Regression Models Course Project

Hatem Jasim Hatem

May 17, 2019

Introduction

This 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

To answer above questions make exploratory analysis to determine relationship between MPG and Transmission type (Automatic, Manual).

Exploratory analysis

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

```
aggregate(mpg~am, data=mtcars, summary)
```

```
##   am mpg.Min. mpg.1st Qu. mpg.Median mpg.Mean mpg.3rd Qu. mpg.Max.
## 1  0 10.40000    14.95000   17.30000 17.14737    19.20000 24.40000
## 2  1 15.00000    21.00000   22.80000 24.39231    30.40000 33.90000
```

The exploratory analysis show difference in **MPG** depend on gear transmission types as shown in figure in appendix. This difference is significant depend on T-test p-value as shown below.

```
t.test(mpg ~ am, data = mtcars)$p.value
```

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

To develop mathematical model between transition and **MPG** used linear regression.

Modeling

```
model_1<- lm(mpg~factor(am), data = mtcars)
summary(model_1)
```

```
##
## 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
```

The manual transition care have 7.245 mpg than automatic car. But $R^2 = 0.3598$ this mean only 0.3598% of mpg variance controlled by gear transmission types. Therefore need to add more variables related with ***MPG*** to increase predict confidence.

```
res <- cor(mtcars)
round(res, 2)
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

The correlation table show there are many variables strongly related to **MPG**.

Therefore must develop **MPG** mathematical model with many variables rather than one variable by using multiple linear regression. Because there are variable does not correlated with **MPG** therefore use **stepwise** function to determine best variable to develop **MPG** mathematical formula.

