Regression Models Analysis - mtcars

Rahul Jain

19/02/2022

Executive Summary

I did analysis for Motor Trend magazine by looking at a data set of a collection of cars namely "mtcars" data-set. We were primarily interested in looking at relationship between MPG and other variables like number of cylinders, displacement etc. Our primary focus was looking at relationship between mpg and automatic transmission analyzing below 2 questions: 1. Is automatic transmission better for mpg or manual transmission 2. Quantifiable difference between the two transmission types

Conclusions

1. Manual transmission has more mpg compared to automatic transmission which has been concluded by using box-plot, t-test and also regression model
2. At 95% confidence level, we can conclude that difference in means from automatic-manual is between -11.3 to -3.2 miles per gallon
3. From simple regression model, we concluded that for manual transmission , mean mpg is 7.245 more than automatic transmission
4. From our best multivariate regression model, we concluded that manual transmission cars has mean mpg of 2.94 mpg more than automatic mpg after adjusting for other variables - wt and qsec

Importing the libraries

```
library(ggplot2)
library(car)
library(GGally)
```

Exploratory Data Analysis

Transmission variables is coded as numeric(0 - automatic, 1 = manual), we will convert to factor variable

```
mtcars$am <- factor(mtcars$am,labels = c("automatic","manual"))
```

As our analysis focus on am variable, we will use a box-plot to analyze whether there is a difference in mpg for different types of transmission. Corresponding box-plot is present in appendix.

From box-plot we can see mpg is higher for manual transmission compared to automatic transmission.

We will do t-test to check whether this difference is statistically significant or not

2-Sample t-test

```
t <- t.test(mtcars$mpg~mtcars$am)
```

From the t-test we can observe that difference between 2 means is statistically significant. Hence, transmission type plays an important role in mpg

We will also look at pairwise scatter plots to observe whether there is any linear relationship between mpg and other variables within the data. Pair-plot is present within appendix

Simple Linear Regression Model

First, we will fit a simple linear regression model between mpg and am and analyze the same.

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

Model Interpretation * Intercept estimate - shows that for automatic transmission mean mpg is 17.147 miles per gallon * am1 estimate - shows for manual transmission , mean mpg is 7.245 more than automatic transmission * p - values - shows that coefficients are statistically significant at 5% alpha * R-squared - is 0.34 which is very low, shows that only 34% variation is captured by fitting mpg with am

Multivariate Rgeression Model

Now, we will fit regression model on mpg with the variables

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

We can observe that all variables apart from wt are not statistically significant. There is a increase in r-squared but we can do better to fit a model. We can look at vif to see which variables could be correlated within each other.

```
v1 <- vif(fit_all)
```

From above, w can observe that below variables have high correlation with other variables because of high vif 1. cyl - 15.37 2. disp - 21.6 3. wt - 15.17

Let's fit different models by including variables one at a time

Stepwise Regression Model

```
fit1 <- lm(mpg~am,data = mtcars)
fit2 <- lm(mpg~wt+am, data = mtcars)
```

From anova test, we can observe that including wt is statistically significant. Let's try including cyl which also had a high correlation with mpg

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

We can observe that adding disp is not significant at 5% alpha. We can try adding different models to see which model will be best

```
fit4 <- lm(mpg~am+wt+disp+hp,data = mtcars)
fit5 <- lm(mpg~am+wt+disp + hp+ cyl,data = mtcars)
fit6 <- lm(mpg~am+wt+disp+hp+cyl+vs, data = mtcars)
```

From above, we can observe that hp variable to regression model had better performance as compared to adding disp. We can automatically select best model using step function in R

```
best_fit <- step(fit_all,direction = "both",trace = FALSE)
```

We can observe that best model had wt,qsec and am as features as part of regression model and had a r-squared of 0.85 with variable am statistically significant. Vif also shows that there is lower co-linearity as compared to model with all variables

Model Diagnostics

```
v2 <- vif(best_fit)
```

Residual plot is present in Appendix

- From the residual plot, we can see that there is no visible pattern between residuals and fitted values
- Q-Q plot shows that residuals are mostly normal but deviate slightly at the end

