Report - Regression Analysis for Final Project in Data Science Course 7 "Regression Models"

Hauer-Glocke

13 Juli 2017

Executive Summary

This report analyses the relationship between miles per gallon (mpg) and automatic/manual transmissions (am). Under consideration of the confounding variables *Displacement (disp)*, *Number of Cylinders (cyl)* and *Weight (wt)*, the dataset *mtcars* shows no significance relation between transmission type and miles per gallon of a car. The observed positive effect of manual transmissions is marginal and the p-value shows insignificance results.

Overview of Report

In this report I examine the relationship between a set of variables and miles per gallon (MPG) (outcome). The analysis is based on the *mtcars*-dataset from the *base* package. For conducting this analysis I used the packages: *dplyr* and *car*. In particular, this analysis addresses the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Exploratory Analysis and Model Selection (Confounding Variables)

The subsequent boxplot, as shown in Appendix A1, suggusts a strong impact of the manual transmission on Miles per Gallon relative to the automatic transmission. The medians of the two subsets with show a differences larger than 5 and the variance in values increase. Also shown in the bigger difference between the 25%- and 75%-Quantile, as indicated by the box sizes.

```
boxplot(mpg ~ am, data=mtcars, main="mtcars - Relation mpg and transmission type",
        xlab="Automatic or Manual Transmissions (am)", ylab="Miles per Gallon (mpg)",
        col=c("green","red"))
```

The following t-test rejects the null hypothesis that there is no effect of transmission type on a 1% significance level. (See P-Value of 0.00137) This also supports the expectation that transmission type is influencing the miles per gallon.

```
test1 <- t.test(mtcars[mtcars$am=="Automatic",1], mtcars[mtcars$am=="Manual",1],
                alternative = "two.sided", paired = FALSE, var.equal = FALSE,
                conf.level = 0.95); test1$p.value
```

```
## [1] 0.001373638
```

The mtcars-dataset contains in total 11 variables, therefore I gonna test if one of the nine remaining variables is suitable to serve as confounding variable. The detailed analysis including regression results of individual models are in Appendix A2.

```r
fit1 <- lm(mpg ~ am, data=mtcars)
fit2 <- lm(formula = mpg ~ ., data=mtcars) #Include all Variables
#The variance inflation factor gives an indication for the best candidates as
#confounding variables as it shows a variables influence on the model's variance.
as.data.frame(vif(fit2))
#Add variables according to their vif to the regression and see when anova
#becomes insignificance.
fit3 <- lm(mpg ~ am + disp, data=mtcars)
fit4 <- lm(mpg ~ am + disp + cyl, data=mtcars)
fit5 <- lm(mpg ~ am + disp + cyl + wt, data=mtcars)
fit6 <- lm(mpg ~ am + disp + cyl + wt + hp, data=mtcars)
fit7 <- lm(mpg ~ am + disp + cyl + wt + hp + carb, data=mtcars)
fit8 <- lm(mpg ~ am + disp + cyl + wt + hp + carb + qsec, data=mtcars)
fit9 <- lm(mpg ~ am + disp + cyl + wt + hp + carb + qsec + gear, data=mtcars)
fit10 <- lm(mpg ~ am + disp + cyl + wt + hp + carb + qsec + gear + vs, data=mtcars)
#cyl, disp and wt are the most interesting candidates. hp, qsec and carb are worth testing.
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9, fit10)
```

The ANOVA-analysis supports the fit5-model with disp, cyl and wt as confounding variables. It has a P-Value of 0.0067, which indicated significance on a 1%-confidence level.

Hypothesis Testing

Based on the preceding Exploratory Analysis and the model selection, we will proceed with the subsequent regression and answer out two research questions based on these results:

```r
#This is only an external example how to shorten the output
fit5 <- lm(mpg ~ am + disp + cyl + wt, data=mtcars)
summary(fit5)$coefficients
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 40.898313414 3.60154037 11.3557837 8.677574e-12
## amManual     0.129065571 1.32151163  0.0976651 9.229196e-01
## disp         0.007403833 0.01208067  0.6128661 5.450930e-01
## cyl         -1.784173258 0.61819218 -2.8861142 7.581533e-03
## wt          -3.583425472 1.18650433 -3.0201537 5.468412e-03
```

Question 1 - Is an automatic or manual transmission better for MPG

After controlling for Displacement (disp), Number of Cylinders (cyl) and Weight (wt), the effect of manual transmission is still positive compared to the automatic transmission. Nonetheless, its p-value is quite high with 0.9229 and therefore, we must conclude that this question cannot be answered based on this dataset.

