Motor Trend Magazine: Influence on the mileage

Executive Summary

We've been asked to analyze the influence of different automobile design parameters on the mileage. To answer this question we used the **mtcars** data set, which includes the fuel consumption (represented by the attribute **MPG**) and 10 additional design aspects of 32 automobiles from 1973-1974. We concentrate our analysis on two particular topics: *Is an automatic or manual transmission better for **MPG*** and the *Quantification of the MPG difference between automatic and manual transmissions*. In the first stage we performed **exploratory data analysis analysis** including a *t-test*. This way we verified our theory for performance differences between cars with manual and automatic transmition. In the next step we built multiple **linear regression models** to identify further dependencies and correlations between the mileage (used as dependent) and the rest of the attributes.

I Exploratory data analysis

To ensure the correctness of the performed analysis a preprocessing step was required. Further some additional R-libraries need to be loaded:

```
library(ggplot2)
library(gridExtra)
summary(mtcars)
str(mtcars)
colnames(mtcars)
#keep the pristine data untouched
d <- mtcars
d$cyl <- factor(d$cyl,levels = c(4,6,8), labels = c("4cyl","6cyl","8cyl"))
d$vs <- factor(d$vs,levels = c(0,1), labels = c("V-engine","S-engine"))
d$am <- factor(d$am,levels=c(0,1), labels = c("Automatic","Manual"))
d$gear <- factor(d$gear,levels = c(3,4,5), labels = c("3gears","4gears","5gears"))
```

To prove our theory that the transmition type has an influence on the mileage we performed a paired t-test for both options: **manual** and **automatic**. We defined the hypothesis:

H0 = No performance difference (MPG) exists between Automatic and Manual cars

H1 = A performance difference exists between Automatic and Manual cars

As the t-test compares the means of two data sets, we compare the two different subsets of the data splitted by the transmition type. The results obtained were (t = -3.7671), (df = 18.332), (p-value = 0.001374). The P-value of 0.001374 allows us to confidently reject the null hypothesis and accept the alternate hypothesis **H1**.

```
t.test(d[d$am=="Automatic","mpg"],d[d$am=="Manual","mpg"])
```

```
##
##  Welch Two Sample t-test
##
## data:  d[d$am == "Automatic", "mpg"] and d[d$am == "Manual", "mpg"]
## t = -3.7671, df = 18.332, p-value = 0.001374
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

However, on the box plots in **Appendix 1** we can observe that the transmition alone doesn't determine the mileage of a car. The values of the **mpg** attributes differs in the different cases, e.g. considering the number of gears or the number of cylinders of the car.

II Linear model selection

To identify an optimal model that describes good enough the mileage of a car, we used a stepwise linear regression. In our test we add (*forward selection*) or remove (*backward selection*) consequently a feature from the model in terms to find the most significant model with the lowest error. For more information consult the R-function **step()**.

```
lmAll <- lm(mpg~.,d)
#foraward selection
mSelF <- step(lmAll,scope = list( upper=lmAll, lower=~1 ),direction = "forward", trace = F)
#backward elimination
mSelB <- step(lmAll,scope = list( upper=lmAll, lower=~1 ),direction = "backward", trace = F)
#bidirectional elimination
mSelBo <- step(lmAll,scope = list( upper=lmAll, lower=~1 ),direction = "both", trace = F)
summary(mSelF)
summary(mSelB)
summary(mSelBo)
```

Model analysis

After consulting the results of the three model selection methods, we chose the one with the most significant p-value: **backward elimination** p-value = 1.21e-11. To analyse further the results we used the residuals plots in **Appendix 2**. In the **Residual vs. Fitted** plot as well in the **Scale-Location** plot we couldn't identify any patterns, they could be an indication for major error in the model. Further the points are randomly dispersed around the horizontal axis. This fact could be interpreted as: the linear regression model is appropriate for the given data set; otherwise, a non-linear model could be more appropriate. The **QQ Plot** gives us also a good comfort for the *normality of the errors*: the points are building a nearly good diagonal plotted against the *Normal distribution*. The last plot **Residuals vs. Levarage** is also a good indication that we do not have any outliers, that could have a big influence on the model. In total the model we chose via the stepwise regressions seems to be appropriate for this data set and the question we've been asked.

Coefficients interpretation

After we validated the model, we want to give you a deeper understanding of it and give you a good overview of the relationships between the different parameters and the mileage.

The **Intercept** is associated with the automatic transmition and gives us the **MPG** index (9.6 Miles) for every unit increase in **Weight** and **quarter miles time**. The negative coefficient for the slope for **wt** indicates that for each weight unit increase, the **MPD** decreases, i.e. the car becomes 3.92 Miles per Galon less efficient. We interpret the positive coefficient for **qsec** as when the "Quarter Mile Time Index" increases by one unit

(indicator that the velocity is going down), the car becomes more efficient by covering an additional 1.23 Miles per Gallon. The last coefficient is giving the change in slope for manual transmition compared to the automatic. Here we observe a positive value of 2.936, which mean that a car with a manual transmition will cover 3 miles more with the same quantity of fuel.

