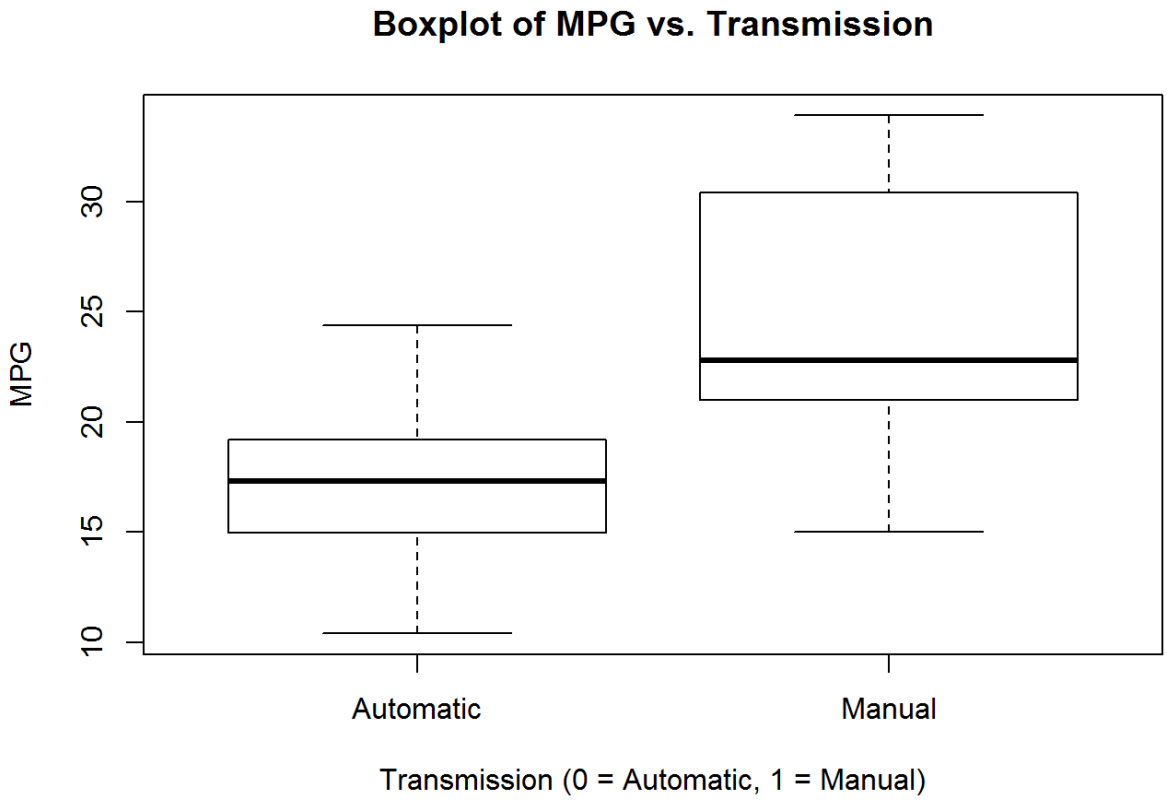
more MPG on average than cars with automatic transmission.

4. Conclusion

According to the residual plots, the following underlying assumptions can be verified: 1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption. 2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line. 3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed. 4. The Residuals vs. Leverage argues that no outliers are present, as all values fall within the 0.5 bands.

