# Regression Models, Course Project

*Aditya Salapaka*

*Tuesday 28 April 2015*

# Regression Models, Course Project

The objective is to explore the relationship between a set of variables and MPG, and attempt to answer the following questions:

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

## Executive Summmary

After some basic exploratory analyses, a regression model was chosen which quantified the effect of various factors in `mpg`. From statistical and regression analysis, the inference was that manual transmission is better for MPG, considering the interaction of other variables such as `cyl`, `wt` and `hp`, which significantly affect the MPG.

## About The Dataset

The `mtcars` dataset was extracted from the 1975 Motor Trend US magazine, and comprises fuel consumption and ten aspects of automobile design and performance for 32 automobiles (1973-74 models).

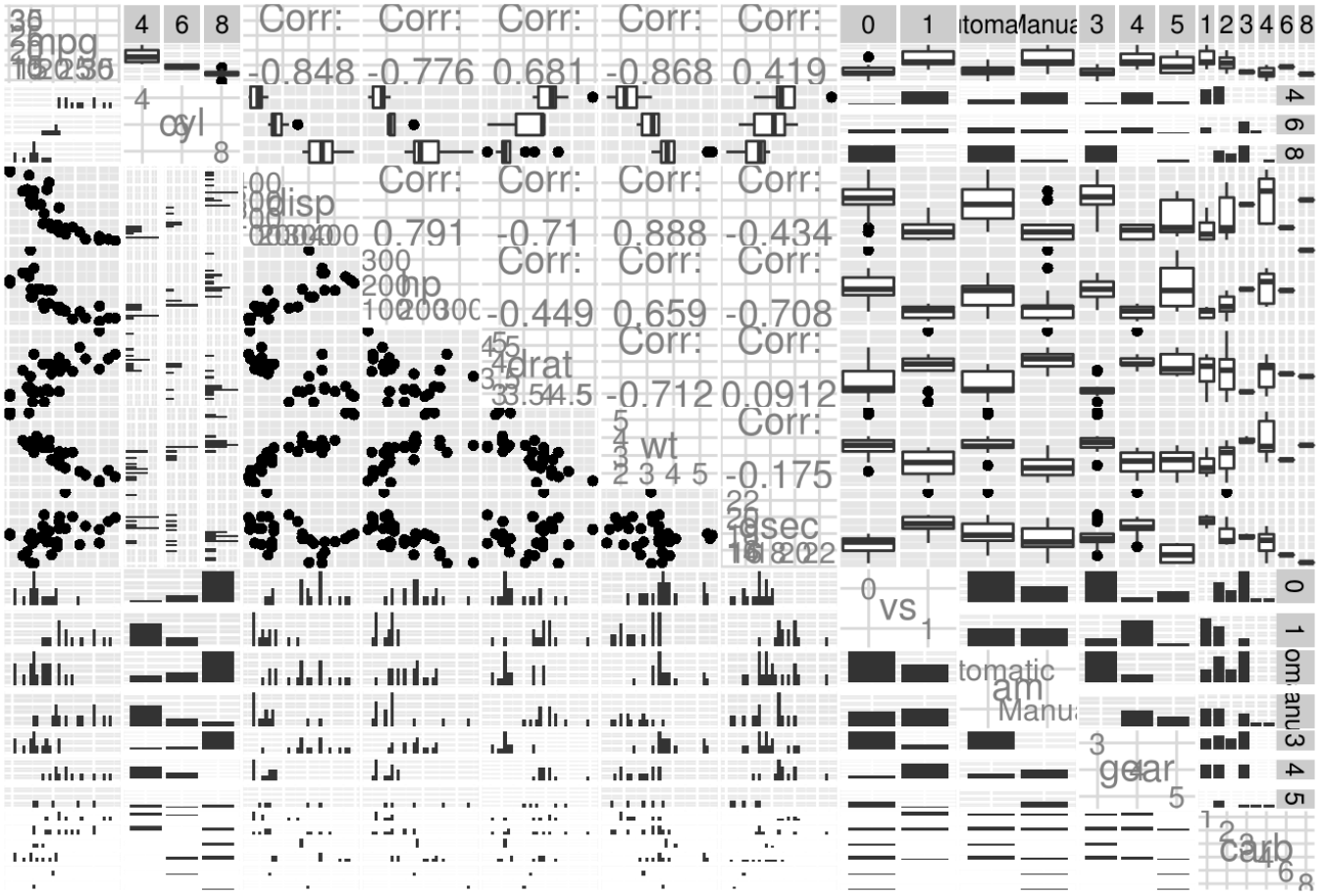
## Data Transformation

The dataset was loaded into R, and the appropriate variables were factored.