```
summary(mSelB)$coef #get the coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
```

Conclusion

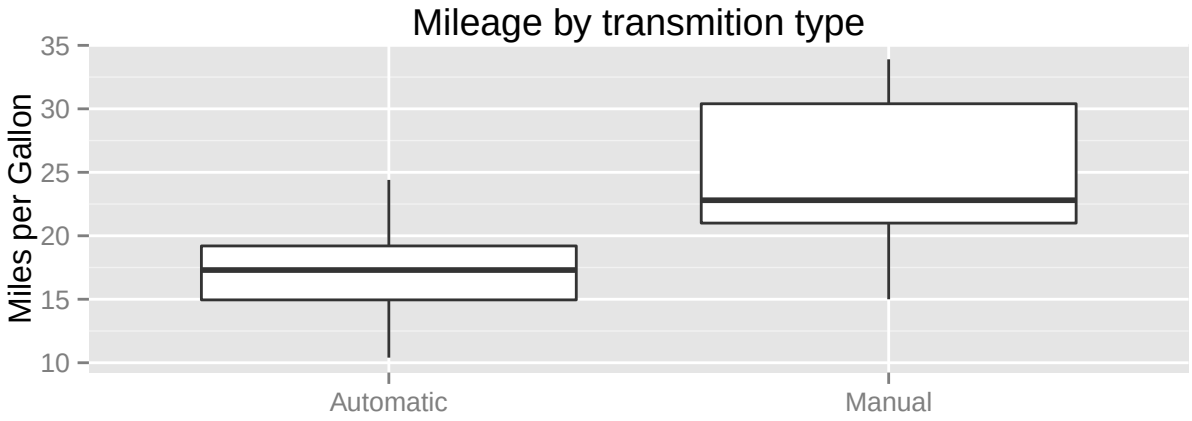
Our statement if a car wit hmanual or automatic transmition is more efficient is, that we can not make that kind of conclusion based only on this attribute. As we've seen in our analysis there multiple additional parameters they need to be considered.

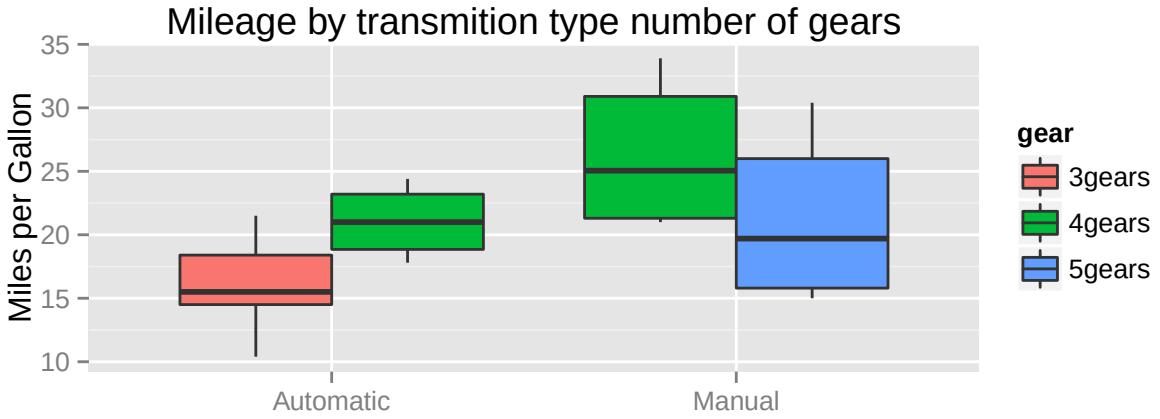
Appendix

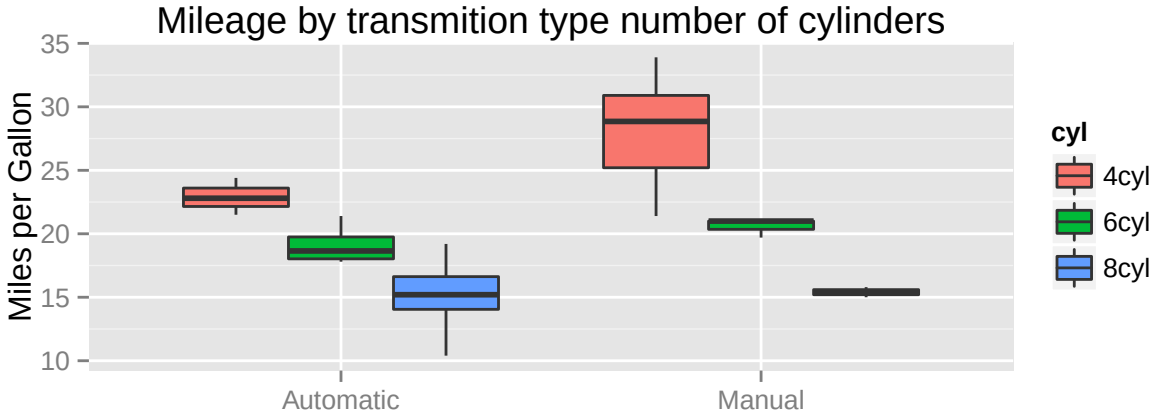
Plots used for the analysis

Appendix 1: Exploratory data analysis

