Regression Models Course Project

SN de Koning

23 maart 2016

Summary

The following report will detail the analysis performed for the Regression Models class for the Data Science Specialization at Coursera and the John Hopkins Institute of Public Health. Several analysis were performed on the base R data-set `mtcars` to answer the following to questions:

1. "Is an automatic or manual transmission better for Miles per Gallon (mpg)"
2. "Quantify the MPG difference between automatic and manual transmissions"

Data Loading and Transformation.

For the code see the appendix.

Data Exploration

The data set consists of 32 observations and 11 variables. A visual inspection of the box-plot of mpg suggests that a manual transmission gives an increase in mpg when compared to an automatic transmission (see figure I). To test if this difference is statistically significant, a students t-test was performed, after the confirmation of the assumption of normality by way of the Shapiro-Wilk test.

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

p > 0.05, so mpg was assumed to follow the normal distribution.

From the t.test we can see that $t(18) = -3.77$, $p < .001$, so we reject the null hypothesis in favor of the alternative hypothesis. From this was concluded that manual transmission has a better mpg than Automatic Transmission.

Regression

First a simple linear regression model was fitted between MPG and transmission type to determine the average change in mpg when switching between transmission types. This model disregards any other variables that might have any influence on the amount of change.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

This model suggests that changing from automatic transmission to manual transmission increases mpg by 7.245. However a visual inspection of the pairs plot (figure II), it seems that there are variables that are highly correlated with each mpg. To create a better fit, the `step` method was used to create a better model with significant predictors

This method suggests that cyl, hp and wt as confounding variables and am as the independent variable.

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

The Adjusted R-squared is 0.8401 according to this summary, so 84.01% of variance can be explained by this model. With ANOVA can been seen that this final model is better ($p < 0.001$) than the model with just the am variable.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics

From the plot (figure III) the following observations can be made:

- In the Residuals vs Fitted plot there is no discernible pattern, so it can be assumed the variables are independent of each other.
- The points in the Normal Q-Q plot fall on a line, therefore it seems the residuals are normally distributed.
- From the Scale-Location plot there is no suggestion of heteroscedasticity, meaning the variance is constant.
- In the Residuals vs Leverage plot there seems to be an outlier with higher leverage on the regression model, which could warrant further investigation.

Conclusions

1. Manual transmission yield a higher mpg than automatic transmission.
2. Disregarding other variables, switching from automatic to manual would yield an increase of 7.25 mpg.
3. A better model includes the amount of cylinders, horsepower and weight, and means that a switch from automatic to manual transmission yields only an increase of 1.80 mpg.

Appendix

code

```
# Setting seed and loading data & dependencies
set.seed(23072016)
data("mtcars")
library(knitr)
# Transforming variables to factor, giving proper labels.
cols <- c("cyl", "vs", "gear", "carb")
mtcars[, cols] <- lapply(mtcars[, cols], as.factor)
mtcars$am <- factor(mtcars$am, levels = c("0", "1"), labels = c("Automatic", "Manual"))
# Testing for the assumption of normality
normal <- shapiro.test(mtcars$mpg)
# Studen t.test for difference in means.
result <- t.test(mpg ~ am, data = mtcars)
# Fitting a linear model between mpg as the dependent, and am as the independent variable
fit <- lm(mpg ~ am, data = mtcars)
summary(fit)$coefficients
# Fitting a linear model between mpg as the dependent variable against all the other variables
fit.all <- lm(mpg ~ ., data = mtcars)
fit.final <- step(fit.all, direction = "both") # Using step method to determine which variables to inclu
summary(fit.final)$coefficients
# Testing the simple lm against the new multivariate lm
anova(fit, fit.final)
# Boxplot fig I
boxplot(mpg ~ am, data = mtcars, main = "Figure I")
#Pairsplot fig II
pairs(mtcars,
      main = "Figure II",
      panel = function(x, y){
          points(x, y)
          abline(lm(y~x), col = "red")})
# Residuals and Diagnostics plot fig III
par(mfrow = c(2,2))
plot(fit.final, main = "Figure III")
```

