Regression Models Peer Assessment

Summary

This study is supposed to explore the 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.

Data Processing

mtcars dataset

data(mtcars)

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 ...
```

The variables present in the dataset are:

mpg - Miles/(US) gallon

cyl - Number of cylinders

disp - Displacement (cu.in.)

hp - Gross horsepower

drat - Rear axle ratio

wt - Weight (lb/1000)

qsec - 1/4 mile time

vs - V/S

am - Transmission (0 = automatic, 1 = manual)

gear - Number of forward gears

carb - Number of carburators

Let's convert am feature into a categorical one:

mtcars$am <- as.factor(mtcars$am)

levels(mtcars$am) <-c("AT", "MT") ## AT for Automatic T## MT for Manual

Exploratory Data Analysis

First, we need to check if mpg has a distribution close to the gaussian distribution so we can apply regression.

The mpg histogram is presented in Appendix.1 and in spite of the reduced observations in mtcars definitely it resembles a normal distribution.

Data Analysis

Next, the means and ranges of mpg for AT versus MT are presented in Appendix.2 boxplot

“Is an automatic or manual transmission better for MPG ?”

To get exact values and confidence interval for fuel consumption by AT vs. MT vehicles, we split the dataset by AT vs. MT

and apply the t-test:

mpg.at <- mtcars[mtcars$am == "AT",]$mpg

```
mpg.mt <- mtcars[mtcars$am == "MT",]$mpg

t.test(mpg.at, mpg.mt)

##

##     Welch Two Sample t-test

##

## data:     mpg.at and mpg.mt

## t = -3.767, df = 18.33, p-value = 0.001374

## alternative hypothesis: true difference in means is not equal to 0

## 95 percent confidence interval:

##     -11.28     -3.21

## sample estimates:

## mean of x mean of y

##           17.15           24.39
```

As the p-value is 0.001374 well bellow 5% or 1%, the alternative hypothesis is true: the difference in means is not equal to 0. So, the mean mileage of automatic transmission is 17.15 mpg and the manual transmission is 24.39 mpg.

The 95% confidence interval of the difference in mean gas mileage is between 3.21 and 11.28 mpg.

Regression Analysis

Let's start with the simplest regression model for mpg with only one predictor: am.

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

summary(model1)

##

## Call:

## lm(formula = mpg ~ am, data = mtcars)

##

## Residuals:
```

```
##        Min                  1Q Median                  3Q             Max

## -9.392          -3.092 -0.297       3.244              9.508

##

## Coefficients:

##                              Estimate Std.    Error          t    value    Pr(>|t|)

## (Intercept)        17.15                  1.12              15.25              1.1e-15 ***

## amMT                        7.24            1.76               4.11      0.00029 ***

## ---

## Signif. codes:    0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 4.9 on 30 degrees of freedom

## Multiple R-squared:    0.36,      Adjusted R-squared:    0.338

## F-statistic: 16.9 on 1 and 30 DF,    p-value: 0.000285
```

The null hypothesis is rejected by p-value = 0.000285 but the regression model covers only 36% of the variance.

So, the model coefficients are:

intercept = 17.15 represents automatic cars mean mpg

am coefficient = 7.24 represents the adjusted estimate for the expected change in mpg comparing AT versus MT.

Next, we will use a more complex regression model:

```
model.all <- lm(mpg ~ ., data=mtcars)
```

and go backwards to the fitest model using step() function:

```
model.best <- step(model.all, trace=0)

summary(model.best)

##

## Call:

## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.481  -1.556  -0.726   1.411   4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## wt            -3.917      0.711   -5.51  7e-06 ***
## qsec           1.226      0.289    4.25  0.00022 ***
## amMT           2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
```

The “Occam's razor” model explains 85% of mpg variance and contains only 3 predictors:

formula = mpg ~ wt + qsec + am

amMT estimated coefficient equals now to 2.9358 and represents the adjusted estimate for the expected change in mpg comparing AT versus MT for this new model containing 2 other predictors besides am.

