Motor Trend - Auto & Manual Transmission: Which is better?

JoseCLee

24 August, 2014

Executive Summary

This report contained the analysis of relationship between transmission type and MPG of the car model, by answering the two questions below:
Is an automatic or manual transmission better for MPG?
Quantify the MPG difference between automatic and manual transmissions

The report will use `mtcars` dataset to perform the analysis.

Data Processing

The mtcar dataset for this report is from the 1974 **"Motor Trend US"** magazine, consisting of fuel consumption measurement (mpg) and 10 different aspects of automoblie design and performance for 32 automobiles (1973-74 models).

```
data(mtcars)
```

The variables of the dataset are listed below:

- **mpg** :Miles/(US) gallon
- **cyl** :Number of cylinders
- **disp** :Displacement (cu.in.)
- **hp** :Gross horsepower
- **drat** :Rear axle ratio
- **wt** :Weight (lb/1000)
- **qsec** :1/4 mile time
- **vs** :V/S
- **am** :Transmission (0 = automatic, 1 = manual)
- **gear** :Number of forward gears
- **carb** :Number of carburetors

Before further analysis, convert `am` to factor variable and label it with `automatic` and `manual` for better readability.

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

Exploratory Data Analysis

Obtain the data distribution and correlatation matrix by using the pairs() provided by R, and a custom function shown in appendix-1. Kindly refer to appendix-2 for the result.

The figure in appendix-2 above provide the information below:
`Cyl`, `hp`, `wt` and `carb` have strong negative correlation with `mpg`
`am`, the variable of the interest is positive correlated with `mpg`

A boxplot has been built to visualize the 5 summary of the data for automatic and manual tranmission type. Kindly refer to appendix-3.

From the boxplot, we can observed a clear different of fuel consumption between ***automatic*** and ***manual*** transmission type by comparing means (17.15 vs 24.39). Base on the figure above, manual transmission seems to have better fuel comsumption. However, statistical testing shall be carry to verify this finding. Since the size of the data is small, t-test is choosen as the test instructment.

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

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean in group Automatic    mean in group Manual
##                   17.15                   24.39
```

The p-value result from the testing above (0.0014) is less than the alpha, we reject the null hypothesis and claim that there is a signficiant difference in the mean MPG between car model with automatic transmission and car model with manual transmission.

Data Analysis using Regression Model

Base Model
Since `mpg` is a continuous variable, linear regression is choosing as the model to quantify the predictor. `am` is the only variable use in the base model as objective is to find our how mpg vary on different transmission type.

```r
basemodel = lm(mpg~am, data=mtcars)
summary(basemodel)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25  1.1e-15 ***
## amManual        7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

The base model only explained 36% of the variance. The finding is similiar with the results show in the boxplot, which saying that, on average, car model with manual transmission have 7.245 more mpg compare to car model with automatic transmission.

Best Model
R provide step() function to iterate each combination for the giving dataset. Using this function does help to find the best model that explain the outcome, `mpg`.

```r
bestmodel = step(lm(mpg~., data=mtcars), trace=FALSE)
```

Let't check if the new model is signigicant different from the base model

```r
anova(basemodel, bestmodel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     28 169  2       552 45.6 1.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result from the ***anova*** test is show that the *Best Model* is siginicantly different from the *Base Model*. Before the details of the new model, is a good practice to run a residuals plot and diagnostic check on the model. From the figure in Appendix-4: *Residuals plot & Diagnostics*, it shown that residuals are normally distributed and homoskedastic. Now, let's look at the details of *Best Model*:

