Regression Analysis of 'mtcars' dataset

Gaurav Tripathi

Introduction

As an exployee for Motor Trend, a magazine about the automobile industry, I analysed the 'mtcars' dataset (a data set 32 cars across 11 measures, captured between 1974-1975), to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). I have tried to answer the following two questions:

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

Executive Summary

Following are the findings from the analysis:

- The impact of manual vs. automatic transmission on MPG is statistically significant
- Manual Transmission cars tend to have higher MPG than Automatic Transmission ones
- However, as the car weight increases, with every 1000lb increase in weight, MPG for manual transmission cars falls faster than that for automatic transmission cars (all other factors remaining constant)
- At about 2800lb weight, the MPG for both Manual and Automatic Transmission cars is same (all other factors remaining constant)
- Below this weight, Manual Transmission Cars are better
- Above this weight, Automatic Transmission Cars are better
- Qsec (which indicates how fast car is driven) also impacts MPG in a significant way

Analysis and Data Processing

First, the library "datasets" was loaded, and then "mtcars" dataset was copied into another variable cars_data. Next, correlation between variables was checked (results shown in Table 1) and the following was observed. This could potentially help when adding / removing variables to the regression model:

- MPG seemed to be highly negatively correlated with Cyl, Disp, HP and WT
- Cyl seemed to be highly correlated with Disp, HP, WT and VS (negatively)

To check the high-level impact of transmission type, a box plot was drawn (Figure 1). It indicated that the impact is significant visually. But the same was tested for statistical significance too (assuming normal distribution and IID), using a Welch 2-sample T-test. A p-value of 0.0013 indicated that the different in impact was > 0 at least 95% times.

Hence, **manual transmission cars are better than automatic transmission cars in 95% cases**.

On closer analysis, it would be proved that the result obtained about was merely a start. A regression analysis between mpg (outcome) and am (variable) would give R-square of 0.36 only. Which means there is scope for addition of more variable, as well as for interaction between few variables. Hence, more models were chosen.

Following code snippet demonstrates the logic used when selecting / Rejecting models:

```r
# AM is insignificant in models below
summary(lm(mpg ~ am + cyl, data = cars_data))
summary(lm(mpg ~ am + disp, data = cars_data))

# The model below has the problem of Heteroscedasticity, as seen from Residuals plot
summary(lm(mpg ~ am + hp, data = cars_data))
fit <- lm(mpg ~ am + hp, data = cars_data)
plot(fit)

# New variables added are all insignificant
summary(lm(mpg ~ am + hp + drat, data = cars_data))
summary(lm(mpg ~ am + hp + qsec, data = cars_data))
summary(lm(mpg ~ am + hp + vs, data = cars_data))
summary(lm(mpg ~ am + hp + gear, data = cars_data))
summary(lm(mpg ~ am + hp + carb, data = cars_data))

# AM is insignificant in model below
summary(lm(mpg ~ am + hp + wt, data = cars_data))

# Since combinations with HP didn't work, removing that variable
# AM is insignificant in model below
summary(lm(mpg ~ am + wt, data = cars_data))
```

The model, which made sense was the one with high R2, AM as a significant variable, homoscedasticity, and no variable insignificant. Since none of the variable combinations chosen had helped, interaction variables were introduced. Interaction between AM and WT seemed to be interesting, a seen from regression analysis, as well as by plotting on a graph.

```r
# Model below seems to make logical sense -> MPG decrease with weight increase,
# assuming transmission type is not changed
summary(lm(mpg ~ am*wt, data = cars_data))

# Same can be validated from the following chart
plot(cars_data$wt, cars_data$mpg, col=cars_data$am, xlab = "Weight of car", ylab = "MPG")
fit_am_wt <- lm(mpg ~ wt*am, data=cars_data)
summary(fit_am_wt)
abline(coef(fit_am_wt)[1], coef(fit_am_wt)[2])
abline(coef(fit_am_wt)[1]+coef(fit_am_wt)[3], coef(fit_am_wt)[2]+coef(fit_am_wt)[4], col="red")
```

To see, if any other variable improved the R2, without compromising on the variable significance or homoscedasticity, multiple models were chosen. The results are as follows:

```r
summary(lm(mpg ~ am*wt + cyl, data = cars_data))    # R2 = 0.877
summary(lm(mpg ~ am*wt + disp, data = cars_data))   # Wt insignificant
summary(lm(mpg ~ am*wt + hp, data = cars_data))     # R2 = 0.87
summary(lm(mpg ~ am*wt + drat, data = cars_data))   # Drat insignificant
summary(lm(mpg ~ am*wt + qsec, data = cars_data))   # R2 = 0.90
summary(lm(mpg ~ am*wt + vs, data = cars_data))     # R2 = 0.87
summary(lm(mpg ~ am*wt + gear, data = cars_data))   # Gear insignificant
summary(lm(mpg ~ am*wt + carb, data = cars_data))   # Carb insignificant
```

Only the model with interaction between AM and WT, and with QSEC added made sense. The results on running that model are shown in Table 3. Its plots are shown in Figure 2. Following are the takeaway from this model:

- The model explains 90% of the results
- Increase in WT decreases MPG of the car
- Manual Transmission is better than Automatic Transmission
- But with increase in WT, manual tranmission car's performance starts to decrease (by a factor of 4.141) per 1000lb increase in WT
- There is no heteroscedasticity or outlier in the residual plot

ANOVA analysis of model with and without interaction indicates significant different (since $P < 0.05$). VIF analysis of the 2 models indicates an increase in variance in the model with interaction, as compared to the one without interaction, but given all other benefits seem from the model, we can live with this variance. Detailed results for ANOVA and VIF analysis are shown in Table 4.