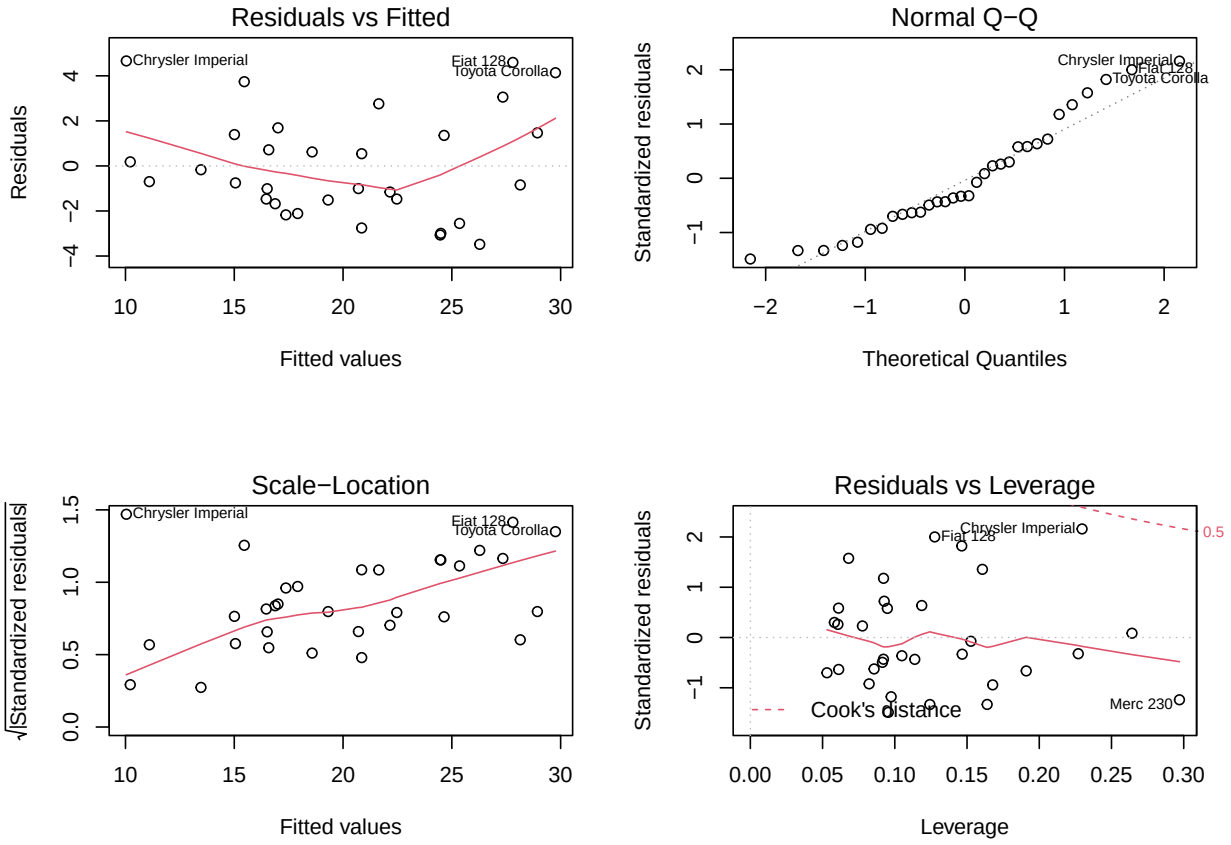
Appendix

```
t
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means between group automatic and group manual is not equa
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group automatic    mean in group manual
##                17.14737                24.39231
```