```r
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## wt            -3.917      0.711   -5.51    7e-06 ***
## qsec           1.226      0.289    4.25  0.00022 ***
## amManual       2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
```

The *Best Model* is taking another 2 variables: `wt` and `qsec`. The new model has a multiple R-square of 0.8497, which explained approx. 85% of the variance. It is a good model to conclude the summary. Base on the coefficients of the model, it explained that: * Every 1000lbs increase on car weight would reduce 3.9165 miles per gallon.
Every second increase in quarter mile time would give extra 1.2259 miles per gallon.
Manual tramission car have 2.9358 more miles per gallon than automatic transmission car.
Since the new model is including `am` as a sigificant predicor, and showing that manual car have better mileage per gallon, the analysis can conclude as below.

Conclusion

Base on the analysis of the given data, manual transmission car do have better mileage per gallon compare to manual transmission car.

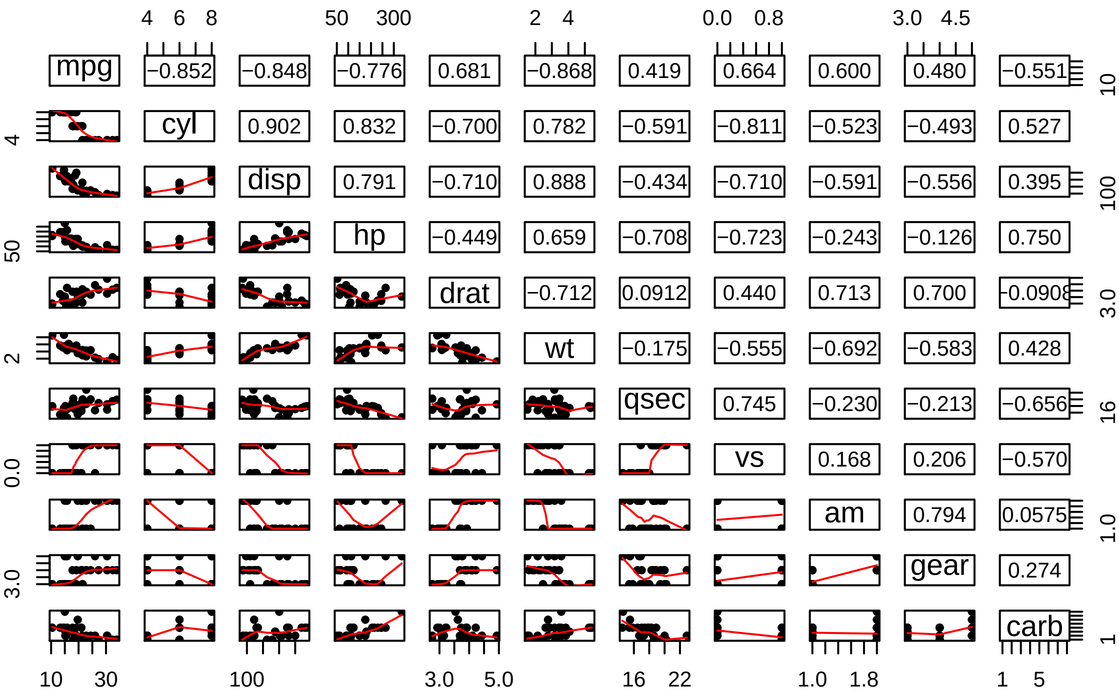
Appendix 1: Panel Correlation Function

```r
panel.cor <- function(x, y, digits=3, prefix="", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- cor(x, y)
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex.cor <- 2/strwidth(txt)
    text(0.5, 0.5, txt)
}
```

Appendix-2: Correlation Matrix with Scatterplot