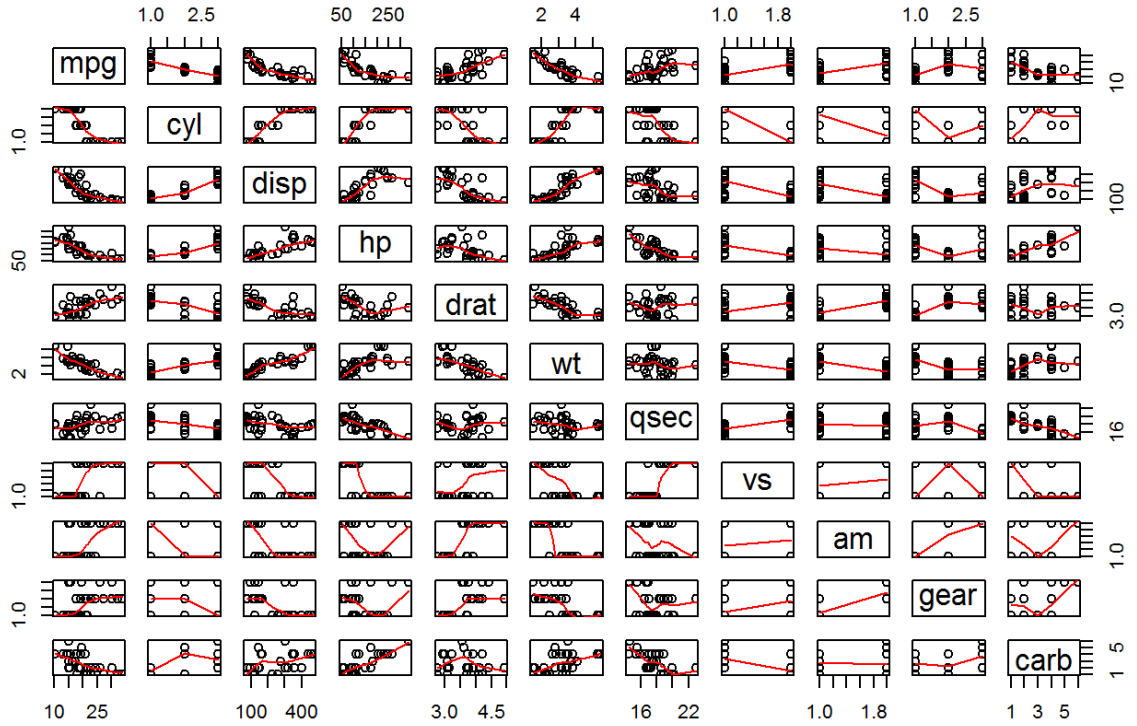
Appendix

```
# Boxplot of MPG vs. Transmission
boxplot(mpg ~ am, data=mtcars,  xlab="Transmission (0 = Automatic, 1 = Manua
l)", ylab="MPG",
        main="Boxplot of MPG vs. Transmission")
```

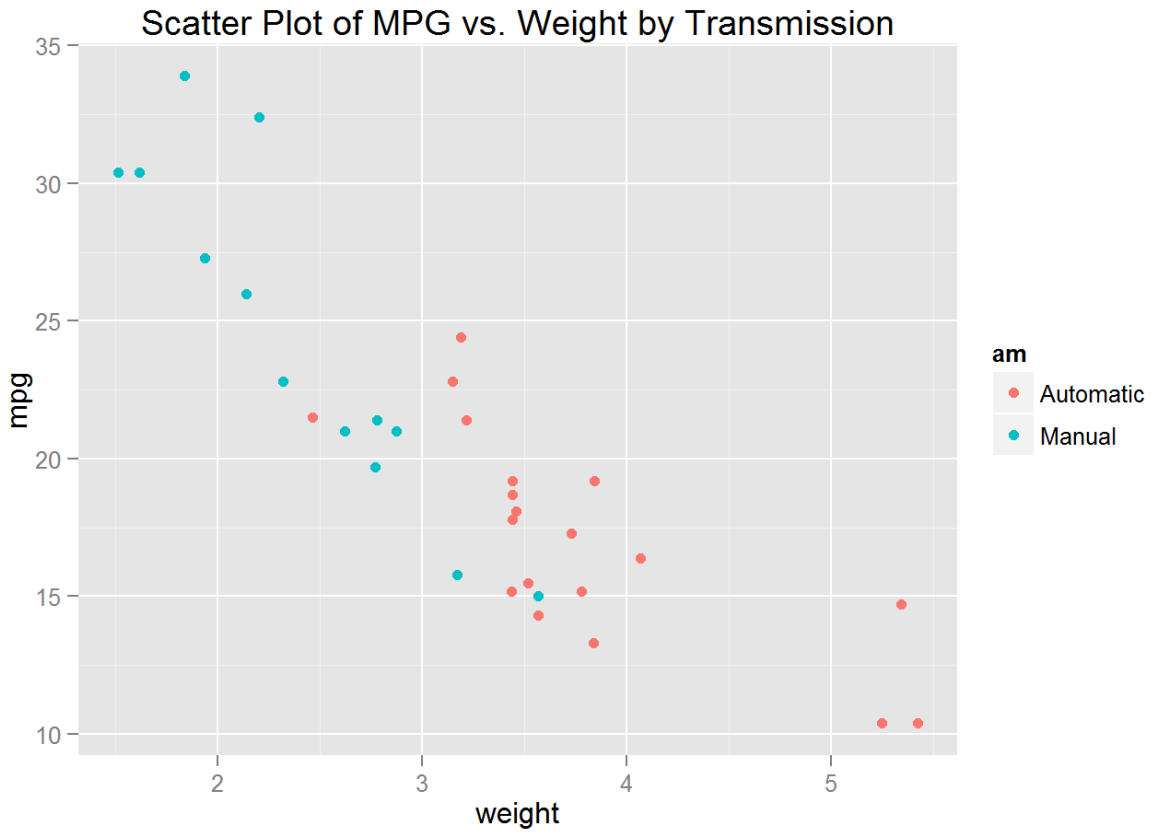



```
# Pair Graph of Motor Trend Car Road Tests
pairs(mtcars, panel=panel.smooth, main="Pair Graph of Motor Trend Car Road T
ests")
```