Figure I

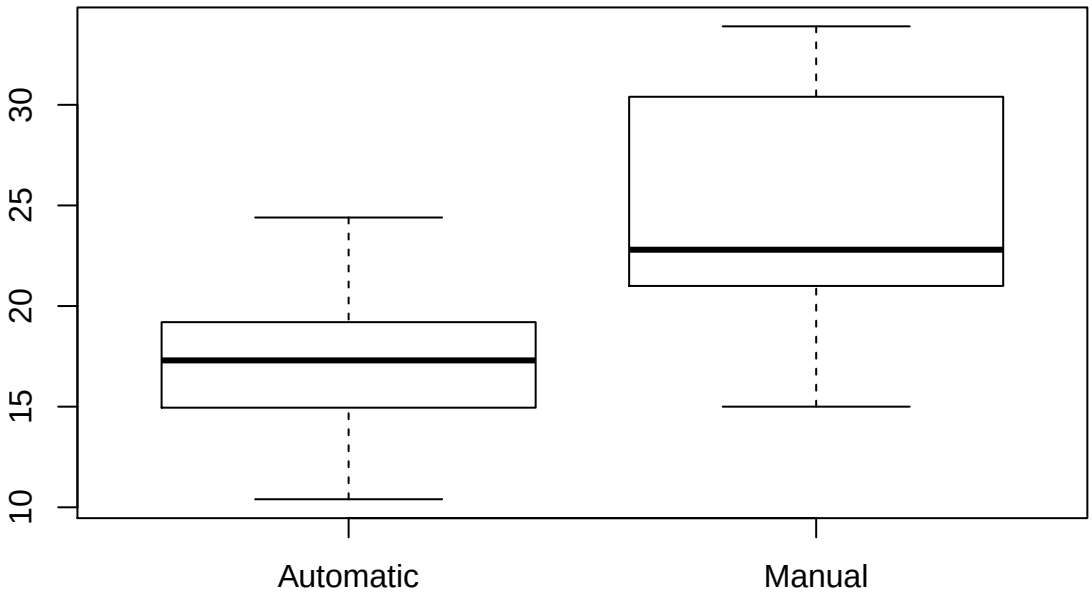
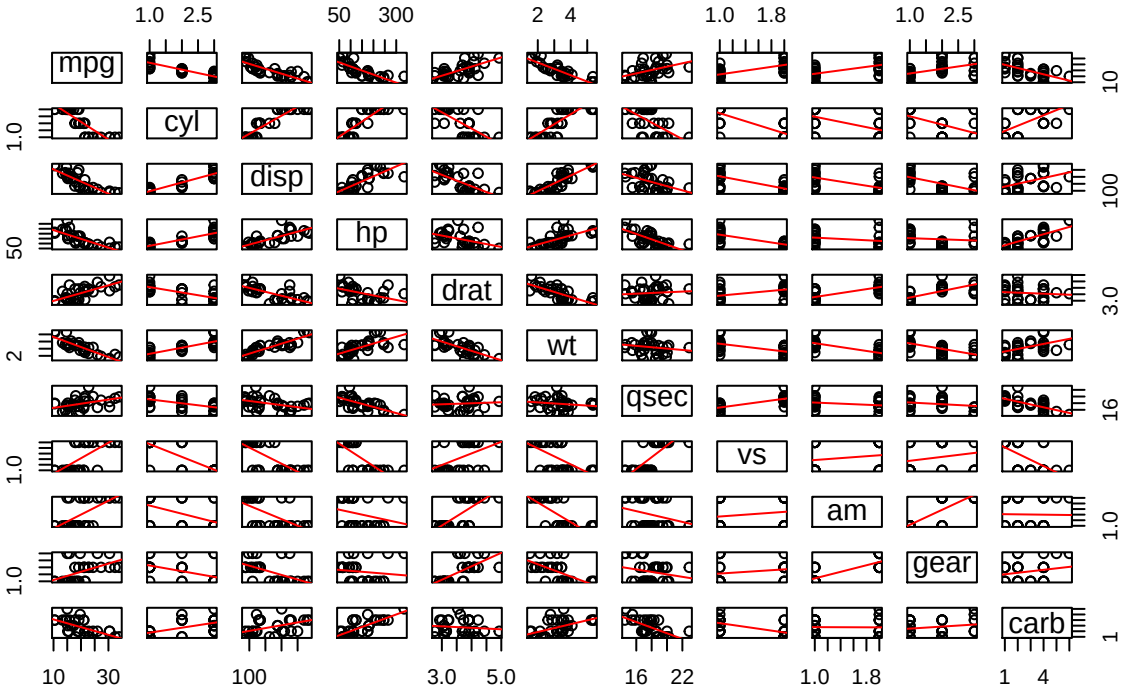