```
data(mtcars)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

## Exploratory Data Analysis

Following is a pairs plot of the `mtcars` dataset.



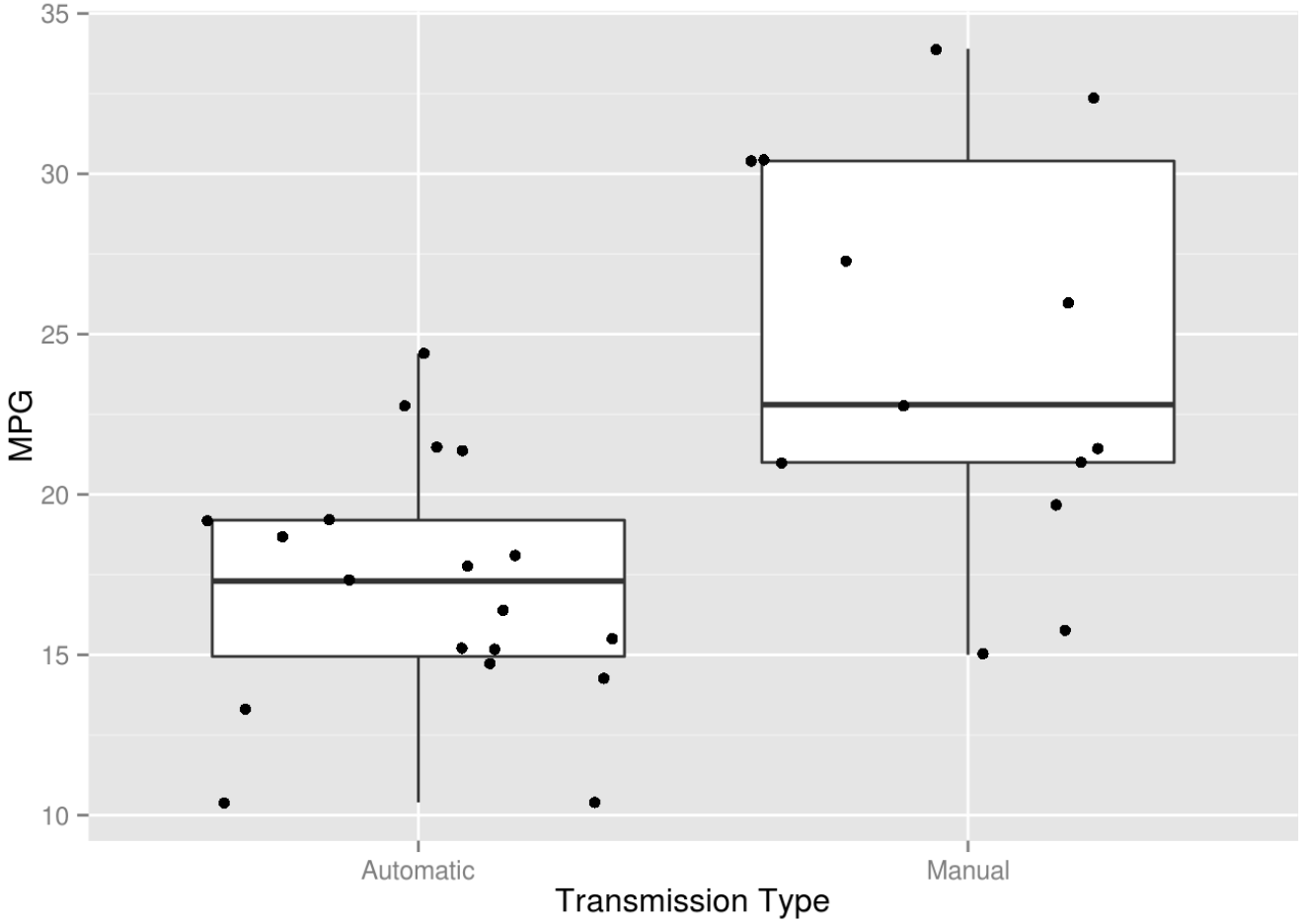

The pairs plot shows that many variables have a correlation with mpg.

Following is a boxplot between `mpg` and `am`.

```
require("ggplot2")
```

```
## Loading required package: ggplot2
```

```
g <- ggplot(mtcars, aes(am, mpg))
g <- g + geom_boxplot() + xlab("Transmission Type") + ylab("MPG") +
        geom_jitter()
print(g)
```




It is clear that manual transmission corresponds to an increase in mpg. This will be proven with a suitable regression model.

# Regression Analysis

Since it is not known which variables have a significant impact, a specific model cannot be chosen immediately. For this, a base model correlating all variables with `mpg` will be created, and the `step` function will iterate multiple regression models to choose the best one.

