Motor Trend Car Road Tests - Effects of transmission on MPG

Executive Summary

This detailed analysis has been performed to fulfill the requirements of the course project for the course Regression Models offered by the Johns Hopkins University on Coursera. In this project, we will analyze the mtcars data set and explore the relationship between a set of variables and miles per gallon (MPG) which will be our outcome.

The main objectives of this research are as follows

- Is an automatic or manual transmission better for MPG?
- Quantifying how different is the MPG between automatic and manual transmissions?

The key takeway from our analysis was

- Manual transmission is better for MPG by a factor of 1.8 compared to automatic transmission.
- Means and medians for automatic and manual transmission cars are significantly different.

(i)Data pre processing

This is shown in appendix.

(ii)Exploratory Data Analysis

In this section, we dive deeper into our data and explore various relationships between variables of interest. Initially, we plot the relationships between all the variables of the dataset (see Figure 1 in the appendix). From the plot, we notice that variables like `cyl`, `disp`, `hp`, `drat`, `wt`, `vs` and `am` seem to have some strong correlation with `mpg`. But we will use linear models to quantify that in the regression analysis section.

Since we are interested in the effects of car transmission type on `mpg`, we plot boxplots of the variable `mpg` when `am` is `Automatic` or `Manual` (see Figure 2 in the appendix). This plot clearly depicts an increase in the mpg when the transmission is `Manual`.

(iii)Regression Analysis

In this section, we start building linear regression models based on the different variables and try to find out the best model fit and compare it with the base model which we have using `anova`. After model selection, we also perform analysis of residuals.

(iv)Model building and selection

Like we mentioned earlier, based on the pairs plot where several variables seem to have high correlation with `mpg`, We build an initial model with all the variables as predictors, and perfom stepwise model selection to select significant predictors for the final model which is the best model. This is taken care by the `step` method which runs `lm` multiple times to build multiple regression models and select the best variables from them using both **forward selection** and **backward elimination** methods by the `AIC` algorithm. The code is depicted appendix. From the above model we see that the **p-value** obtained is highly significant and we reject the null hypothesis that the confounder variables `cyl`, `hp` and `wt` don't contribute to the accuracy of the model.

(v)Residuals and Diagnostics

In this section, we shall study the residual plots of our regression model and also compute some of the regression diagnostics for our model to find out some interesting leverage points (often called as outliers) in the data set.

From the Figure 3: residual plots, we can make the following observations, * The points in the `Residuals vs. Fitted plot` seem to be randomly scattered on the plot and verify the independence condition. * The `Normal Q-Q` plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed. * The `Scale-Location` plot consists of points scattered in a constant band pattern, indicating constant variance. * There are some distinct points of interest (outliers or leverage points) in the top right of the plots. Regression diagnostics is performed to support the above results as shown in appendix

Conclusion

Based on the observations from our best fit model, we can conclude the following,

- Cars with `Manual` transmission get more miles per gallon `mpg` compared to cars with `Automatic` transmission. (**1.8** adjusted by hp, cyl, and wt).
- mpg will decrease by **2.5** (adjusted by hp, cyl, and am) for every **1000 lb** increase in `wt`.
- `mpg` decreases negligibly with increase of `hp`.
- If number of cylinders, `cyl` increases from `4` to `6` and `8`, `mpg` will decrease by a factor of `3` and `2.2` respectively (adjusted by hp, wt, and am).

Appendix

(i)Data processing and transformation

We load in the data set, perform the necessary data transformations by factoring the necessary variables and look at the data, in the following section.

```
data(mtcars)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

```
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
#str(mtcars)
```

Figure 1: Pairs Plot for mtcars dataset

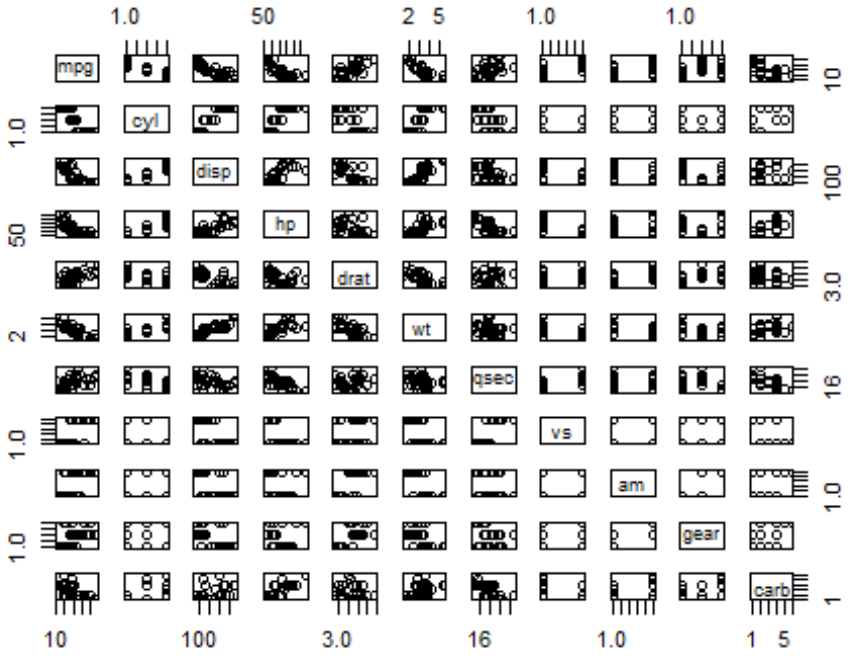
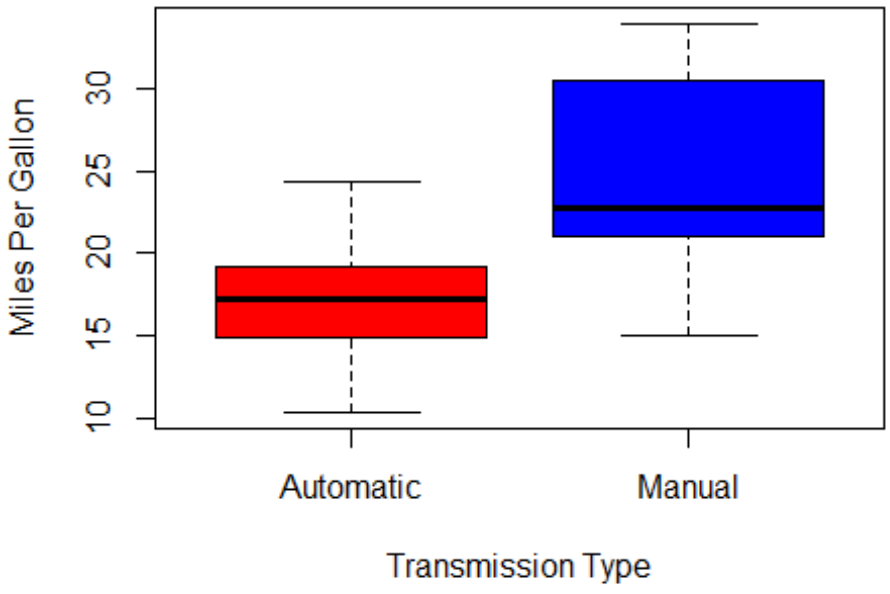


Figure 2: Car MPGs by Transmission type



(iV)Model building and selection

