Regression Models Course Project

Authors: Jianlei Sun, November 6 2016

Executive Summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

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

Conclusions:

- The type of transmission is significant for the fuel consumption (mpg), where manual transmission had on average 7.24 mpg higher consumption than automatic. The manual transmission contributes 1.81 times higher consumption, with assumption that all other variables are equal to zero.
- In addition, it is observed that the transmission type was not so significant as other variables such as weight, horsepower and number of cylinders.

Exploratory Data Analysis

Load the data and rename the variables:

```
data(mtcars)
names(mtcars) <- c("mpg", "cylinders", "displacement", "horsepower", "axleratio", "we
ight",
"qmiletime","vs","transmission", "gears","carburetors")
```

Convert the data to factor variables:

```
mtcars$transmission <- factor(mtcars$transmission)
mtcars$cylinders <- factor(mtcars$cylinders)
mtcars$gears <- factor(mtcars$gears)
mtcars$carburetors <- factor(mtcars$carburetors)
mtcars$vs <- factor(mtcars$vs)
levels(mtcars$transmission) <- c("automatic", "manual")
```

Show the data summary:

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg         : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cylinders   : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ displacement: num  160 160 108 258 360 ...
##  $ horsepower  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ axleratio   : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ weight      : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qmiletime   : num  16.5 17 18.6 19.4 17 ...
##  $ vs          : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ transmission: Factor w/ 2 levels "automatic","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gears       : Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carburetors : Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Test fuel consumption (mpg) between transimission types:

```
t.test(mtcars$mpg ~ mtcars$transmission)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  mtcars$mpg by mtcars$transmission
## 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 in group automatic    mean in group manual 
##                17.14737                24.39231
```

The p-value is less than 0.05, which means the difference is significant. The mean mpg difference between automatic and manual transmissions is (24.39-17.15) = 7.24 mpg

Figure 1 (see Appendix) shows a boxplot of mpg vs. transmission type.

Model Regression

1. Model selection from stewise regession option:

During the model selection, it is important to determine which variables have greatest impact on fule consumption and thus should be included. We will use the “backward stepwise regression” option, which starts with all predictors, and then removes those that are not statistically signifcant.

```
fullModel <- lm(mpg ~ ., data = mtcars)
bestmodel <- step(fullModel, direction="backward", k=2, trace=0)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + horsepower + weight + transmission,
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        33.70832    2.60489  12.940 7.73e-13 ***
## cylinders6         -3.03134    1.40728  -2.154  0.04068 *
## cylinders8         -2.16368    2.28425  -0.947  0.35225
## horsepower         -0.03211    0.01369  -2.345  0.02693 *
## weight             -2.49683    0.88559  -2.819  0.00908 **
## transmissionmanual  1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

- the p-value equal to 1.506E-10 is less than 0.05, indicating that the model might be significant;
- the adjusted R-squared value equal to 0.8401 means that the model explains 84.01% of variance;
- it is observed that the “weight”, “horsepower” and “cylinder” are more significant than the transmission type.

2. test the significance of the final model:

```
initialModel<- lm(mpg ~ transmission, data = mtcars)
finalModel <- lm(mpg ~ cylinders + horsepower + weight + transmission, data = mtcars)
anova(initialModel, finalModel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ transmission
## Model 2: mpg ~ cylinders + horsepower + weight + transmission
##   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
```

The above p-value is much lower than 0.05, which indicates the final selected model is statisticaly significant.

3. Residual Analysis

Figure 2 (see appendix) shows the plot of residuals vs. fitted are randomly scattered without an obvious pattern. The Q-Q plot show most of points follow the line trend, indicating that residuals are normaly distributed.

Appendix

Figure 1:

```r
library(ggplot2)
g <- ggplot(mtcars, aes(transmission, mpg))
g <- g + geom_boxplot(aes(fill = transmission))
g <- g + labs(title = "The boxplot of mpg vs. transmission types", x = "Transmission
Types",
              y = "The mpg values")
g
```

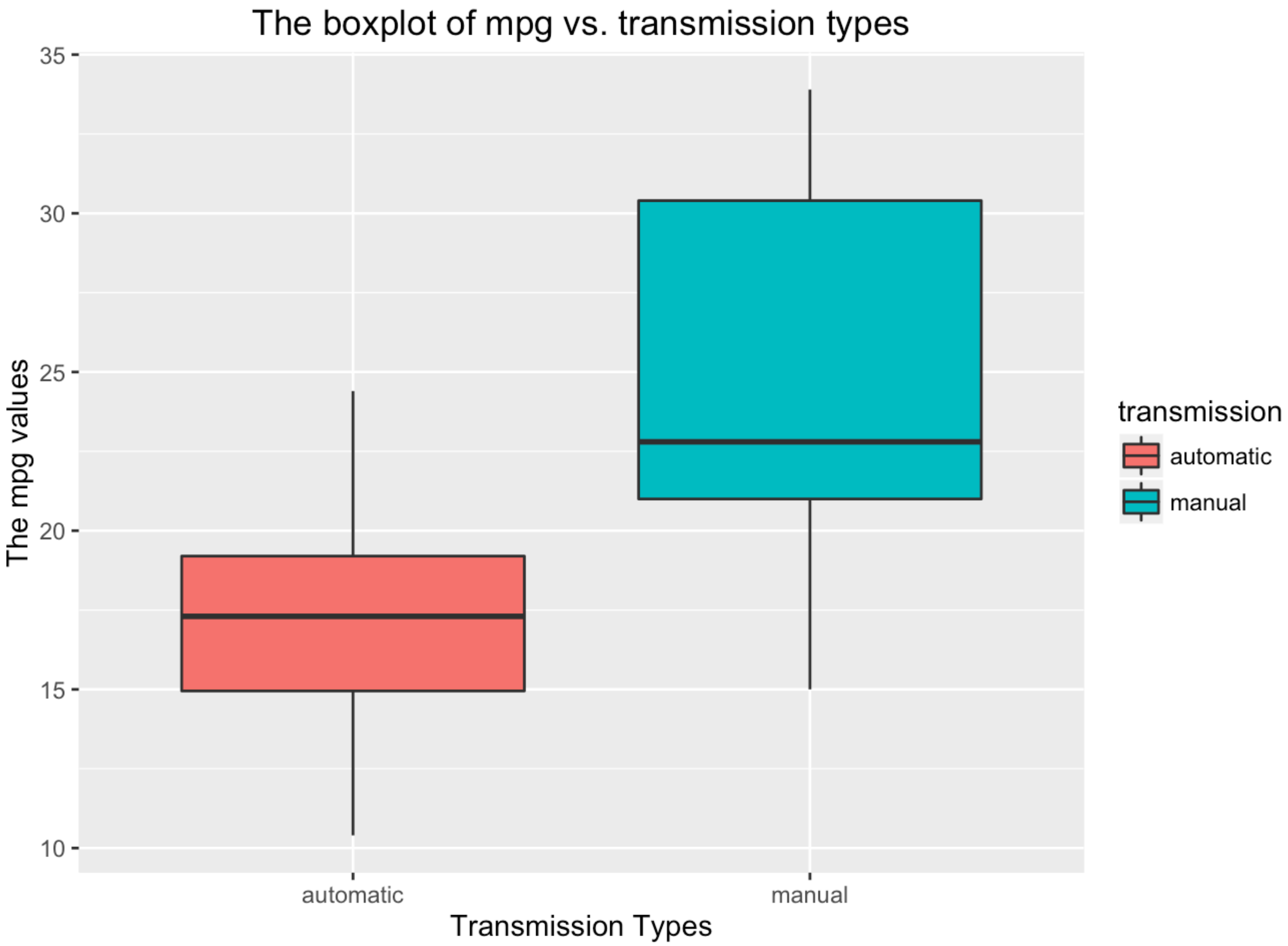
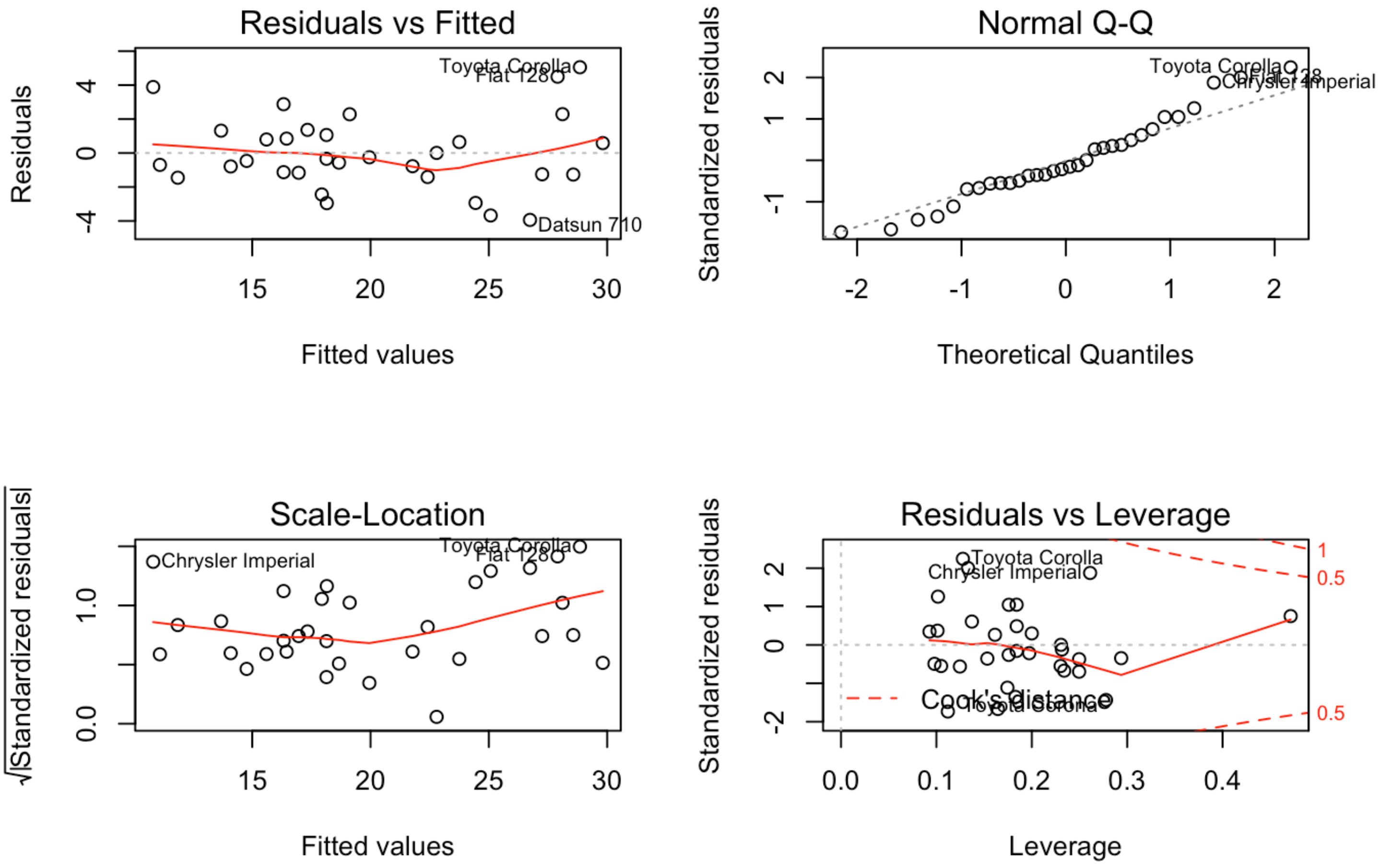


Figure 2:

```r
par(mfrow=c(2, 2))
plot(finalModel)
```


Residuals vs Fitted
Residuals
4
0
-4
Fitted values
15
20
25
30
Toyota Corolla
Fiat 128
Datsun 710
Normal Q-Q
Standardized residuals
2
1
-1
Theoretical Quantiles
-2
-1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale-Location
√|Standardized residuals|
1.0
0.0
Fitted values
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
2
1
0
-2
Leverage
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5