Regression Model on Motor Trend

Devara Izaz Fathan

8/28/2020

1. Excecutive Summary

In this assignment, we 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:

1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

2. Loading Data and Library

The data set that we use in this project is from the : Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411. We also need to assign some libraries that are useful for our analysis.

```
data(mtcars)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4

## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

3. Exploratory Data Analysis

First we look at the first lines of the data. We can is in the appendix 1

Then we look at the structure of the data, we can is in the appendix 2

The data contains 32 observations on 11 variables. The detailed variables translate is in appendix 4

For summarizing the data the we are interested to, we need to look the distribution of the mpg and adjusting the am variable through the boxplot. We can look at the appendix 3. Based on this plot, we can see that mpg for manual transmission is higher than mpg for automatic transmission. We will look at the mean of mpg for the auto and manual transmission

For further analysis, we have to take a look on the correlation of other variables with the pairs function. You can see in the appendix 5

4. Regression Modelling

Model 1: Simple Model

In Model 1, we only use variable mpg and am. We can see the

```
model1 <- lm(mpg~am, mtcars)
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## am             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
```

Model 1 shows us that the value of the intercept(beta0) is the mean value of auto transmission and the coefficient(beta1) value means the increasing mpg value when using manual transmission. The R-squared value=0.3385, it means that am variable only give 33,85% proportion of variance to the mpg as the dependent variable. So we have to fit the data into another model.

Model 2 : Initial Model

In this model, we include all variables into the regression model. This is the result :

```
model2 <- lm(mpg~., mtcars)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

From the result above, we see that the coefficient for wt is very dominant. The increasing per 1 unit value of wt reduce -3.715 to the mpg value. Based on this, we can not use all of the variables on this data. If we use all of the data, we may have overfitting problem that our model will not give us nearly the real value. So we have to reduce the variables we will include in the data.

Model 3 : Best Model

In this model, we will use step function in R to get best model automatically. This is taken care by the step method which runs lm multiple times to build multiple regression models and select the best variables from them using both forward selection and backward elimination methods by the AIC algorithm. The smaller the value of the AIC then better the model. For the detail tracing, we can see in the appendix.

```
model3 <- step(model2, direction=’both’, trace=FALSE)
summary(model3)
```

```
##
## 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
```

The model include wt, qsec, and am variable. The value of the intercept(beta0) when all of the independent is 0 is 9.6178. If we increase wt variable per one unit(1000 lbs), then the mpg will decrase -3.9165. Then if we increase the qsec per one unit(1 second) then the mpg will increase 1.2259. Last, if the transmission(am) is manual, the mpg value will increase 2.9358 from the automatic transmission.

The R-squared value=0.8336, it means that am, qsec, and wt variable give 83.36% proportion of variance to the mpg as the dependent variable. So we can use it as our model because R squared is big enough to represent the result of the mpg value. Our adjusted R squared value in model 3 close to the Multiple R squared. Thus we can use model 3 as our regression model for this project.

To evaluate the inference, we need to see the p value for all of the independent variables. All of the p values are less than 0.05, so we can say that all of the variables reject the null hypothesis that there is no relationship between the mpg and each variables. Therefore, we an look the confident interval of the variables.

```
confint.lm(model3)
```

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

5. Diagnostics and Residuals

For this analysis, the plot is shown in the appendix 6

- The Residuals vs Fitted plot, the points are randomly plotted, so there is no systematic error such heteroscedasticity.
- In the Normal Q-Q plot, we see that the point is located in the normal line. So the residuals are normally distributed.
- The Scale-Location plot consists of points located in a constant nearly constant line pattern, indicating constant variance.
- Based on the Residuals vs Leverage plot, we can see that there are no point outside the cook distance's, so we can say no point that have high influence in this model.

6. Conclussion

To make conclion, let's look at the original two questions:

1. Is an automatic or manual transmission better for MPG?

As we can see above from each of all three regression models, manual transmission is clearly better for MPG than automatic transmission.

2. Quantify the difference between automatic and manual transmissions.

Our regression models have quantified how much better manual transmission is than automatic transmission by three different amounts:

model 1 = 7.245 model 2 = 2.52023 model 3 = 2.9358

At the modelling analysis we have conclude that we use model 3 as our regression model. Thus we conclude that the differences between manual and automatic transmission is 2.9358