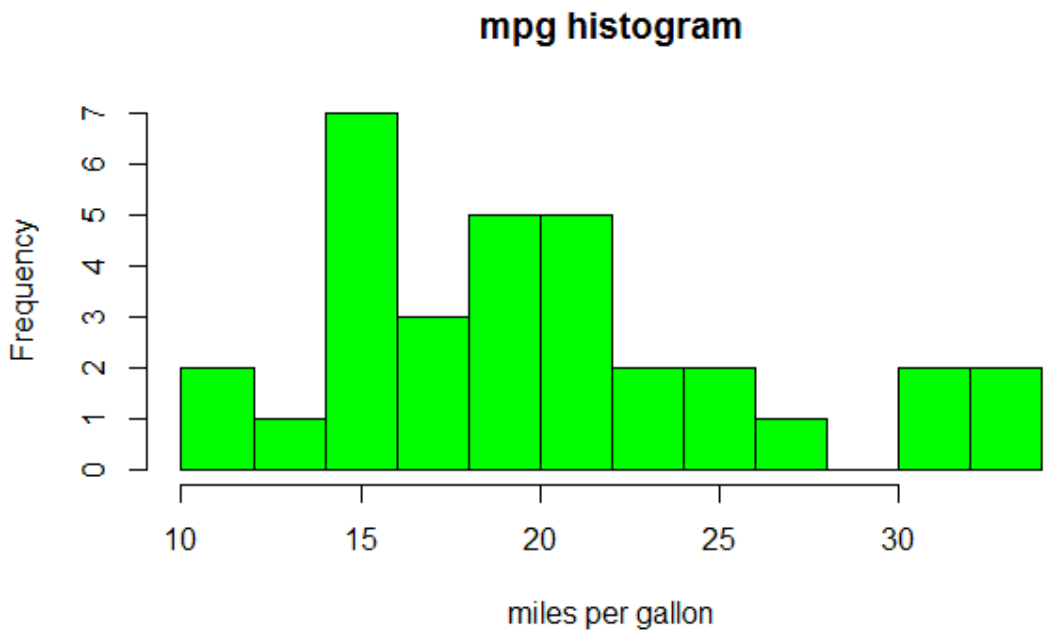
-amMT estimated coefficient is the answer to the second question.

Best model residuals are depicted in Appendix.3

First graphic, “Residuals vs. Fitted values” is not quite a straight line, proof of some outliers.

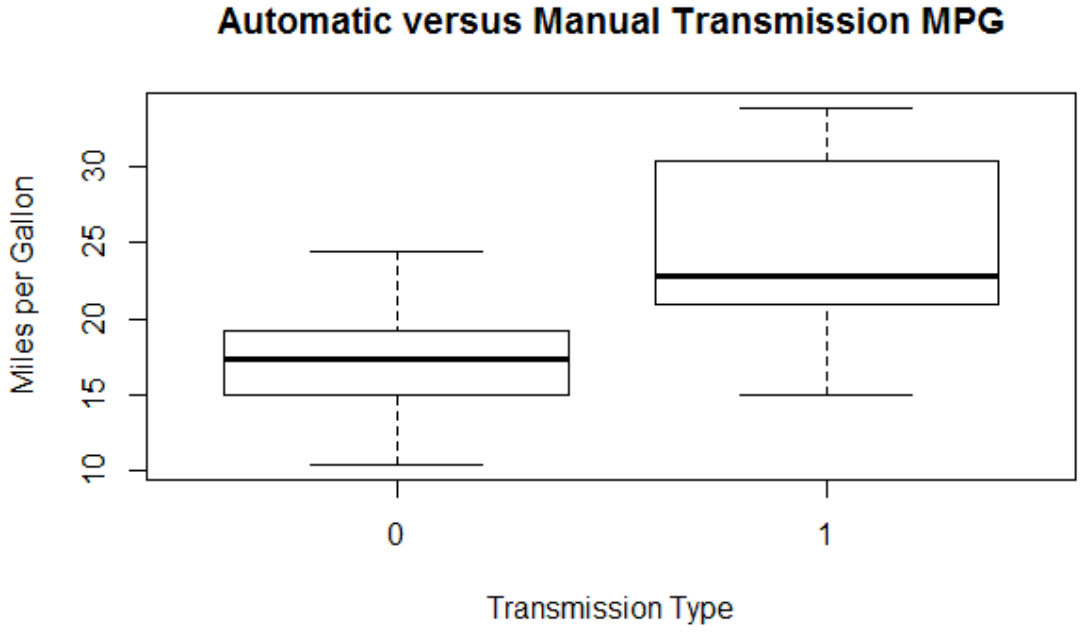
Appendix.1 Miles per Gallon Histogram

```
hist(mtcars$mpg, breaks=12, xlab="miles per gallon", main="mpg histogram", col="green")
```