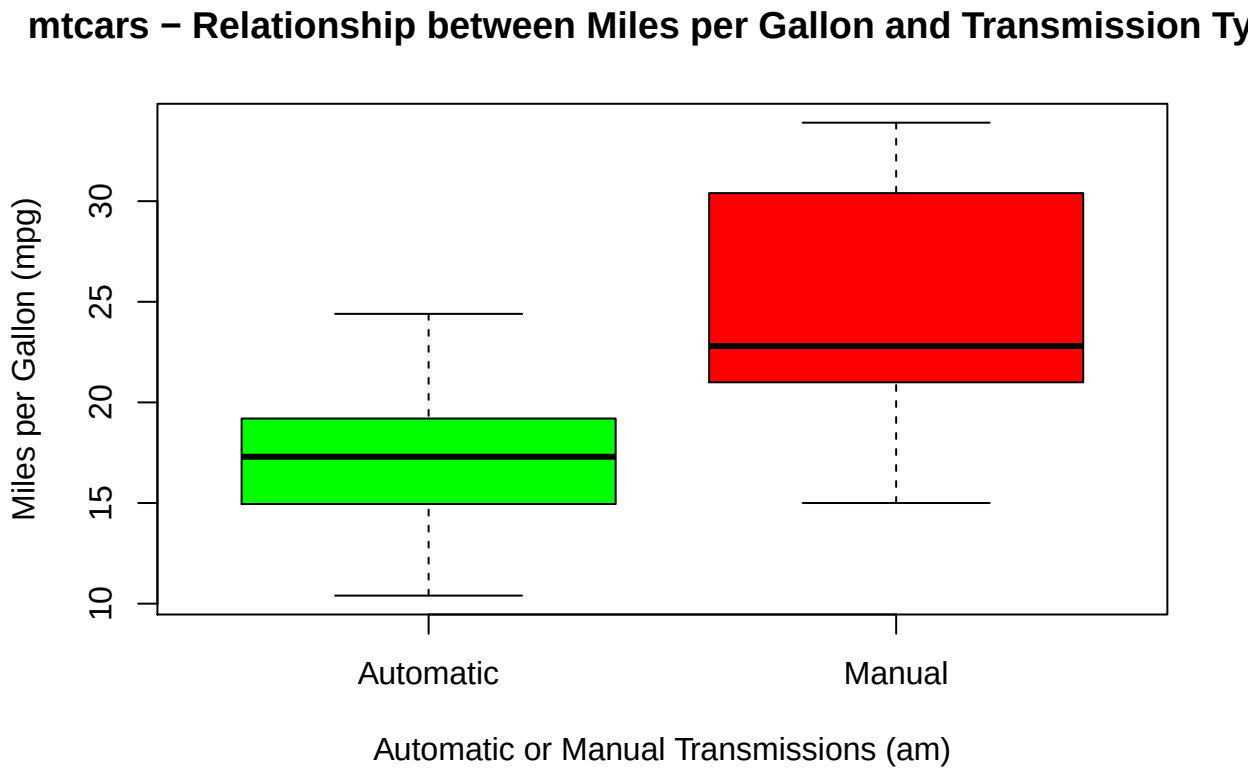
Question 2 - Quantify the MPG difference between automatic and manual transmissions

As stated before, the regression result is insignificance, therefore we cannot quantify the difference. Nonetheless, if we assume the p-value indicates significance for the am-coefficient, than the coefficient of would indicate a 0.1291 Miles per Gallon increase comparing a car with manual transmission to a car with automatic transmission.

Appendix

A1 - Boxplot for the relationsship between Miles per Gallon and Transmission Type

This is part of the Explanatory Analysis. Here you see the Boxplot.



A2 - Model Selection and Confounding Variables

```
fit1 <- lm(mpg ~ am, data=mtcars)
fit2 <- lm(formula = mpg ~ ., data=mtcars)

#The variance inflation factor gives an indication for the best candidates as
#confounding variables as it shows a variables influence on the model's variance.
as.data.frame(vif(fit2))
```

```
##      vif(fit2)
## cyl  15.373833
## disp 21.620241
## hp    9.832037
## drat  3.374620
## wt   15.164887
## qsec  7.527958
```

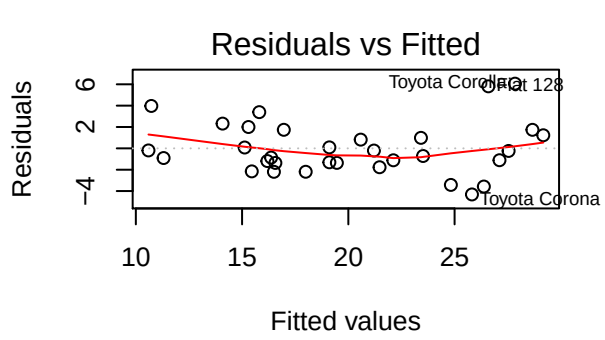
```
## vs    4.965873
## am    4.648487
## gear  5.357452
## carb  7.908747
```

