Course Project: Regression models

Koen Vermeulen

10-9-2019

Management summary

In this project we will explore some features of cars that affect fuel consumption in miles per gallon (MPG), transmission type in particular (automatic vs. manual). We are looking a dataset of a collection of cars (mtcars - Motor Trend Car Road Tests), and are interested in exploring the relationship between a set of variables and Miles Per Gallon (MPG). In particularly we want answer two questions: 1. Is an automatic or manual transmission better for MPG? 2. What is the difference in MPG between cars with an automatic and manual transmission?

Loading & preparating data

Load the dataset and convert categorical variables to factors.

```
library(ggplot2)
data(mtcars) # loading the data
head(mtcars, n=5) # first look at the data
dim(mtcars) # dimensions are 32 records by 11 variables
mtcars$cyl <- as.factor(mtcars$cyl); mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am); mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
levels(mtcars$am) <- c("Auto", "Manual") # for easier analysis
```

Exploratory data analysis

To answer our questions we are initially interested in the relation between the two parameters: transmission (am) and Miles per Gallon (mpg). **Appendix 1** is a boxplot that compares automatic and manual transmission MPG. The graph shows that the miles per gallon is larger for manual than for automatic. But it still is possible that the other variables play a role in the determination of MPG.

Appendix 2 is a pairs graph that shows that MPG has correlations with other variables than just **am**. To obtain a more accurate model, we need to predict MPG with more variables than **am** alone.

Statistical inference

```
ttest_MPG_AM <- t.test(mtcars$mpg ~ mtcars$am)
ttest_MPG_AM$p.value
```

The results rejects the null hypothesis that the difference between transmission types is 0.

Regression analysis

The **first model** (below) explains the MPG variability by the transmission type (**am**) alone.

```
fit_MPG_AM <- lm(mpg ~ am, mtcars)
summary(fit_MPG_AM)$coef
summary(fit_MPG_AM)$r.squared
```

Eventhough the p-value is low (0.000285), the R-Squared is low as well (0.3385). Therefore before making any conclusions on the effect of transmission type on fuel efficiency, we have to take a look at models inlcuding the other variables in the dataset.

The **second model** fits all variables on MPG.

```
fit_MPG_all <- lm(mpg ~ ., mtcars)
summary(fit_MPG_all)
```

Now, eventhough the R-squared has increased (0.8931), none of the coefficients is significant at the 0.05 level. We have to find the optimal model somewhere in the middle.

The **step-function** can be used to do the variable selection for the optimal model.

```
fit_step <- step(fit_MPG_all,direction="both",trace=FALSE)
summary(fit_step)
```

The new model has 4 variables (cylinders, horsepower, weight, transmission). The R-squared value of 0.8659 confirms that this model explains about 87% of the variance in MPG. The p-values also are statistically significantly because they have a p-value less than 0.05. The coefficients conclude that increasing the number of cylinders from 4 to 6 with decrease the MPG by 3.03. Further increasing the cylinders to 8 with decrease the MPG by 2.16. Increasing the horsepower is decreases MPG 3.21 for every 100 horsepower. Weight decreases the MPG by 2.5 for each 1000 lbs increase. A Manual transmission improves the MPG by 1.81.

Residual Analisys

Appendix 3 shows the residual plots of the last model (fit_best). These show: 1. The randomness of the Residuals vs. Fitted plot, which supports the assumption of independence 2. The points of the Normal Q-Q plot following closely to the line, which concludes that the distribution of residuals is normal 3. The Scale-Location plot random distribution, which confirms the constant variance assumption 4. Since all points are within the 0.05 lines, the Residuals vs. Leverage, which concludes that there are no outliers

