Regression Models Project - Motor Trend Magazine

Rajaram

30/08/2020

Executive Summary

As part of this project, we are looking at a data set of collection of cars and interested in exploring relationship between the set of variables and miles per gallon (MPG).

We are interested in addressing the following 2 questions using linear regression:

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

Data Description

We are going to use mtcars dataset for the regression modeling to illustrate how automatic or manual transmission cars affect MPG.

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

This dataset consists of 11 variables, all are numeric.

- 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

Data Preparation

In this section, we load in the data and do necessary transformation, which is coercing some variables to factor variables.

```
mtcars <- mtcars
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
```

```r
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
levels(mtcars$am) <- c("Auto","Manual")
```

Exploratory Data Analysis

In this section, we are going to check if transmission type has any relationship with MPG.

Refer to Fig.1 in Appendix.Boxplot clearly shows that manual transmission cars provides better mileage than automatic transmission cars.

We will check statistically if the above information derived from boxplot is correct using t tests.

```r
autoTrans <- mtcars[mtcars$am=="Auto",]
manualTrans <- mtcars[mtcars$am=="Manual",]
t.test(autoTrans$mpg, manualTrans$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  autoTrans$mpg and manualTrans$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

As per T test results, p-value is around 0.001 which is less than significance level of 0.05. Hence, we reject the null hypothesis that there is no difference between mpg for manual vs auto transmission cars.

Regression analysis

We are going to perform regression analysis with single variable to check if there is a relationship between to support there is a relationship between transmission type and MPG.Based on the results, we will perform regression analysis with multiple variables.

```r
lrModel <- lm(mpg ~ am, data = mtcars)
summary(lrModel)
```

```
##
## 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
```

As per summary of results, it shows there is a relationship between transmission type and MPG. Estimated MPG of automatic transmission cars is 17 and manual transmission cars is 7.2 more than automatic transmission cars.However, R-squared score of 0.3598 suggests that transmission type accounts for only 35% of the performance of MPG. Hence, there might be some other factors which influence mpg.

We use analysis of variance and step function for our model selection.

```
t_var <- aov(mpg ~ ., data=mtcars)
summary(t_var)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of analysis of variance tests, We can consider cyl, disp and wt variables along with am for regression model.

We will perform step function to cross check, if there are any other factors to be considered.

As per scatterplot in Appendix also confirms the same about relationship of mpg with these variables.

```
fullModelFit <- lm(mpg ~ ., data = mtcars)
stepFit <- step(fullModelFit)
```

```
summary(stepFit)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Step function identifies cyl, hp and wt variables to consider for the slope and this produces R-Squared score of 0.8659. Hence we will go with this combination for multiple variable linear regression.