APPENDIX 1

```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

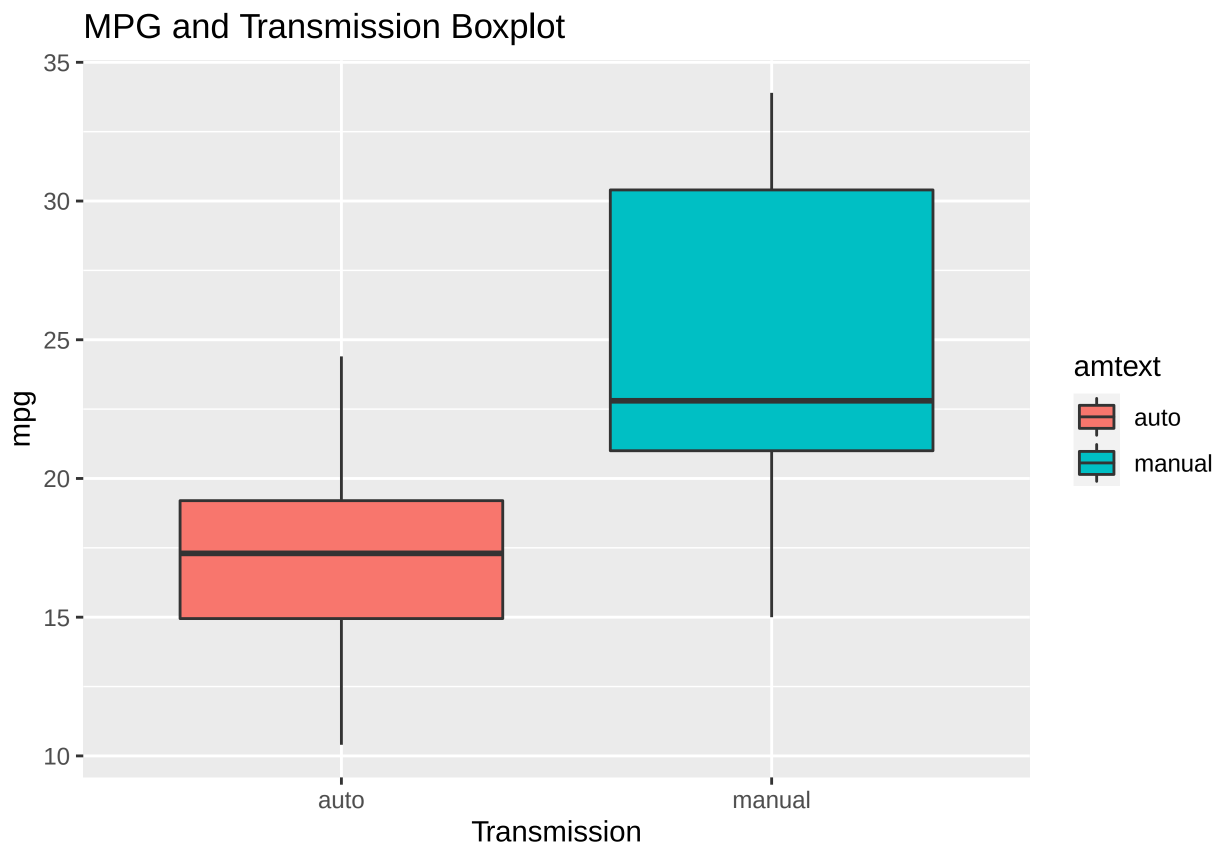
APPENDIX 2

```
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 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

APPENDIX 3

```
cars <- mtcars %>% mutate(amtext = ifelse(am== 1, ’manual’, ’auto’))
g <- cars %>% ggplot(aes(amtext, mpg)) +
      geom_boxplot(aes(group=amtext, fill=amtext)) +
      ggtitle(’MPG and Transmission Boxplot’) +
      xlab(’Transmission’)
g
```

