

Regression Models Course Project

Relationship between transmission type and miles per gallon (MPG)

jin.jiangli

4/17/2020

Executive Summary

In this report, we will analyze `mtcars` data set in R and explore the relationship between a set of variables and miles per gallon (MPG). We are particularly interested in the following two questions:

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

From our analysis we can conclude that cars with manual transmission get more miles per gallon than cars with automatic transmission by a rate of 1.8 adjusted by number of cylinders, gross horsepower and weight.

Exploratory Data Analysis

We explore some relationships between variables of the data set and our outcome. First we plot the relationship between all variables of the data set (see Figure 2 in the appendix). From the plot we can see a strong correlation between variables such as: `disp`, `hp`, `drat`, `wt`, and `am` with our outcome `mpg`.

Since we are actually interested in quantifying the MPG difference between automatic and manual transmissions we also plot a boxplot between `mpg` and `am` (see Figure 1 in the appendix). We can see that there is an increase in `mpg` when the transmission is `manual`.

Regression

In this section we build a linear regression using different variables as predictors and `mpg` as outcome. To find the best model we perform a stepwise selection using both forward selection and backward elimination. The code below takes care of this.

```
first_model <- lm(mpg ~ ., cars)
best_model <- step(first_model, direction = "both")
```

We can see that the best model uses the variables `cyl`, `hp` and `wt` as confounders and `am` as the independent variable.

```
best_model$call
```

```
## lm(formula = mpg ~ cyl + hp + wt + am, data = cars)
```

We can also use `anova` to compare a base model, that only uses `am` as a predictor, and the best model that was found performing stepwise selection.

```
base_model <- lm(mpg~am, cars)
anova(base_model, best_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      30 720.90
## 2      26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the **p-value** is highly significant we reject the null hypothesis that the variables **cyl**, **hp** and **wt** do not contribute to the model.

Diagnostics

Finally we study the residual plots of the regression model and do some diagnostics. From the residual plot (see Figure 3 in the appendix) we can see that the **Residuals x Fitted** plot seems randomly scattered, what verifies the independence condition. The **Normal Q-Q** plot indicates that the residuals are normally distributed and the **Scale-Location** plot indicates constant variance. We can also notice some outliers or leverage points from the plots, let's take a look in the hatvalues and dfbetas to see if they are the same as the ones showing in the plots.

```
hat_values <- hatvalues(best_model)
df_betas <- dfbetas(best_model)
tail(sort(hat_values),3)
```

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

```
tail(sort(df_betas[,6]), 3)
```

```
## Chrysler Imperial      Fiat 128      Toyota Corona
##      0.3507458      0.4292043      0.7305402
```

We can see from the analysis above that the cars showing in the residual plots are the same as the ones found in our diagnosis.

Inference and Conclusion

By performing a t-test we can conclude that manual and automatic transmission are significantly different.

```
t.test(mpg~am, cars)
```

```
##
## Welch Two Sample t-test
##
```

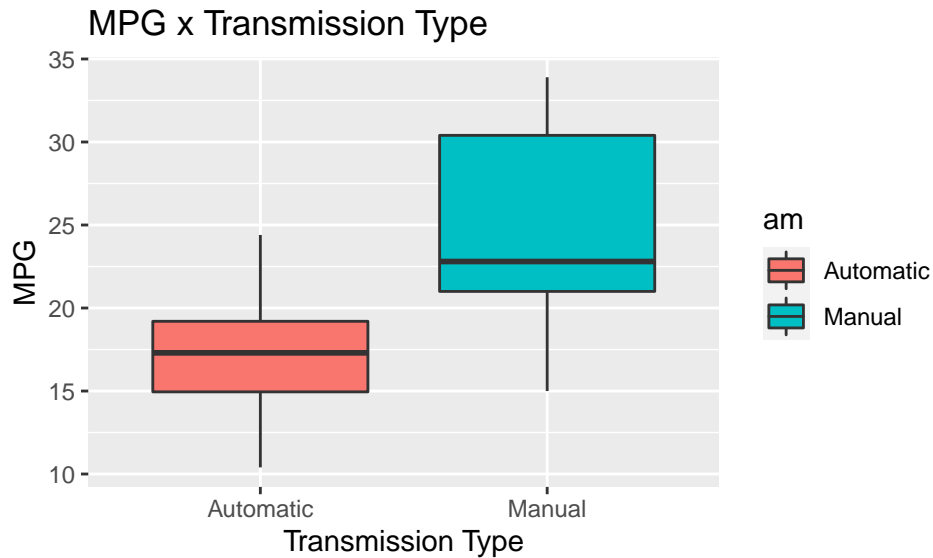


Figure 1: Boxplot between mpg and am

```
## data: mpg by am
## 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 in group Automatic    mean in group Manual
##          17.14737          24.39231
```

Let's take a look at the coefficients of our best model and draw some conclusions from it.