Appendix.2 Automatic vs Manual Transmission MPG

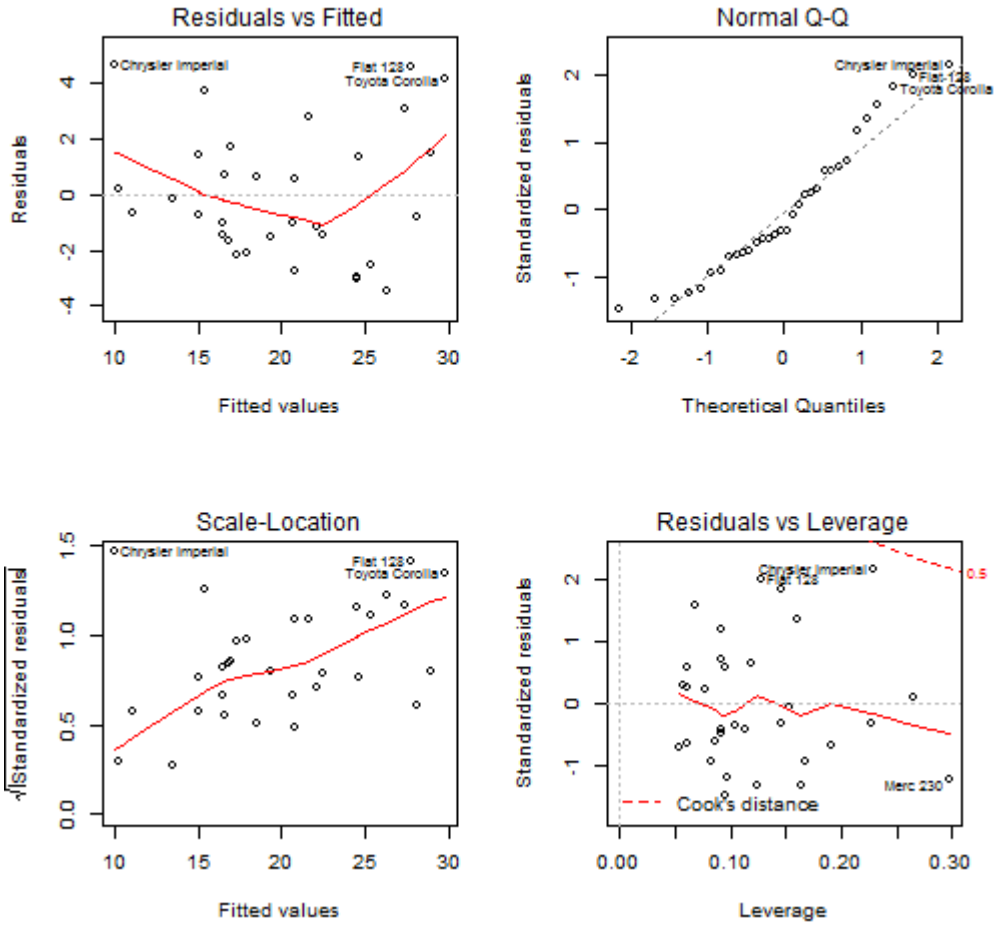
```
boxplot(mpg ~ am, data=mtcars, xlab="Transmission Type", ylab="Miles per Gallon",
        main="Automatic versus Manual Transmission MPG")
```



Appendix.3 Best model residuals

par(mfrow = c(2,2))

plot(model.best)



Final conclusion

From the selected model (fit2), we conclude that the cars with manual transmission have MPG 2.94 higher than the cars with automatic transmission.