```
model_2<- step(lm(mpg ~ ., data = mtcars), trace=0)
summary(model_2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, 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
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## 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
```

Conclusion

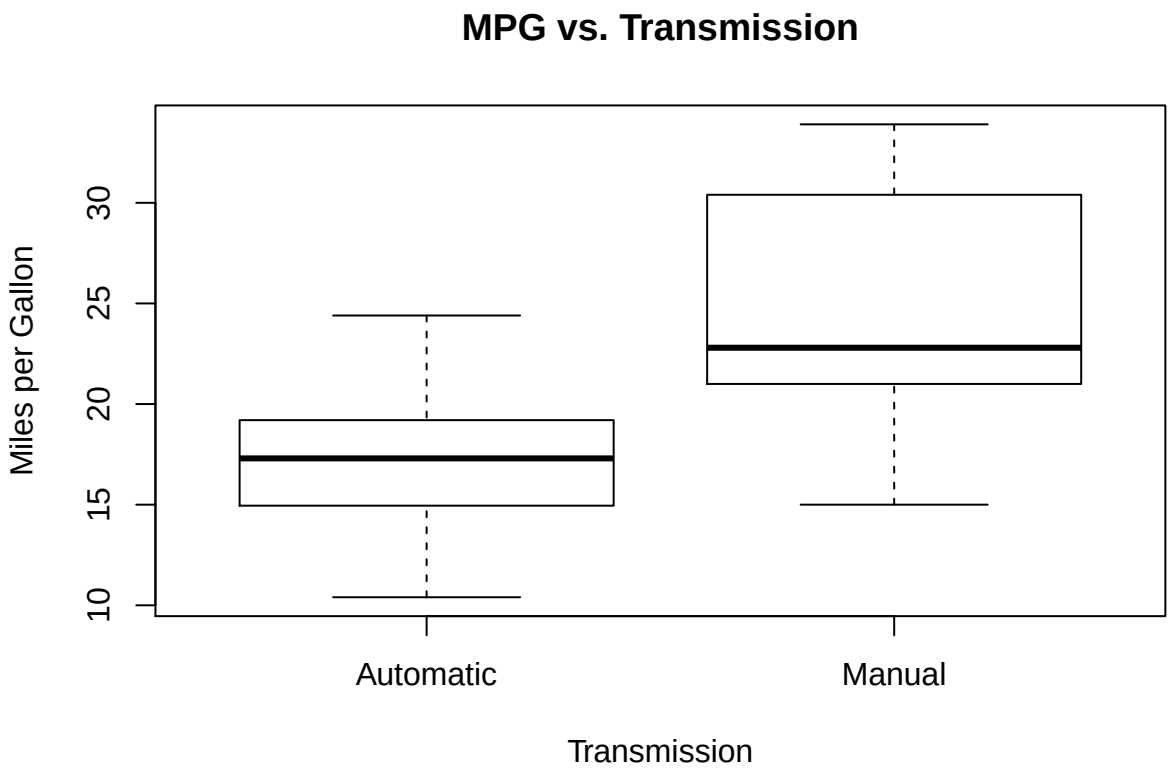
According to summary of multiple linear regression (with, qsec, am) beset variable to predict **MPG** with variance 0.8497 % based on p.value and R^2

- Increase care wight decrease **MPG** by $-3.9165 * wt$ with probability $1 - 2x6.95e^6$
- Increase care accerlation increase **MPG** by $1.2259 * qsec$ with probability $1 - 2x0.000216$.
- Transition mode increase **MPG** by 2.9358 if manual or not effect if automatic with probablity $1 - 2x0.046716$.
- the final formula is for **MPG**:

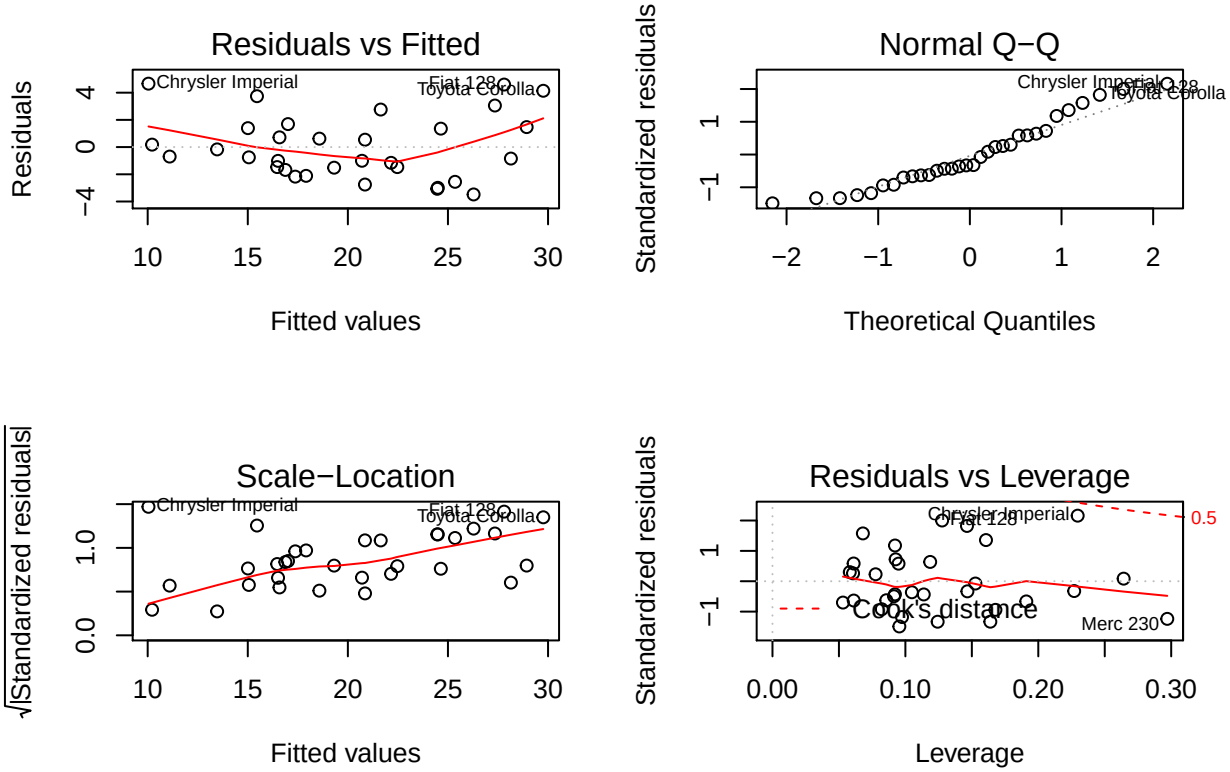
$$mpg = 9.6178 - 3.9165wt + 1.2259qsec + 2.9358am + \epsilon$$

Appendix

```r
boxplot(mpg ~ factor(mtcars$am,labels=c('Automatic','Manual')),
        data=mtcars,
        xlab="Transmission",
        ylab="Miles per Gallon",
        main="MPG vs. Transmission")
```




```
par(mfrow = c(2,2))
plot(model_2)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
Merc 230
Leverage