Regression Models Course Project

Dmitri Perov

January 28, 2016

Executive Summary

In this project we explored the relationship between a set of variables and miles per gallon (MPG) in the mtcars dataset and answered the following questions:

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

First, we made some exploratory data analyses and fit simple linear model.

Simplest model explains only 34% of total data variability and then can not be considered as good.

Then more complex models had been tested. It had been found that other variables, especially car weight and qsec parameter (1/4 mile time), have a significant impact to MPG.

This model showed that if we compare cars with similar weight and performance (qsec parameter), then a car with manual transmission has in average 2.8 better mpg (with 95% confidence interval from 0.04 to 5.83 mpg)

The conclusion can be made that manual transmission is better for MPG.

Analisys

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

First investigate relationship of MPG and transmission type disregarding all other variables.

```
auto <- mtcars[mtcars$am==0,]$mpg
manual <- mtcars[mtcars$am==1,]$mpg
summary(auto)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
summary(manual)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

```r
test <- t.test(auto, manual)
test$conf.int
```

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

Result shows that there is a significant impact (pvalue: 0.0013736) of the transmission type to MPG.

Evaluate a simple linear model that considers transmission type only.

```r
fit0 <-lm(mpg~factor(am),mtcars)
summary(fit0)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)1    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```r
confint(fit0)
```

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## factor(am)1  3.64151 10.84837
```

This simple model explains only 34% (Adjusted R-squared value) of the total variation and can not be considered as a good model.

Let's try to select set of other variables from the dataset that have significant impact to MPG.

Fit multivariate model with all the dataset variables as a factor and examine the model coefficient. More influential variable has a larger absolute value of the slope coefficient.

```r
fit1 <- lm(mpg~., data = mtcars)
fit1$coefficient[order(abs(fit1$coefficient), decreasing = TRUE)]
```

```
## (Intercept)          wt          am        qsec        drat        gear
## 12.30337416 -3.71530393  2.52022689  0.82104075  0.78711097  0.65541302
##          vs        carb         cyl          hp        disp
##  0.31776281 -0.19941925 -0.11144048 -0.02148212  0.01333524
```

Here we see that variable wt and qsec has comparable impact to MPG. Build another model including these variables as factors.

```
fit3 <- lm(mpg ~ factor(am) + wt + qsec, data = mtcars)
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
confint(fit3)
```

```
##                    2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## factor(am)1  0.04573031  5.825944
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
```

This model explains 85% of data variability (Adjusted R-squared value) and can be considered as a good one. The interpretation of model coefficient may be the follows:

If we compare cars with similar weight and performance (qsec param), then a car with manual transmission has in average 2.8 better mpg (with 95% confidence interval from 0.04 to 5.83 mpg)

Figure3 shows the diagnostic plots of the model. There are no signs of patterns or Heteroskedasticity in the data.