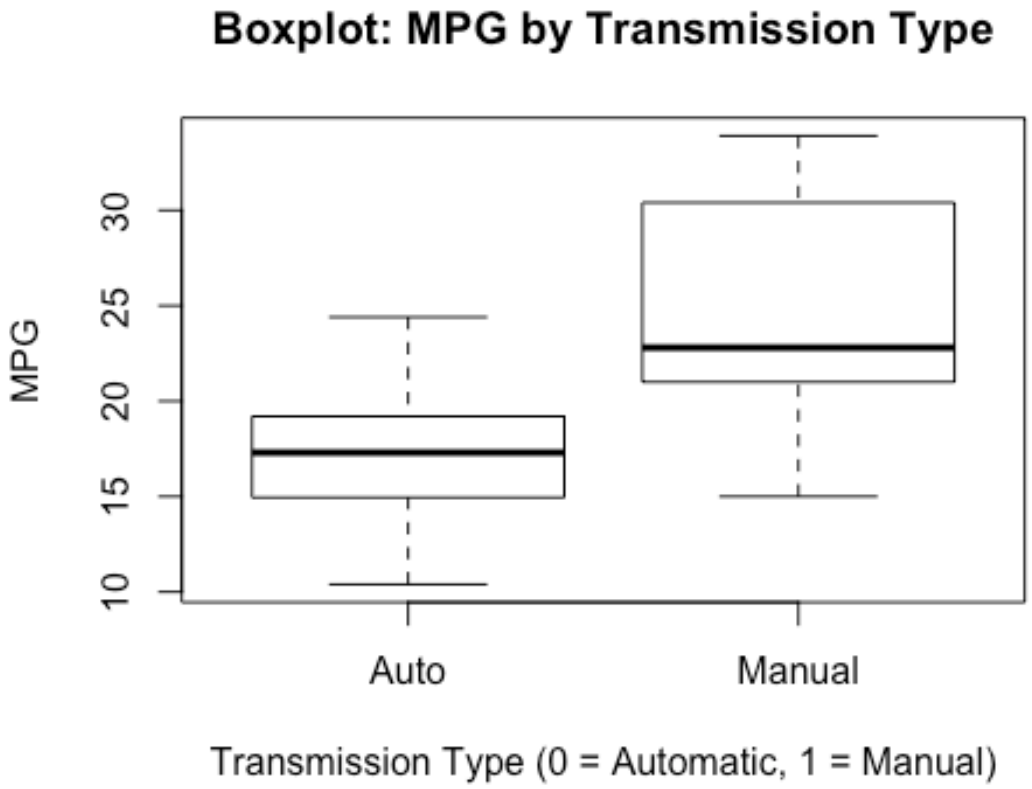
Conclusion

There is a difference in MPG based on transmission type. A manual transmission will have a slight MPG boost. However, it seems that weight, horsepower, & number of cylinders are more statistically significant when determining MPG. In the step-model the manual transmission improves the MPG by 1.81.

Appendices

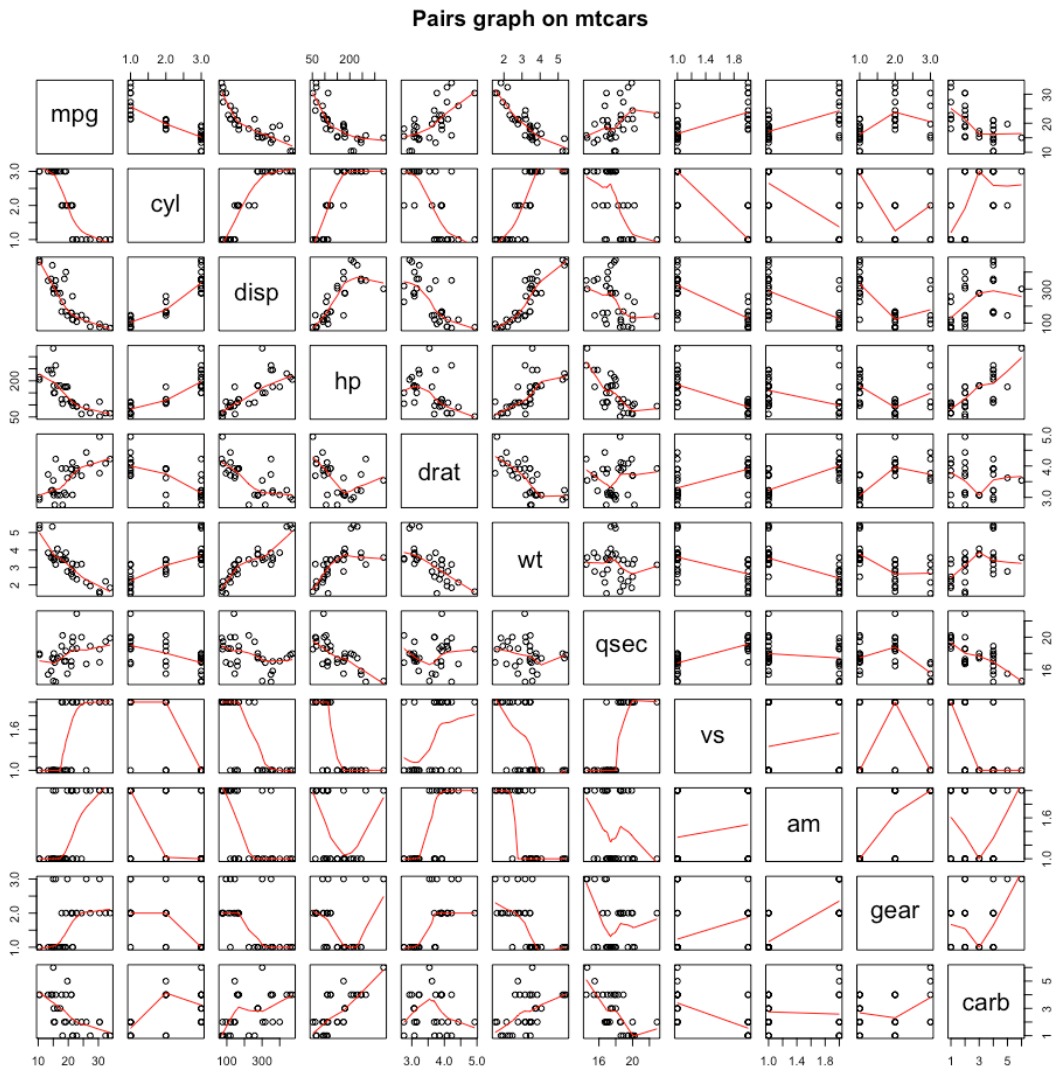
Appendix 1

```
  boxplot(mtcars$mpg ~ mtcars$am,
          xlab="Transmission Type (0 = Automatic, 1 = Manual)",
          ylab="MPG",
          main="Boxplot: MPG by Transmission Type")
```