```
b1<-qplot(am,mpg, data=d, geom=c("boxplot"),
    main="Mileage by transmition type",
   xlab="", ylab="Miles per Gallon")

b2<-qplot(am,mpg, data=d, geom=c("boxplot"),
   fill=gear, main="Mileage by transmition type number of gears",
   xlab="", ylab="Miles per Gallon")

b3<-qplot(am,mpg, data=d, geom=c("boxplot"),
   fill=cyl, main="Mileage by transmition type number of cylinders",
   xlab="", ylab="Miles per Gallon")

grid.arrange(b1,b2,b3)
```

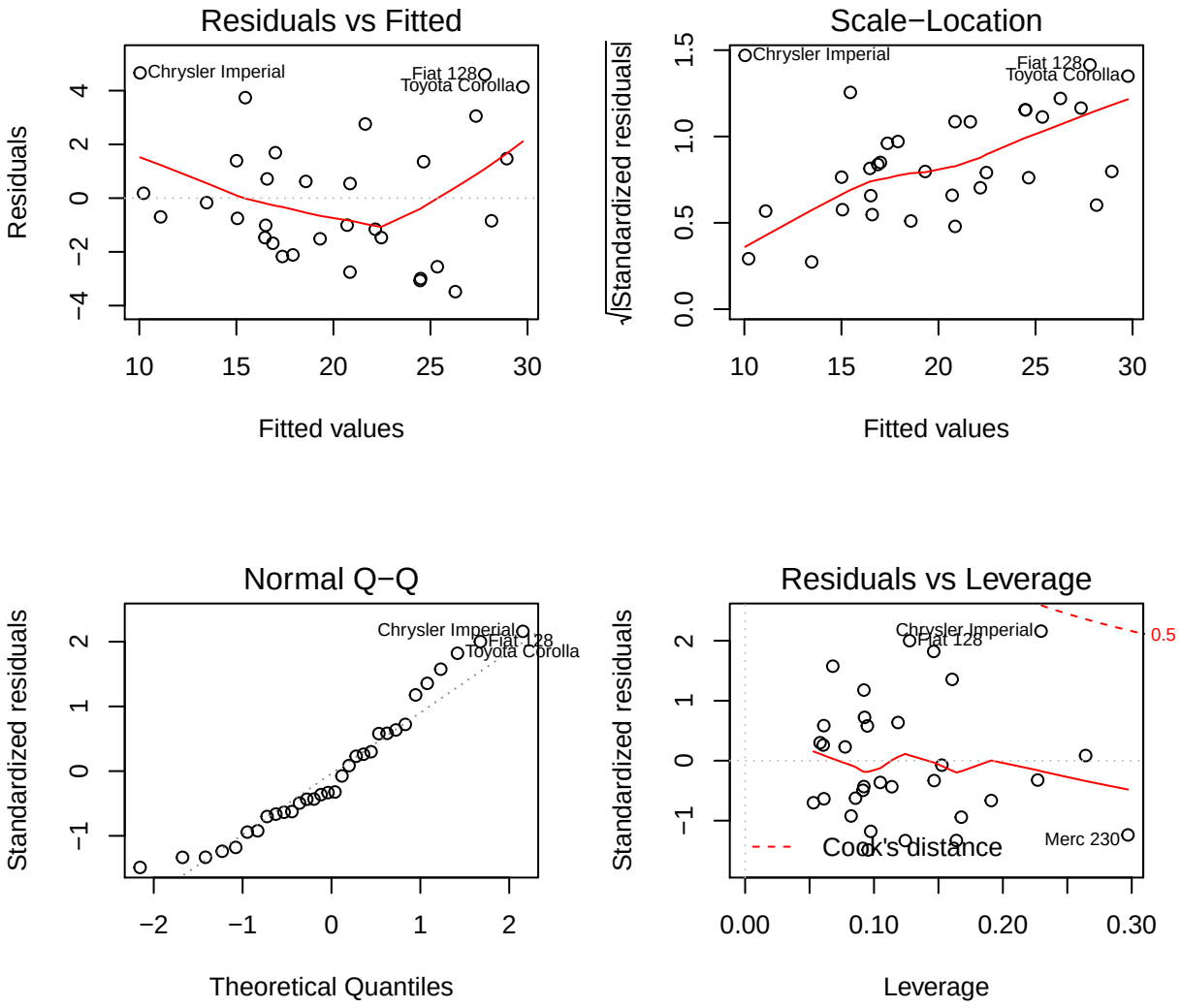






Appendix 2: Residual analysis

```
par(mfcol=c(2,2))
plot(mSelB)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale−Location
√|Standardized residuals|
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Residuals vs Leverage
Leverage
Cook's distance
0.5
Merc 230