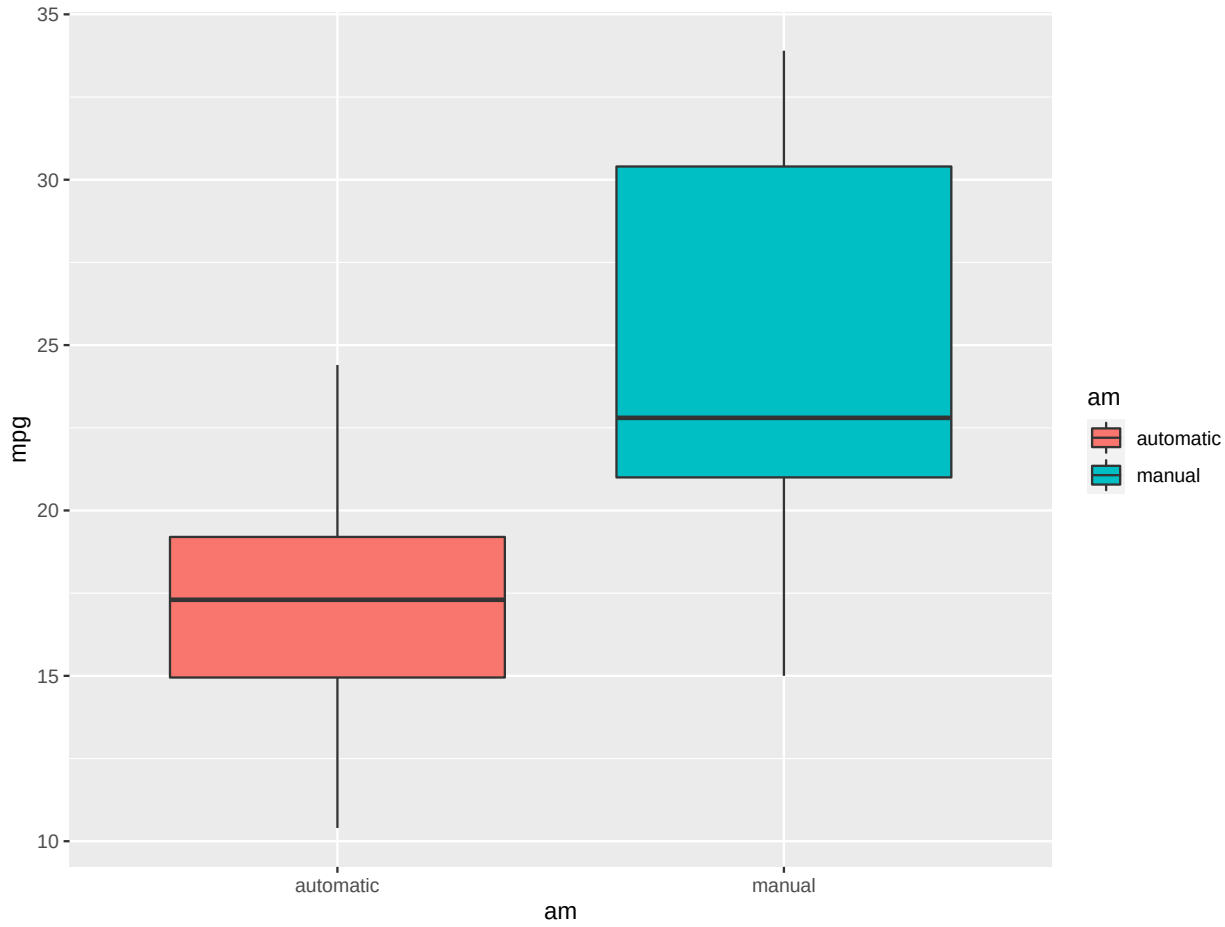
Regression Diagnostic plot

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



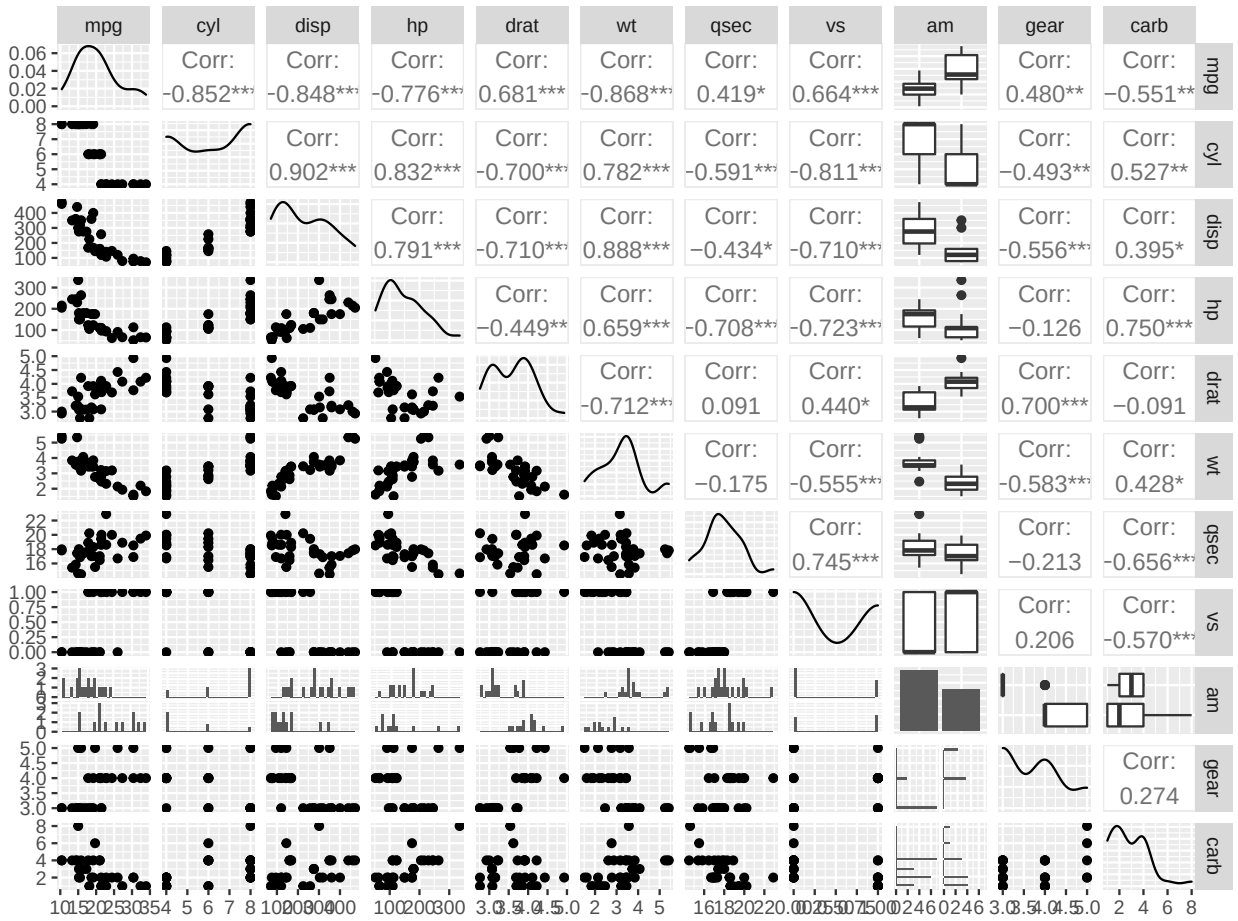
Box-plot for (mpg vs am)

```
g <- ggplot(data = mtcars, aes(x = am, y = mpg,fill = am))
g + geom_boxplot()
```



s Pair-plot

```
ggpairs(mtcars)
```



Summary of Simple Linear Regression Model

```
summary(fit_am)
```

```
##
## 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 ***
## ammanual       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
```

Summary from multivariate-regression model fitting all variables

```
summary(fit_all)
```

```
##
## 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
## ammanual     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
```

```
v1
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
##      gear      carb
##  5.357452  7.908747
```

Summary of best-fit model

```
summary(best_fit)
```

```
##
## 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 ***
## ammanual      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
```

```
v2
```

```
##       wt     qsec       am
## 2.482952 1.364339 2.541437
```

Anova Diagnostics for Step-wise multivariate models

```
anova(fit1,fit2,fit3,fit4,fit5,fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ am + wt + disp
## Model 4: mpg ~ am + wt + disp + hp
## Model 5: mpg ~ am + wt + disp + hp + cyl
## Model 6: mpg ~ am + wt + disp + hp + cyl + vs
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 69.3600 1.12e-08 ***
## 3     28 246.56  1     31.76  4.9779 0.034878 *
## 4     27 179.91  1     66.65 10.4451 0.003436 **
## 5     26 163.12  1     16.79  2.6310 0.117344
## 6     25 159.52  1      3.60  0.5639 0.459698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```