```
mrModel <- lm (mpg ~ am+cyl+hp+wt, data=mtcars)
summary(mrModel)
```

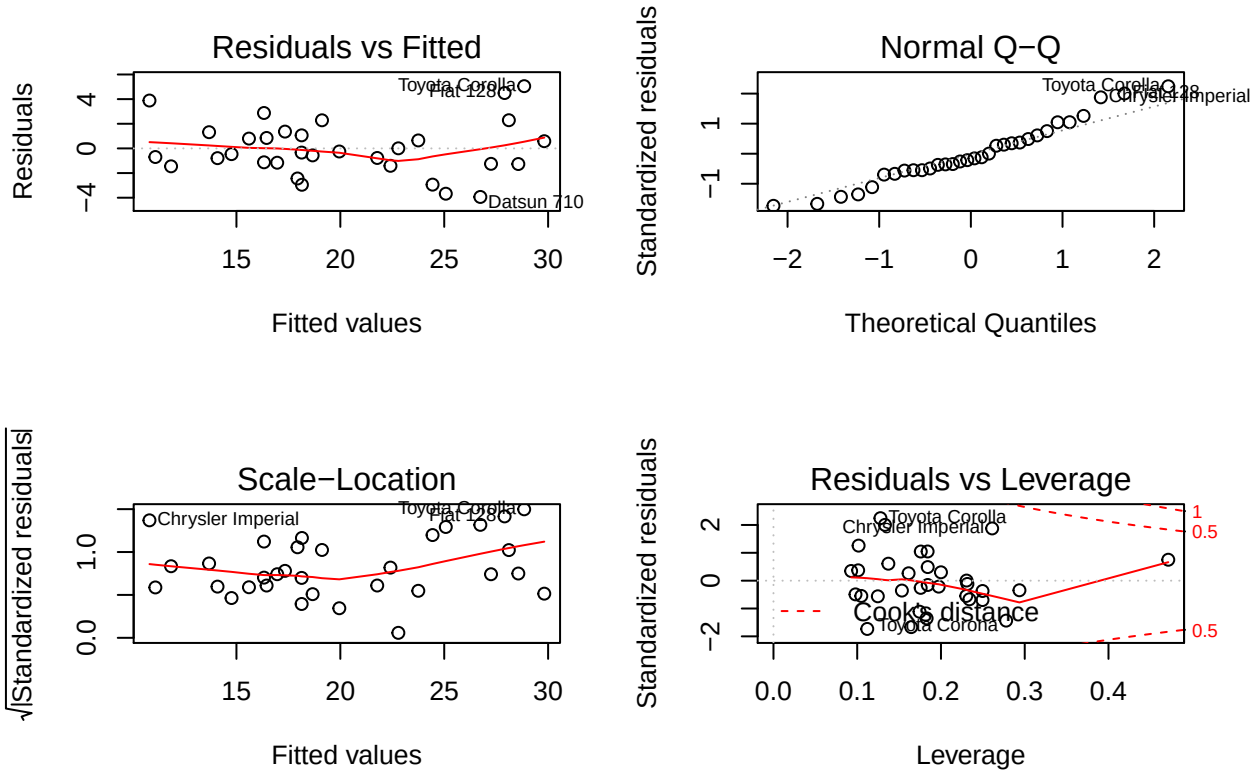
```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## amManual     1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

From the above, we observe that R-squared value is 0.8659. We conclude that more than 86% of MPG variability comes from these variables.

Residual and Diagnostics

In this section, we examine residual plots of our regression model and compute some diagnostics to identify outliers in dataset

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



The following observations can be made from the above plots:

- Residual vs Fitted plot is randomly scattered and verify the independence
- Normal Q-Q plot points mostly fall on the line indicating results are normally distributed
- Scale-Location plot points scattered in a constant band pattern
- Residuals vs Leverage plot points shows some outliers

We will compute diagnostics of our model now

```
leverage <- hatvalues(mrModel)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##          0.2777872           0.2936819          0.4713671
```

```
influential <- dfbetas(mrModel)
tail(sort(influential[,6]),3)
```

```
##       AMC Javelin     Toyota Corona Chrysler Imperial
##         0.2301107         0.3643262         0.9389082
```