Appendix

Figure 1

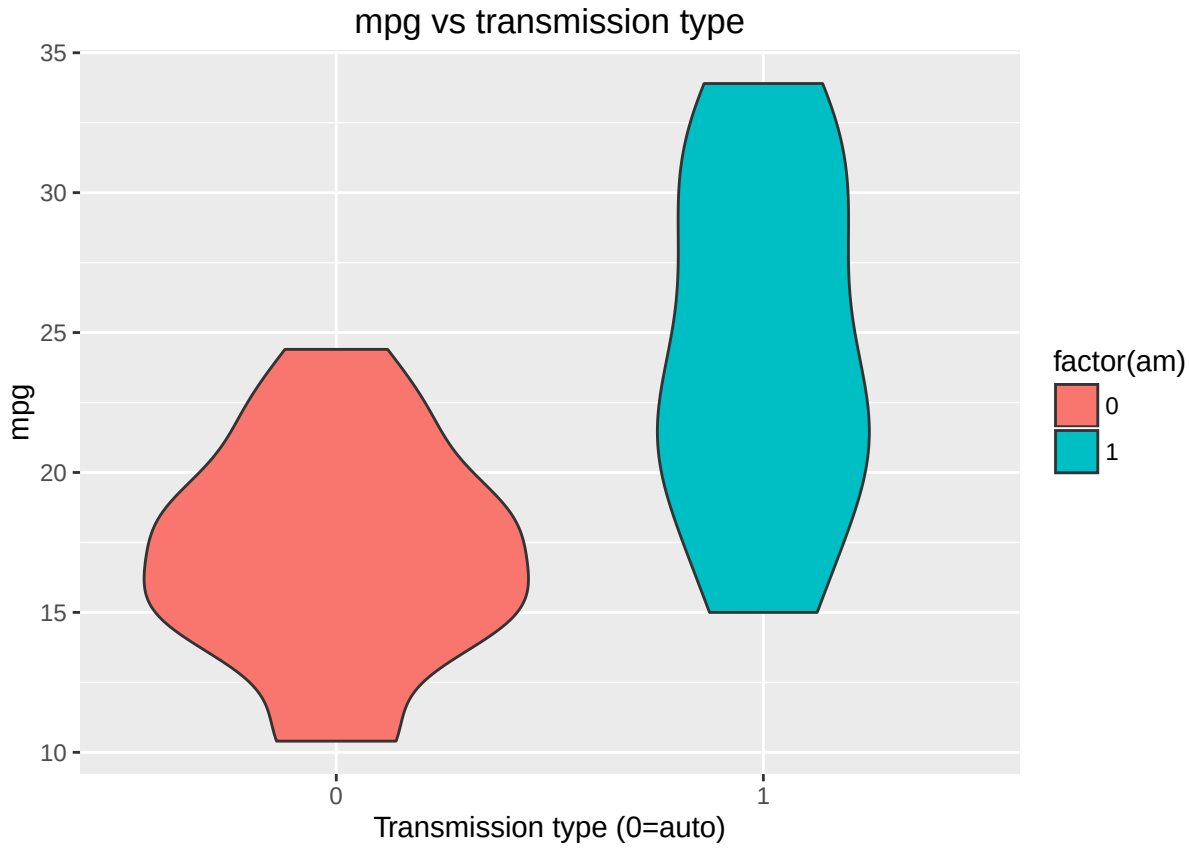


Figure 2

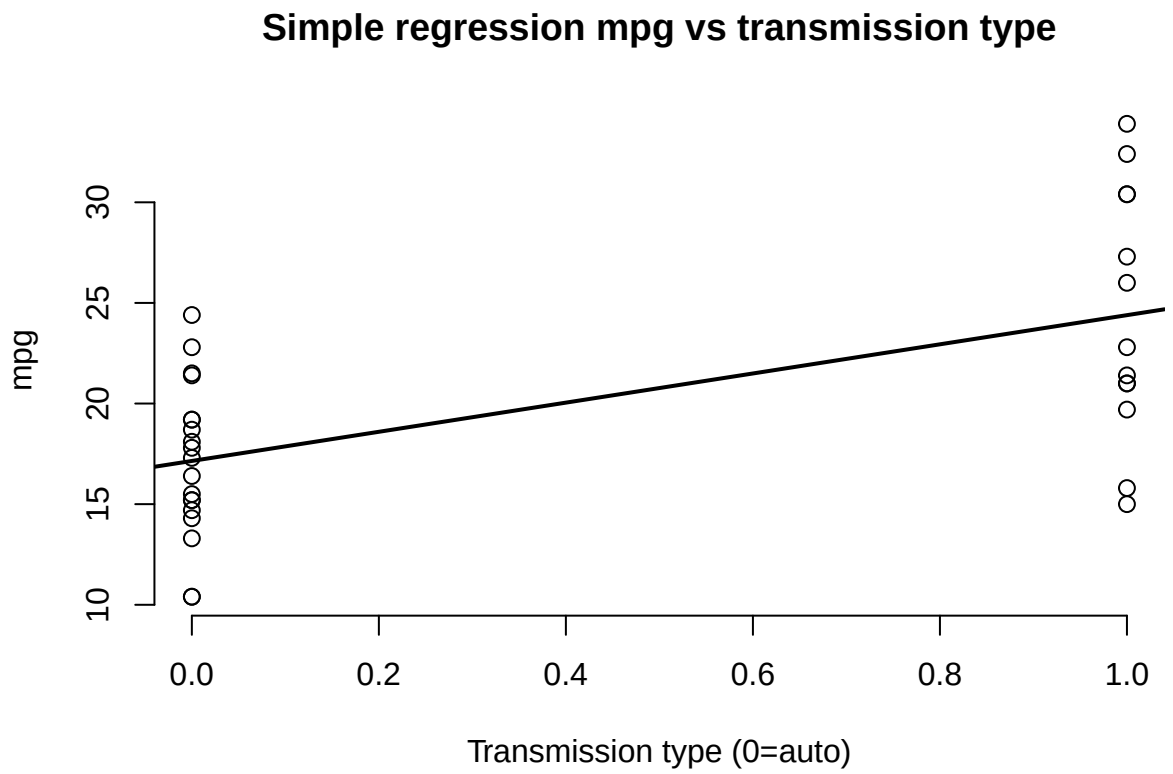
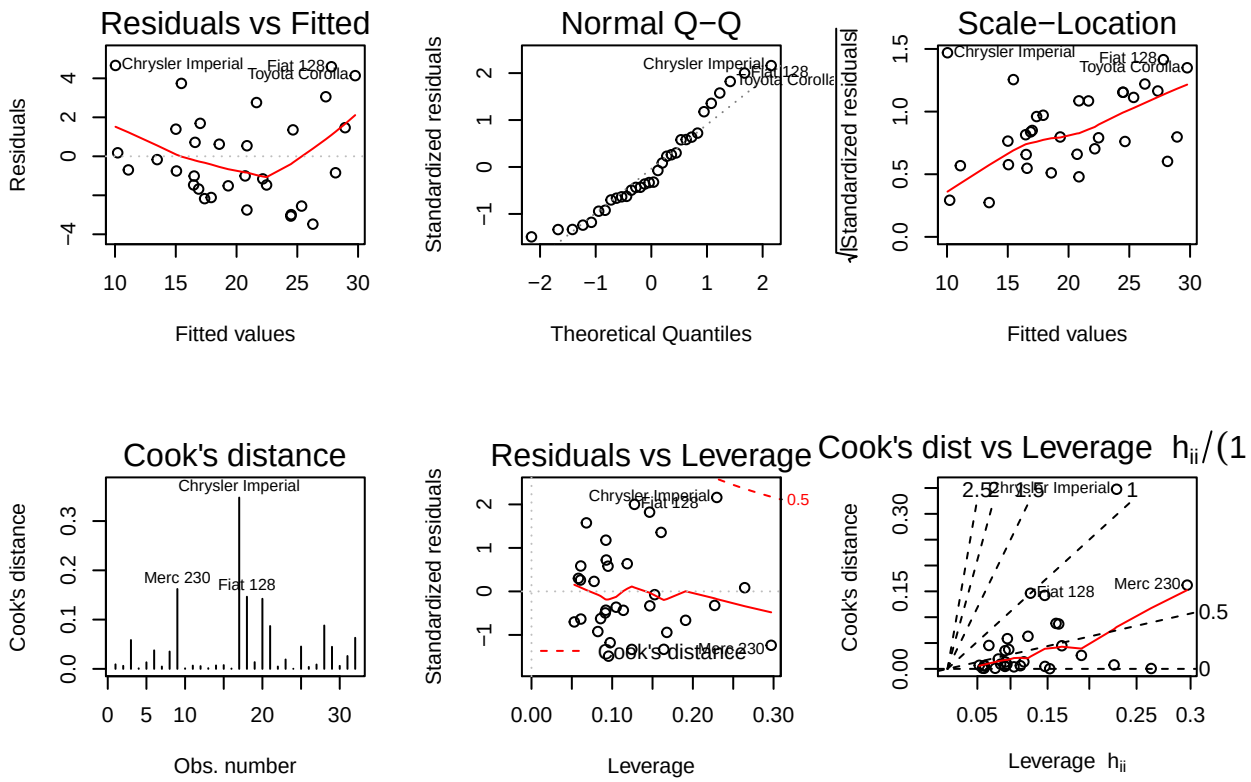

Simple regression mpg vs transmission type
mpg
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0
Transmission type (0=auto)

Figure 3



Markdown source ("https://github.com/dperov/RegressionModels_CourseProject/blob/master/Project.Rmd")