Regression Models Course Project

Kurt Fitz

February 16, 2020

An Analysis of MPG comparing Automatic and Manual Transmissions.

Executive Summary

Using the 'cars' data set, the following provides an analysis of MPG while taking into account Manual and Automatic vehicle transmission. The goal is to answer the question of whether automatic or manual tranmissions get a better MPG rating. The results of the analyses strongly suggest that manual transmission result in a higher MPG rating then automatic transmission.

Exploratory Analyses

The appendix contains plots for the exploratory analyses. The first plot is the pairs plot for the dataset. The second plot is a boxplot showing the relationship between mpg and transmission type. From the boxplot, we can cleary see that tranmission type has an impact on mpg. It appears that automatic transmissions have a lower MPG rating the manual transmission.

```r
#Load the dependencies
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```r
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
```

```
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```r
data <- mtcars
data$am <- factor(data$am, labels=c("Automatic","Manual"))
```

Model Analysis and Diagnostics

From our exploratory analysis, we can see that manual tranmissions are related to an increase in MPG. Therefore, our hypotheses is that on average, manual transmissions have a better MPG rating then automatic. We can perform a T-test to test this hypotheses, and we see that the difference in means is statistically significant.

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

```
##
##  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
```

Next we will compare two linear models for the their ability to explain the variance in MPG. Our first model will use only transmission type as the predictor, and our second will use “hp”, “disp”, and “cyl”, in addition to tranmission type.

```r
fit <- lm(mpg ~ am, data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```r
fit2 <-lm(mpg ~ am + hp + disp + cyl, data=mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + hp + disp + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6537 -2.2037 -0.3065  1.5581  6.0095
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.476440   2.865526  10.636 3.72e-11 ***
## am           3.445269   1.453855   2.370   0.0252 *
## hp          -0.032962   0.015614  -2.111   0.0442 *
## disp        -0.007745   0.010716  -0.723   0.4761
## cyl         -0.834497   0.757091  -1.102   0.2801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.831 on 27 degrees of freedom
## Multiple R-squared:  0.8079, Adjusted R-squared:  0.7794
## F-statistic: 28.38 on 4 and 27 DF,  p-value: 2.54e-09
```

From the summary output, we see that our first model only explains about 36% of the variation observed, while the second model explains about 80% of the variance we see. Further, we compare the two models using the anova function.

```r
anova(fit, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp + disp + cyl
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     27 216.37  3    504.53 20.986 3.177e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value of 3.177e-07 shows us that our second model is significantly better at explaining variability.

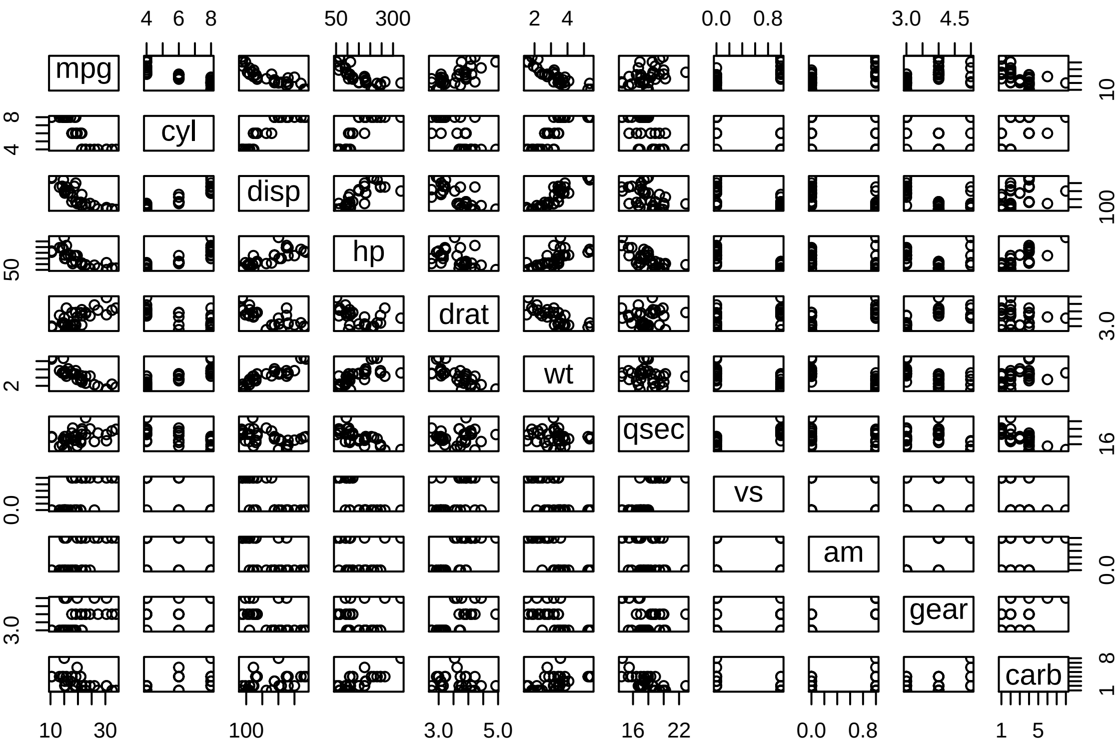
```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Appendix