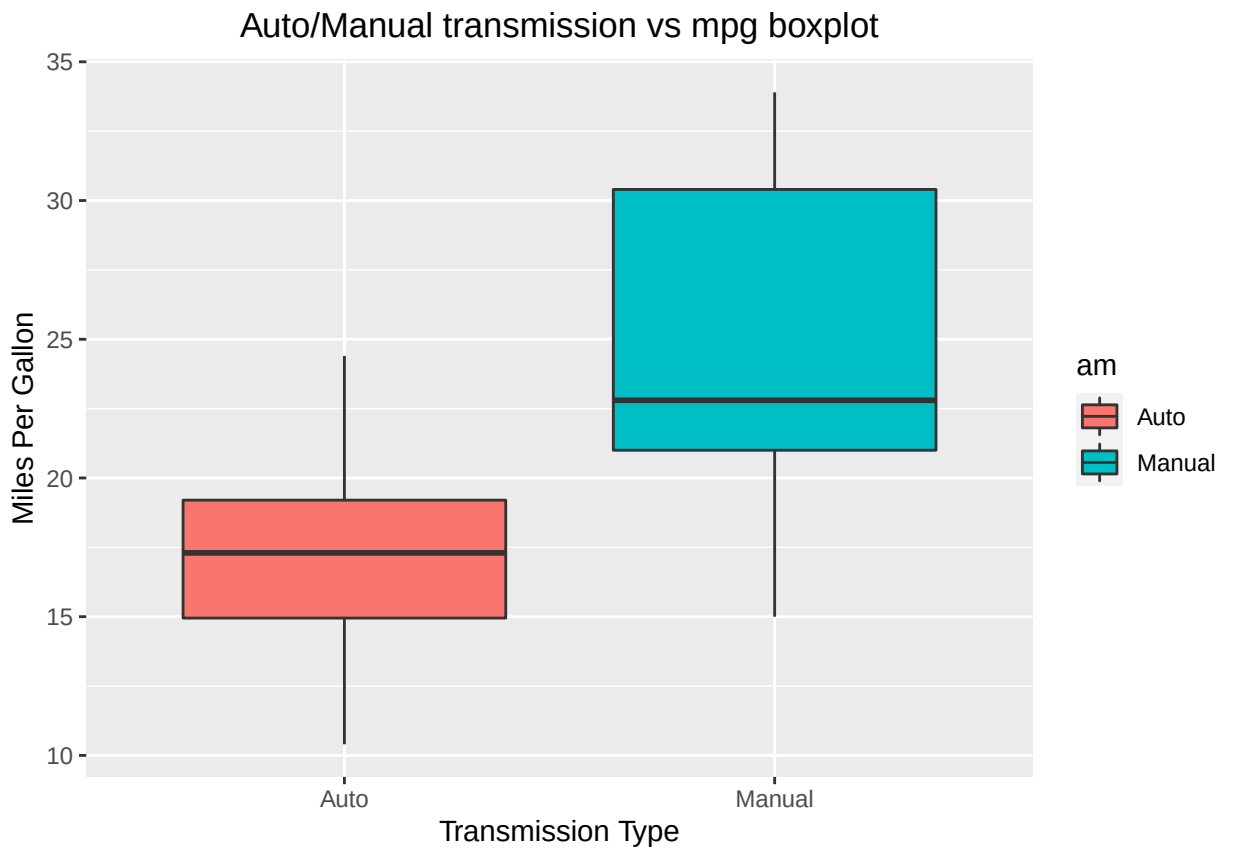
Conclusion

- Exploratory data analysis is conducted first to identify the relationship between MPG and manual/automatic transmission types.
- We infer manual transmission type cars gives 7.25 MPG better than automatic transmission types
- We then performed linear regression with one variable to check the relationship between MPG vs AM.
- Eventhough there was a relationship, R-Squared value is around 0.35 which explains variance of 35% can be accounted based on transmission type.
- Hence, Simple linear regression is not a very good model to answer definitively.
- We incorporate multiple variables linear regressions by conducting analysis of variance and step functions
- From the above tests, we identified there is a need to consider Number of Cylinders, Gross horsepower and Weight of the car.
- This model seems to be a very good which explains these variables account for 86% variance of MPG

Appendix

Boxplot of MPG vs Transmission Types

```
library(ggplot2)
transtyp <- ggplot(mtcars, aes(x=am, y=mpg)) + geom_boxplot(aes(fill=am))
transtyp <- transtyp + labs(title ="Auto/Manual transmission vs mpg boxplot")
transtyp <- transtyp + theme(plot.title=element_text(hjust = 0.5))
transtyp <- transtyp + xlab("Transmission Type") + ylab("Miles Per Gallon")
transtyp
```



scatter plot matrix for mtcars dataset

```
pairs(mpg ~ ., data = mtcars)
```