```
init_model <- lm(mpg ~ ., data = mtcars)
best_model <- step(init_model, direction = "both",trace=FALSE)
```

The best model obtained from the above computations consists of the variables, `cyl`, `wt` and `hp` as confounders and `am` as the independent variable. Details of the model are depicted below.

```
summary(best_model)
```

```
##
## 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
```

From the above model details, we observe that the adjusted R^2 value is **0.84** which is the maximum obtained considering all combinations of variables. Thus, we can conclude that more than 84% of the variability is explained by the above model.

In the following section, we compare the base model with only `am` as the predictor variable and the best model which we obtained earlier containing confounder variables also.

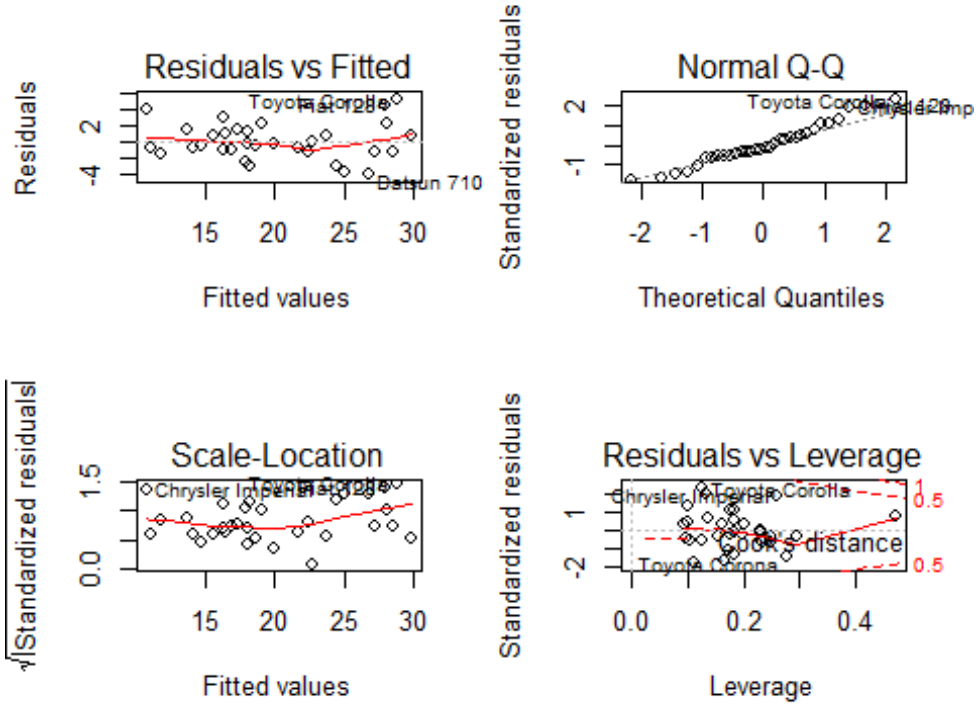
```
base_model <- lm(mpg ~ am, data = mtcars)
anova(base_model, best_model)
```

```
## 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
```

Looking at the above results, the **p-value** obtained is highly significant and we reject the null hypothesis that the confounder variables `cyl`, `hp` and `wt` don't contribute to the accuracy of the model.

(v)Residuals and Diagnostics

Figure 3: Residual Plots



Regression diagnostics is performed as shown in the following section to support the conclusions mentioned in Residuals and Diagnostics Section. We compute top three points in each case of influence measures.

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

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

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

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

Looking at the above cars, we notice that our analysis was correct, as the same cars are mentioned in the residual plots.
We also perform a t-test assuming that the transmission data has a normal distribution and we clearly see that the manual and automatic transmissions are significatively different.

```
t.test(mpg ~ am, data = mtcars)
```