```
initial <- lm(mpg ~ ., data = mtcars)
best <- step(initial, direction = "both")
```

```
summary(best)
```

```
##
## 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
```

The best model has `mpg` depending on `cyl`, `hp`, `wt` and `am`. It has an adjusted R-squared value of 0.8400875, which is the highest achievable value.

To check that variables other than `am` also contribute to a change in `mpg`, the `anova` test can be run. The null hypothesis is that `cyl`, `hp`, `wt` and `am` do not affect `mpg`.

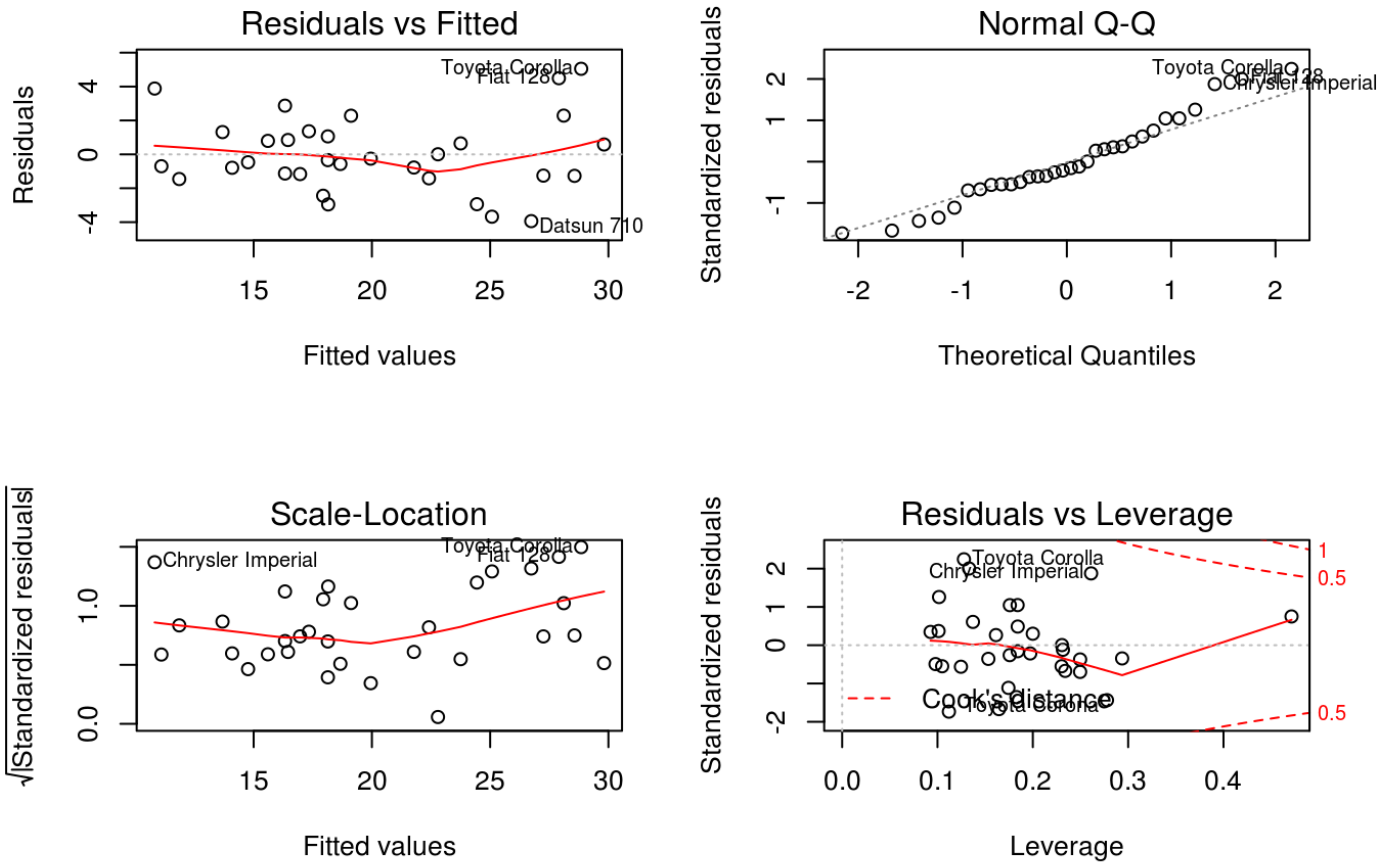
```
base <- lm(mpg ~ am, data = mtcars)
anova(base, best)
```

```
## 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
```

A p-value of 1.69e-08 helps rejects the null hypothesis.

# Residual Plot and Diagnostics

Following is a plot of the best model.




The residuals vs fitted plot is very scattered, indicative of independence. The normal Q-Q plot is a straight line, indicative of the normality of the data.

# Outlier Analysis

```
tail(sort(hatvalues(best)), 3)
```

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

```
tail(sort(dfbetas(best)[,6]), 3)
```

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

The results are the same as seen in the plot, which verifies the analysis.

# Statistical Inference

Assuming normality of the data, t-test is performed, with the null hypothesis that automatic or manual transmission have no effect on `mpg` .

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

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

A p-value of 0.0013736 helps in rejecting the null hypothesis. The difference in `mpg` based on transmission is as follows.

```
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

# Conclusion

The objective was to answer two questions:

## “Is an automatic or manual transmission better for MPG?”

**Manual Transmission is better for MPG.** The boxplot and t.test show that automatic transmission gives an `mpg` of 17.1473684, while manual transmission gives an `mpg` of 24.3923077.

## “Quantify the MPG difference between automatic and manual transmissions”

Based on the regression model obtained, it can be concluded that:

- `mpg` increases by 1.8092114 in Manual Transmission, when considering the effect of `cyl`, `hp` and `wt`.