```
best_model$coefficients
```

```
## (Intercept)      cyl6      cyl8      hp      wt      amManual
## 33.70832390 -3.03134449 -2.16367532 -0.03210943 -2.49682942  1.80921138
```

From this coefficients we can conclude the following:

- Cars with manual transmission get more miles per gallon than cars with automatic transmission by a rate of 1.8, adjusted by number of cylinders, gross horsepower and weight.

Appendix

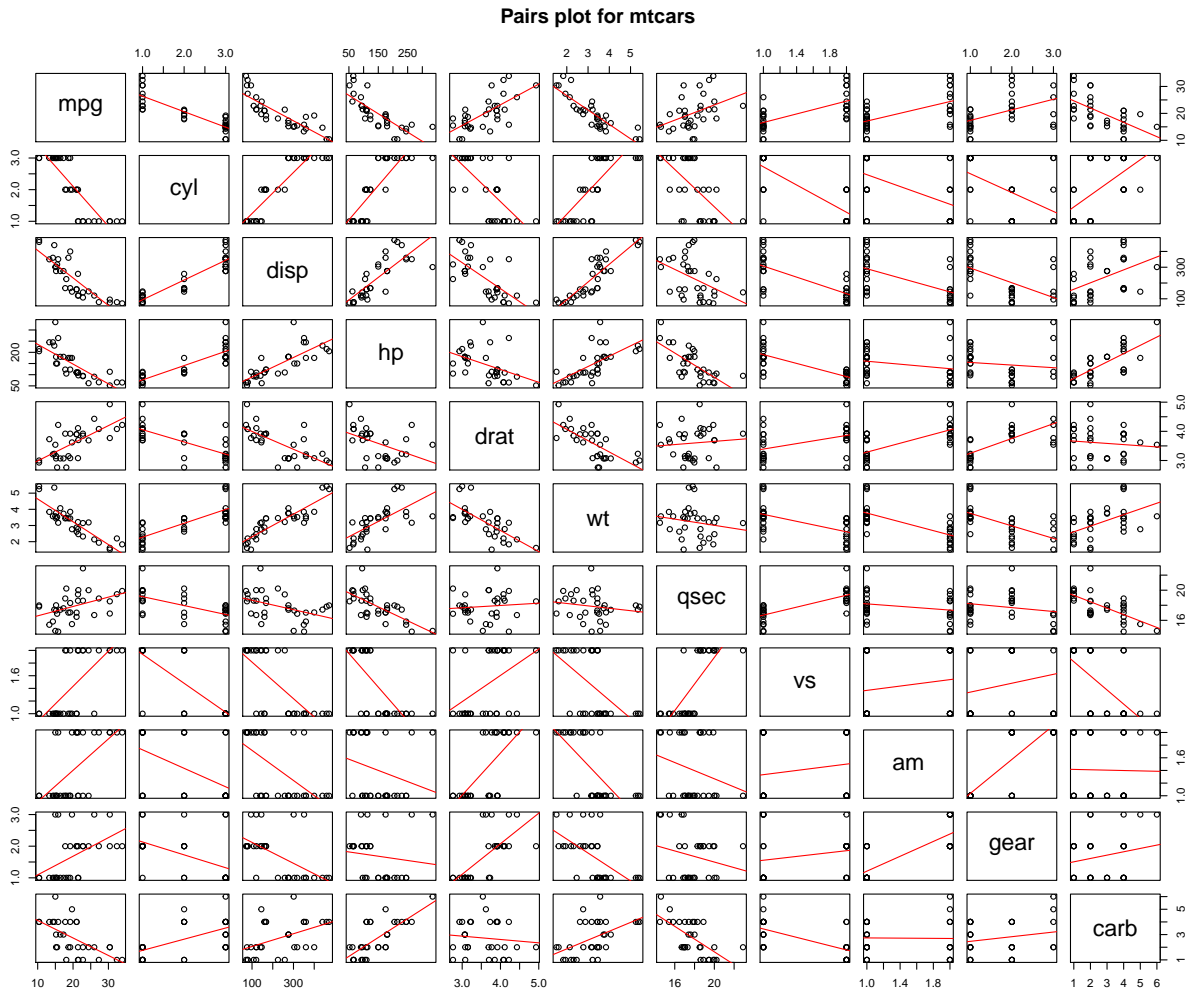


Figure 2: Pairs plot for mtcars

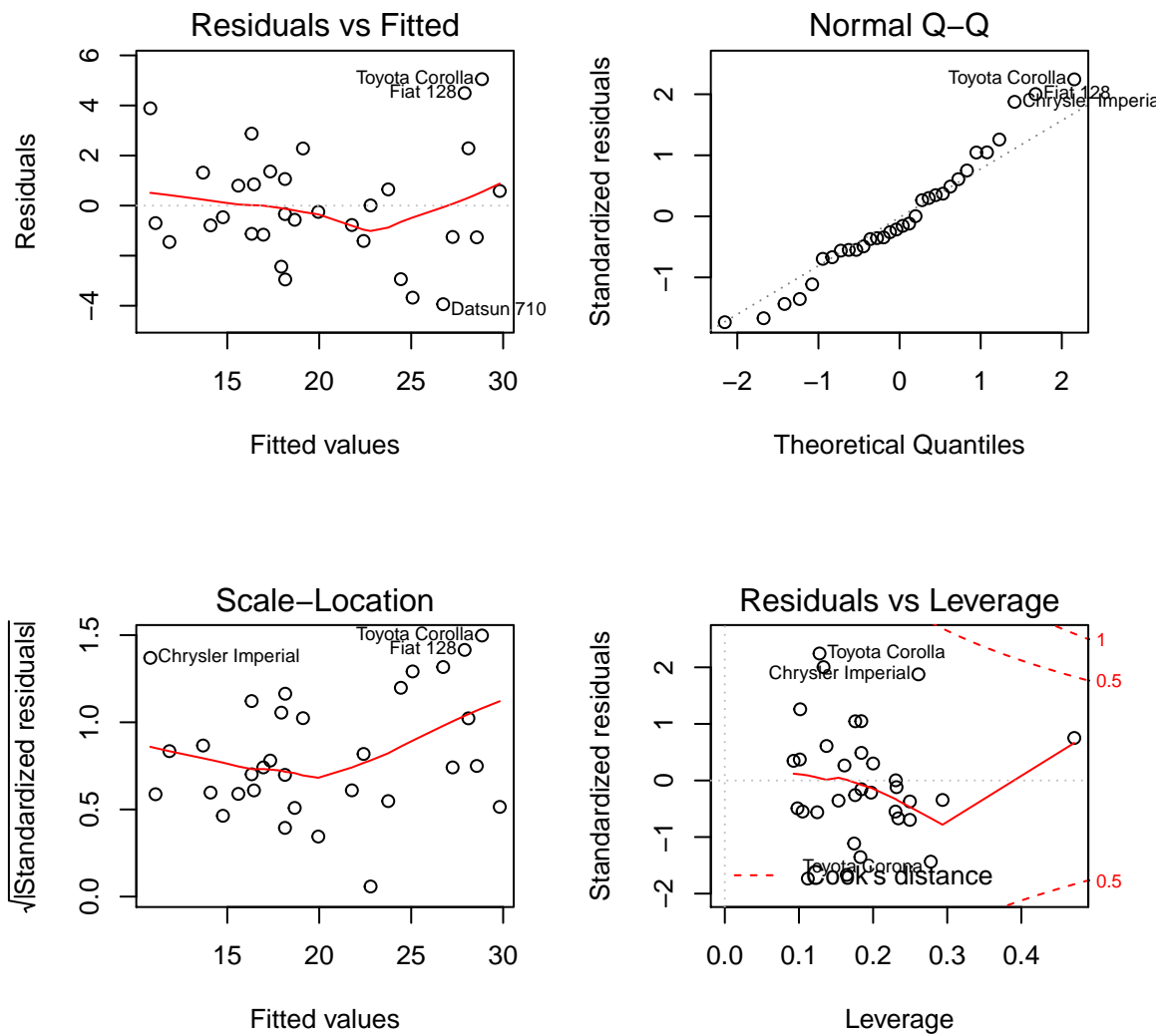


Figure 3: Residual plots