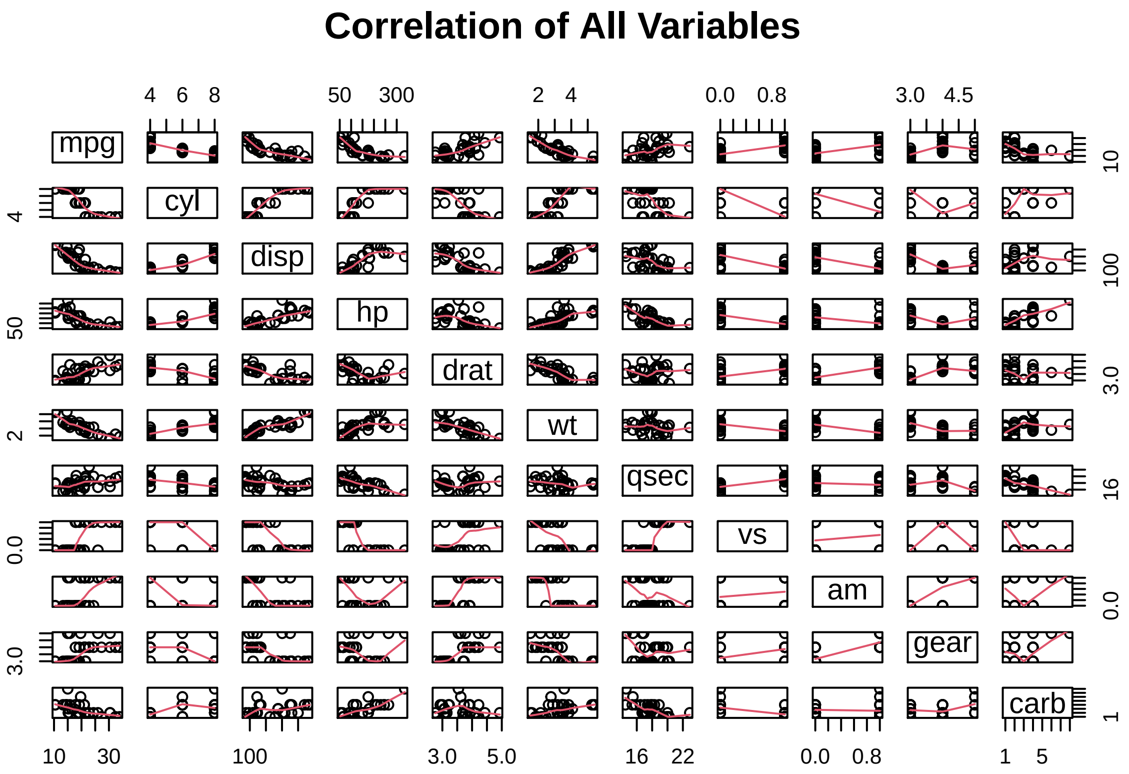


APPENDIX 4

mpg = Miles/(US) gallon
cyl = Number of cylinders
disp = Displacement (cu.in.)
hp = Gross horsepower
drat = Rear axle ratio
wt = Weight (1000 lbs)
qsec = 1/4 mile time
vs = Engine (0 = V-shaped, 1 = straight)
am = Transmission (0 = automatic, 1 = manual)
gear = Number of forward gears
carb = Number of carburetors

APPENDIX 5

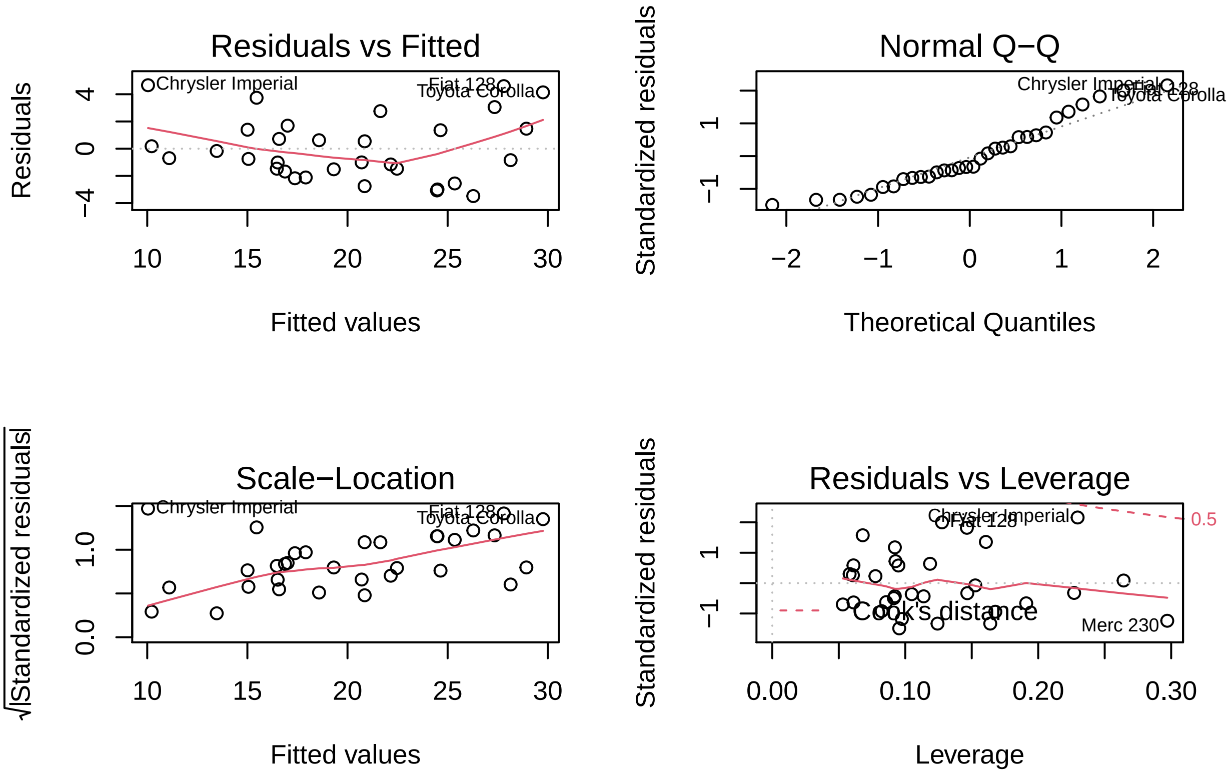
```
pairs(mtcars, panel=panel.smooth, main= 'Correlation of All Variables')
```



Correlation of All Variables

APPENDIX 6

```
par(mfrow=c(2,2))
plot(model3)
```


Residuals vs Fitted
Residuals
Fitted values
4
0
−4
10
15
20
25
30
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
1
−1
−2
−1
0
1
2
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale−Location
√|Standardized residuals|
Fitted values
1.0
0.0
10
15
20
25
30
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
Leverage
1
−1
0.00
0.10
0.20
0.30
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5