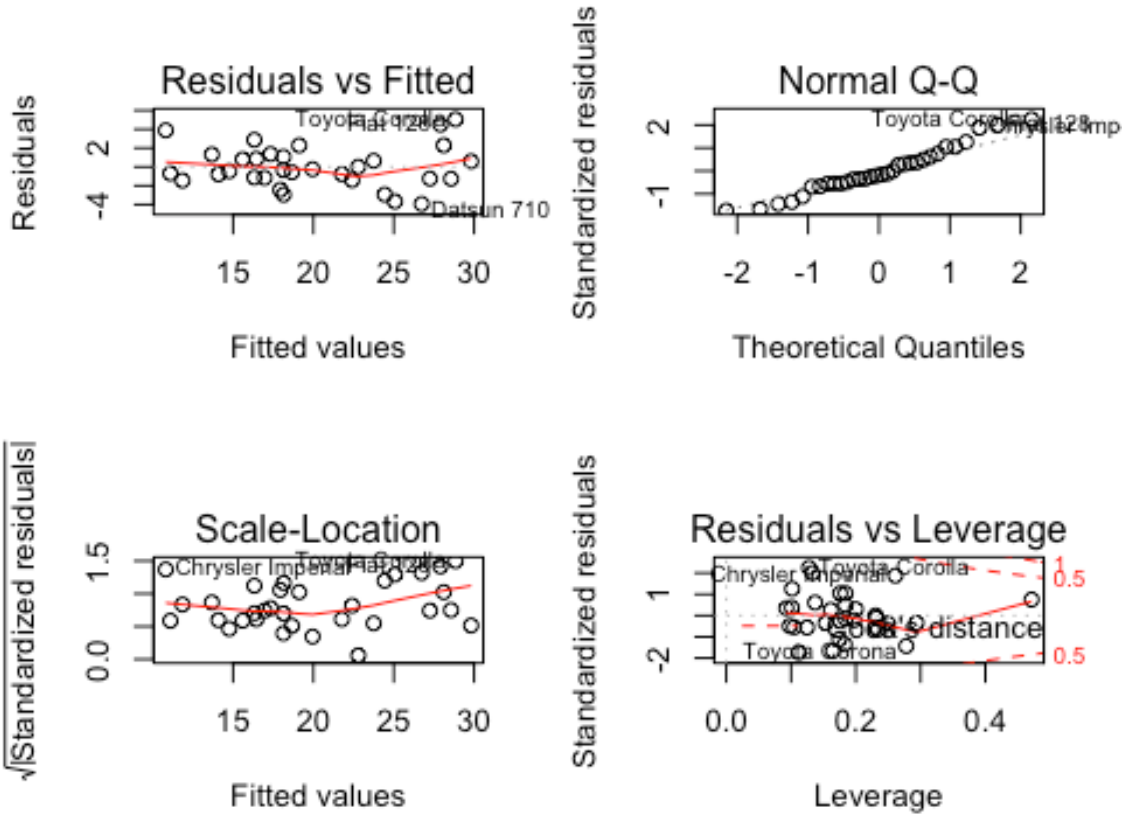
Appendix 2

```
pairs(mtcars, panel = panel.smooth, main = "Pairs graph on mtcars")
```



Appendix 3

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


Residuals vs Fitted
Residuals
Fitted values
Datsun 710
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corona