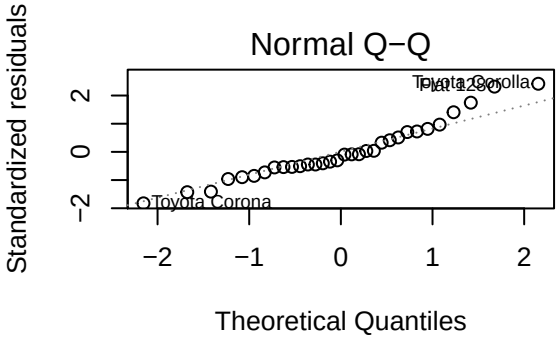
```
##Add variables according to their vif to the regression and see when anova
#becomes insignificance.

fit3 <- lm(mpg ~ am + disp, data=mtcars)
fit4 <- lm(mpg ~ am + disp + cyl, data=mtcars)
fit5 <- lm(mpg ~ am + disp + cyl + wt, data=mtcars)
fit6 <- lm(mpg ~ am + disp + cyl + wt + hp, data=mtcars)
fit7 <- lm(mpg ~ am + disp + cyl + wt + hp + carb, data=mtcars)
fit8 <- lm(mpg ~ am + disp + cyl + wt + hp + carb + qsec, data=mtcars)
fit9 <- lm(mpg ~ am + disp + cyl + wt + hp + carb + qsec + gear, data=mtcars)
fit10 <- lm(mpg ~ am + disp + cyl + wt + hp + carb + qsec + gear + vs, data=mtcars)

#cyl, disp and wt are the most interesting candidates. hp, qsec and carb are worth testing.
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9, fit10)
```

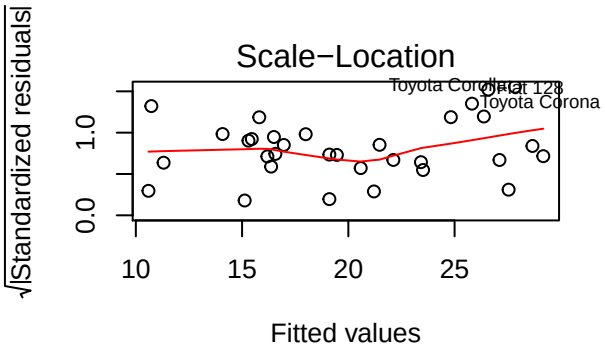
```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model  3: mpg ~ am + disp
## Model  4: mpg ~ am + disp + cyl
## Model  5: mpg ~ am + disp + cyl + wt
## Model  6: mpg ~ am + disp + cyl + wt + hp
## Model  7: mpg ~ am + disp + cyl + wt + hp + carb
## Model  8: mpg ~ am + disp + cyl + wt + hp + carb + qsec
## Model  9: mpg ~ am + disp + cyl + wt + hp + carb + qsec + gear
## Model 10: mpg ~ am + disp + cyl + wt + hp + carb + qsec + gear + vs
##    Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1      30 720.90
## 2      21 147.49  9    573.40 9.0711 1.779e-05 ***
## 3      29 300.28 -8   -152.79 2.7192   0.03157 *
## 4      28 252.08  1     48.20 6.8627   0.01601 *
## 5      27 188.43  1     63.66 9.0631   0.00666 **
## 6      26 163.12  1     25.31 3.6030   0.07151 .
## 7      25 161.44  1      1.68 0.2394   0.62973
## 8      24 150.96  1     10.48 1.4917   0.23549
## 9      23 149.31  1      1.65 0.2347   0.63310
## 10     22 149.12  1      0.19 0.0274   0.87022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

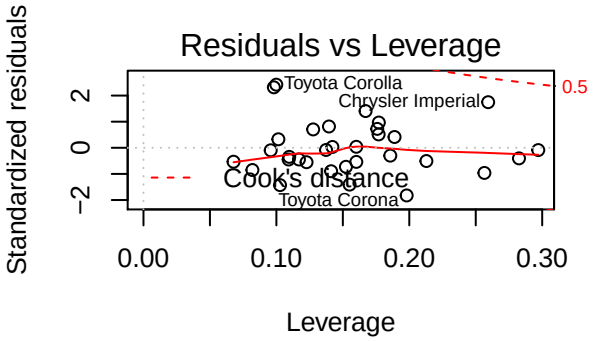
Residual Plot and Diagnostics

```
par(mfrow = c(2,2))
plot(fit5)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Toyota Corona


Scale−Location
√|Standardized residuals|
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona


Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5