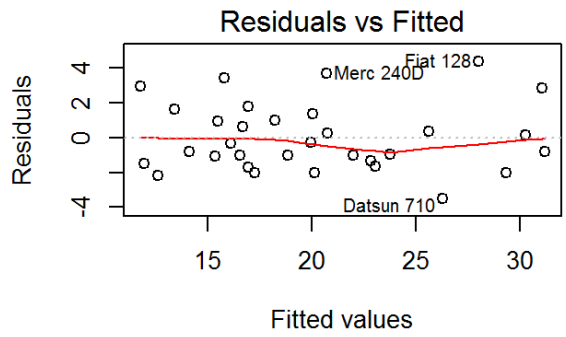
Pair Graph of Motor Trend Car Road Tests

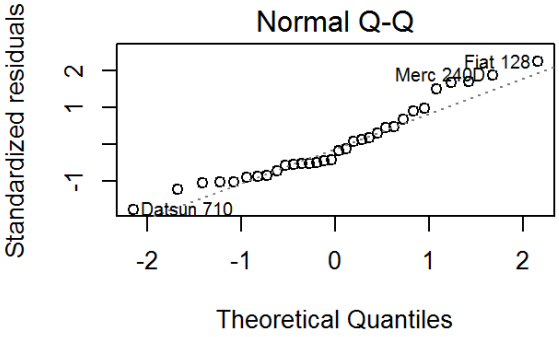


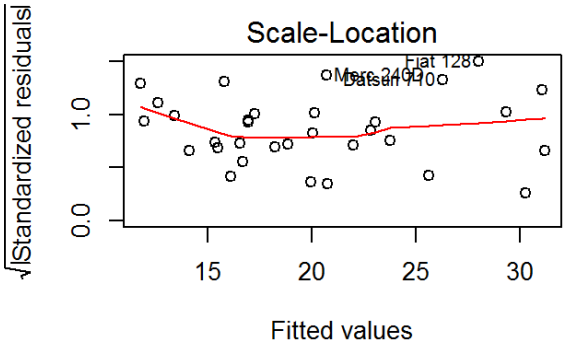

```
# Scatter Plot of MPG vs. Weight by Transmission
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + ge
om_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("Scatter Plot of MPG vs. Weight by Transmission")
```

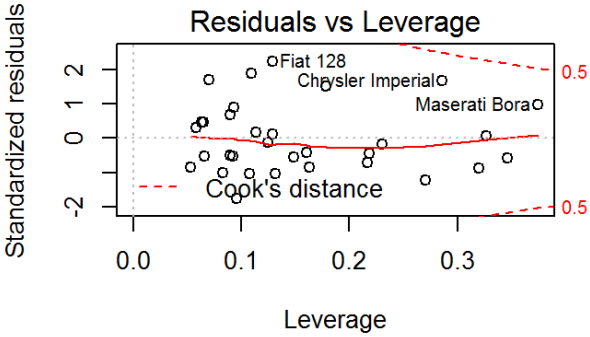



```
# Residual Plots
par(mfrow = c(2, 2))
plot(amIntWtModel)
```


Residuals vs Fitted
Residuals
Fitted values
Merc 240D
Fiat 128
Datsun 710


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710


Scale-Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Datsun 710


Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5