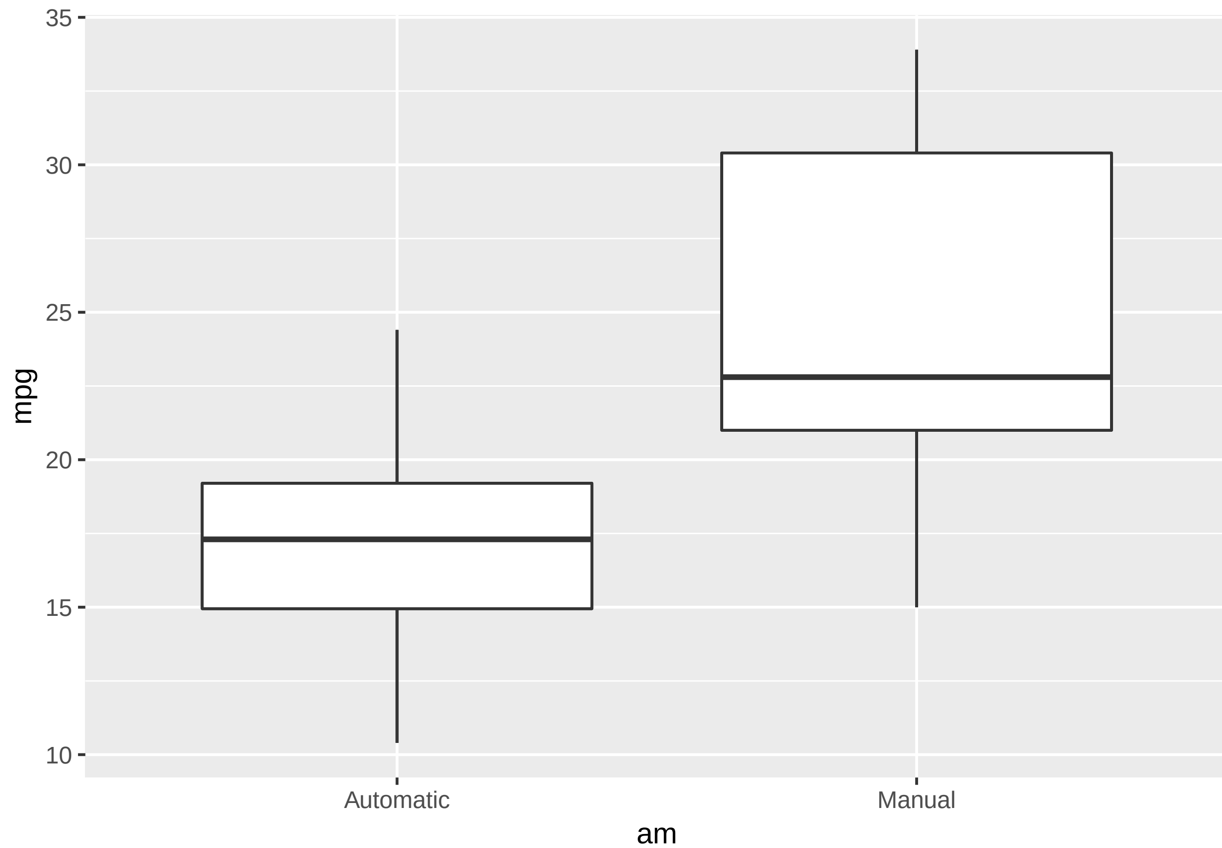
Plot 1 - Pairs plot for the data set.

```r
pairs(mpg ~ ., data=mtcars)
```