```r
pairs(mpg~., data=mtcars, lower.panel=panel.smooth,
      upper.panel=panel.cor, pch=20,
      main="Motor Trend Cars Scatterplot Matrix")
```

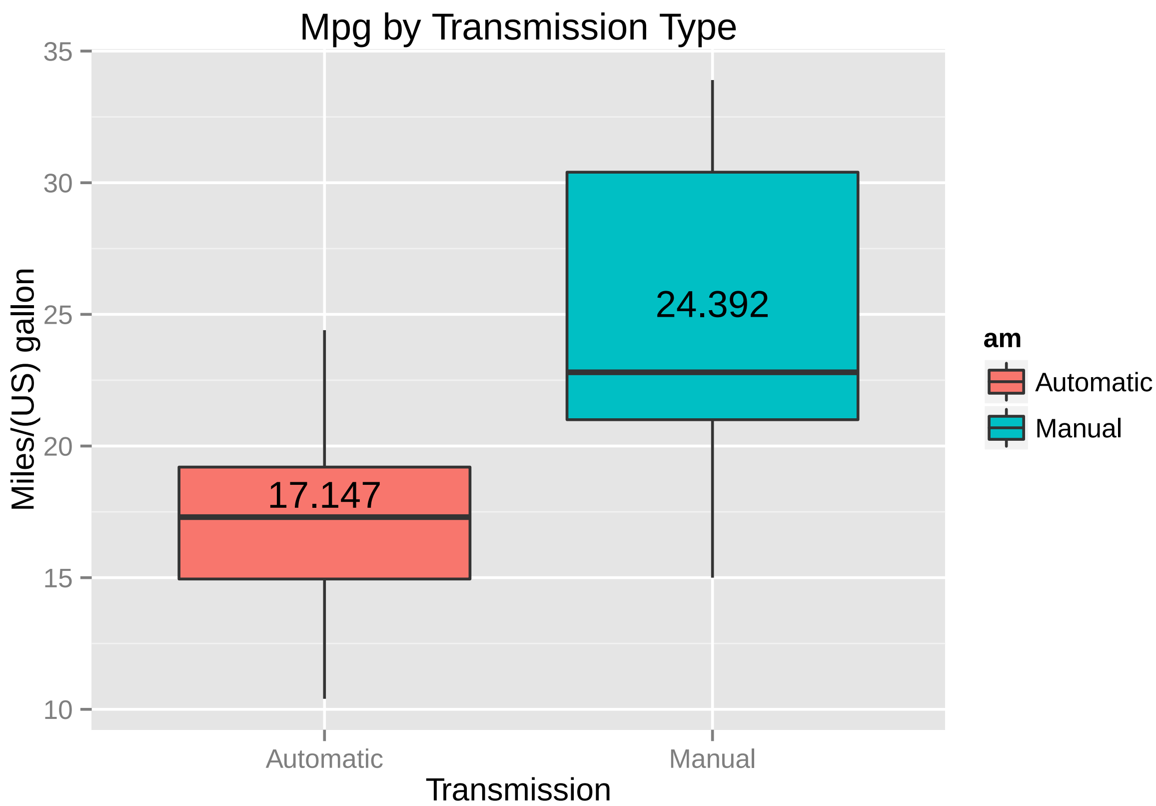
Motor Trend Cars Scatterplot Matrix



Scatterplot Matrix.pdf

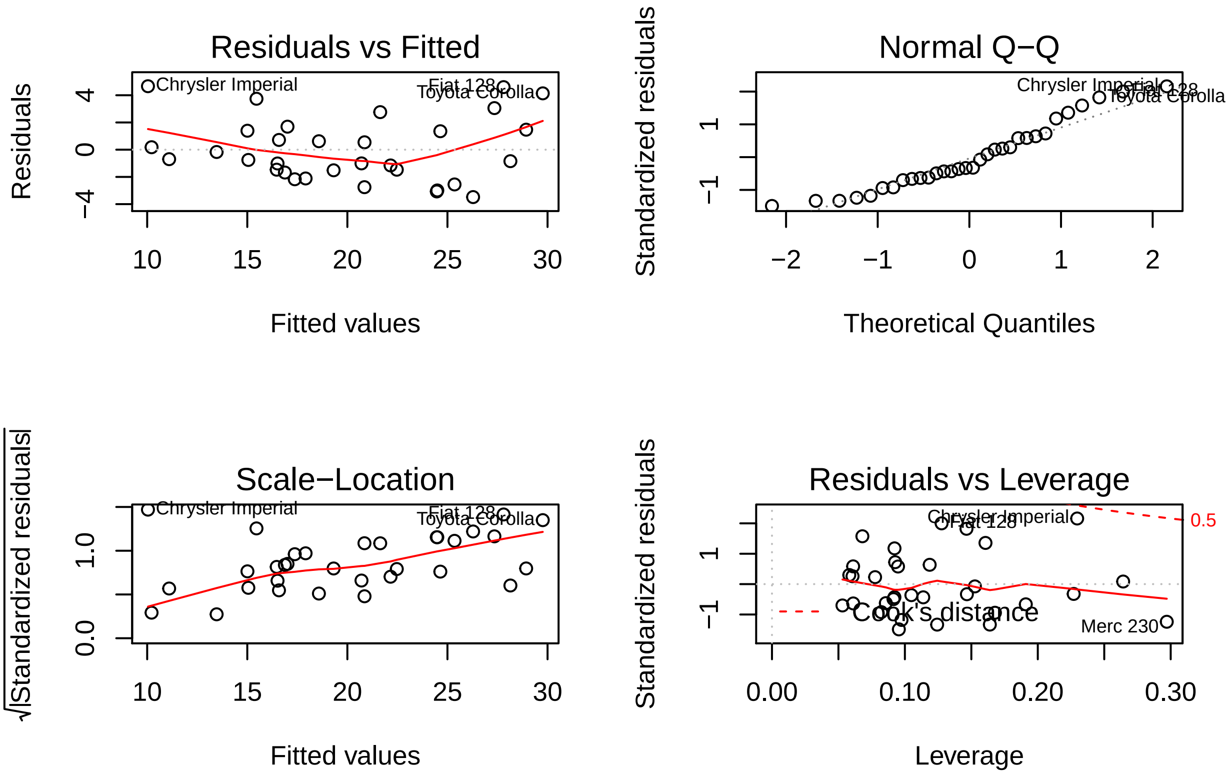
Appendix-3: Boxplot

```
library(ggplot2)
ggplot(data=mtcars, aes(x=am, y=mpg, fill=am)) +
        geom_boxplot(bin="identity") +
        xlab("Transmission") + ylab("Miles/(US) gallon") +
        ggtitle("Mpg by Transmission Type") +
        stat_summary(fun.y=mean, colour="black",
                     geom="text", show_guide = FALSE, vjust=-0.5,
                     aes( label=round(..y.., digits=3)))
```



Appendix 4: Residuals plot & Diagnostics

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


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Merc 230
0.5