Figure II



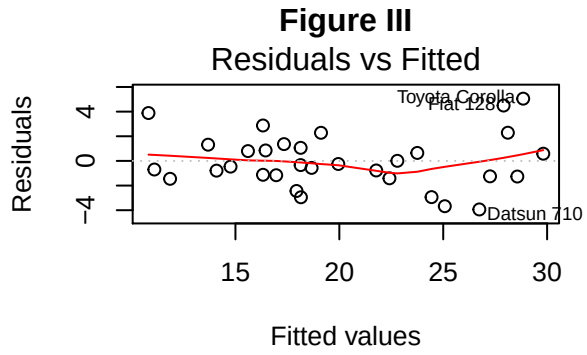


Figure III
Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710

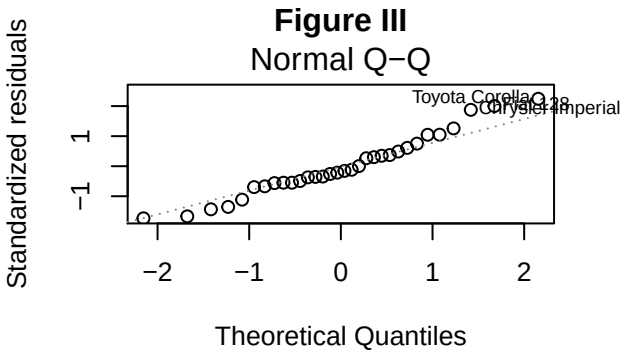


Figure III
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial

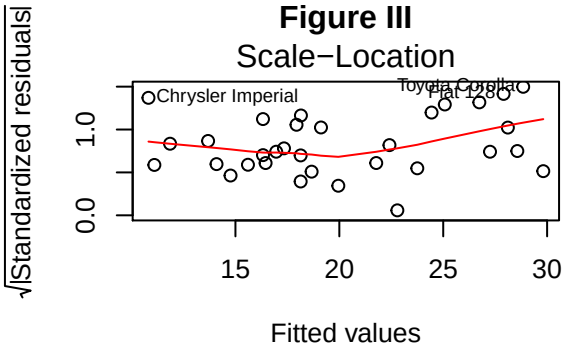


Figure III
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128

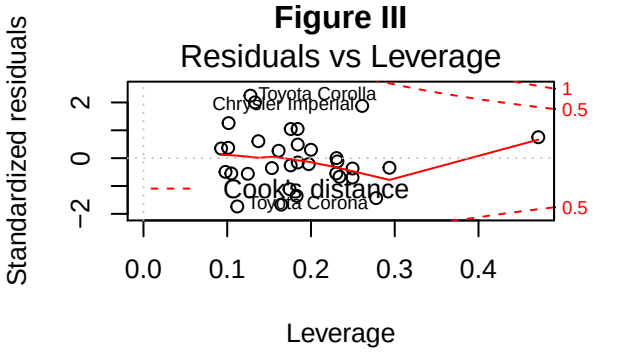


Figure III
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corolla
Chrysler Imperial
Toyota Corona