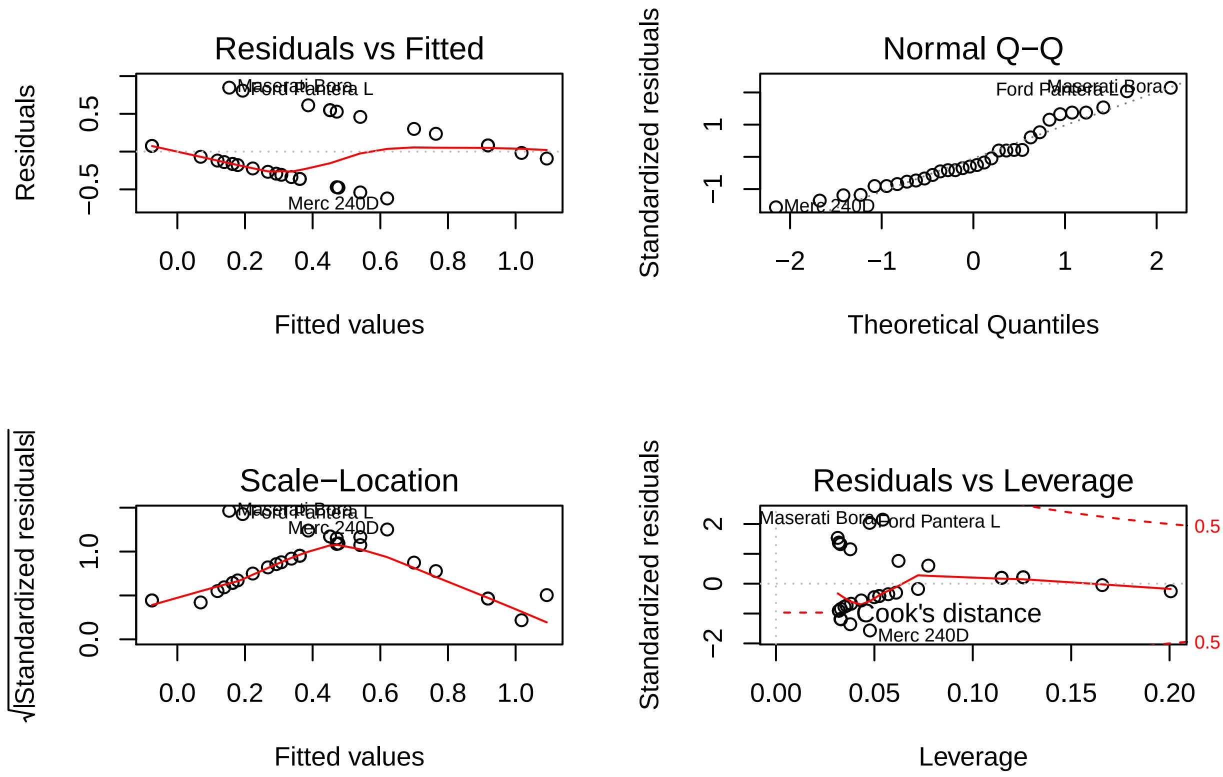
Plot 2 - Boxplot of MPG for Transmission.

```r
data <- data
ggplot(data, aes(x=am, y=mpg, ylab="MPG", xlab="Transmission Type")) + geom_boxplot(fill='#A4A4A4', colo
```



Plot 3 - Residuals Plot

```
fit <- lm(am ~ mpg, data=mtcars)
par(mfrow = c(2,2))
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
Maserati Bora
Ford Pantera L
Merc 240D
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
0.5