Appendix

Table 1

```
##        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
```

Figure 1

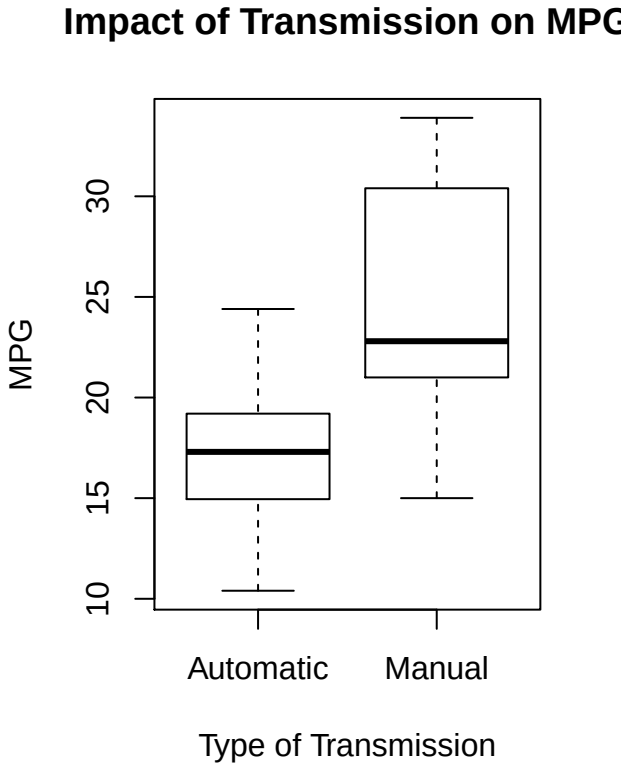

Table 3

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   9.723053  5.8990407  1.648243 0.1108925394
## amManual     14.079428  3.4352512  4.098515 0.0003408693
## wt           -2.936531  0.6660253 -4.409038 0.0001488947
## qsec          1.016974  0.2520152  4.035366 0.0004030165
## amManual:wt -4.141376  1.1968119 -3.460340 0.0018085763
```

```
## [1] "R square:"
```

```
## [1] 0.8958514
```

Figure 2

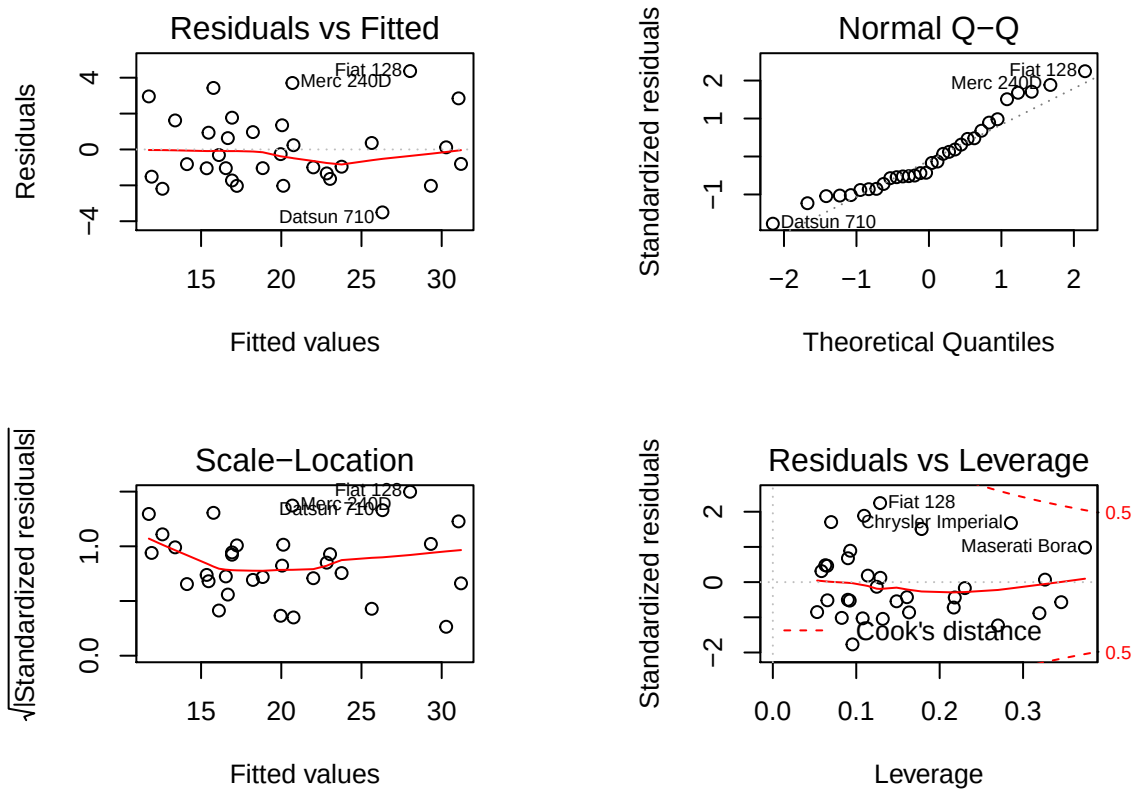

Table 4

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am * wt + qsec
## Model 2: mpg ~ am + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     27 117.28
## 2     28 169.29 -1    -52.01 11.974 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] "VIF analysis:"

##        am        wt      qsec     am:wt
## 20.970925  3.030963  1.447406 16.302453

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