Regression Models Project

Xiaoqing Tan

October 30, 2018

Executive Summary

In the report, I estimated the relationship between a set of variables and miles per gallon (MPG) (outcome). I investigated into two questions specifically: 1. "Is an automatic or manual transmission better for MPG"; 2. "Quantify the MPG difference between automatic and manual transmissions".

Exploratory Analysis

```
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
```

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

# Test difference between two transmission groups
automatic <- mtcars[mtcars$am == "Automatic",]
manual <- mtcars[mtcars$am == "Manual",]
summary(t.test(automatic$mpg, manual$mpg)$p.value)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.001374 0.001374 0.001374 0.001374 0.001374 0.001374
```

From the above test, p-value < 0.05. We can tell there is a significant difference between two transmission group.

Model Fitting

Model 1

```
# Use transmission type as the only predictor
fit <- lm(mpg~factor(am), mtcars); summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)Manual    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
```

The result shows that the average MPG for automatic is 17.1 MPG, while manual is 7.2 MPG higher, which means the average MPG for manual is 24.3 MPG. The R-squared value is 0.36, which tells that this model only explains 36% of the variance.

Therefore, a model with more than one predictor is needed. The below built a multivariate model. From the Plot 2 we can see that the variables “cyl”, “disp”, “hp”, “wt” have strong correlation with the dependent variable “mpg”. So I built a new model combining these variables along with the variable “am”.

Model 2

```
fit2 <- lm(mpg~am+cyl+disp+hp+wt, mtcars); summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## amManual     1.55649    1.44054   1.080  0.28984
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## hp          -0.02796    0.01392  -2.008  0.05510 .
## wt          -3.30262    1.13364  -2.913  0.00726 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

“wt” and “am” have strong and significant correlation with “mpg”. The difference between automatic and manual transmissions is 1.56 MPG. The R-squared value is 0.8551 so the model explains 85.51% of the variance.

Check the normality and homonity of Model 2 in Plot 3.

Compare two models

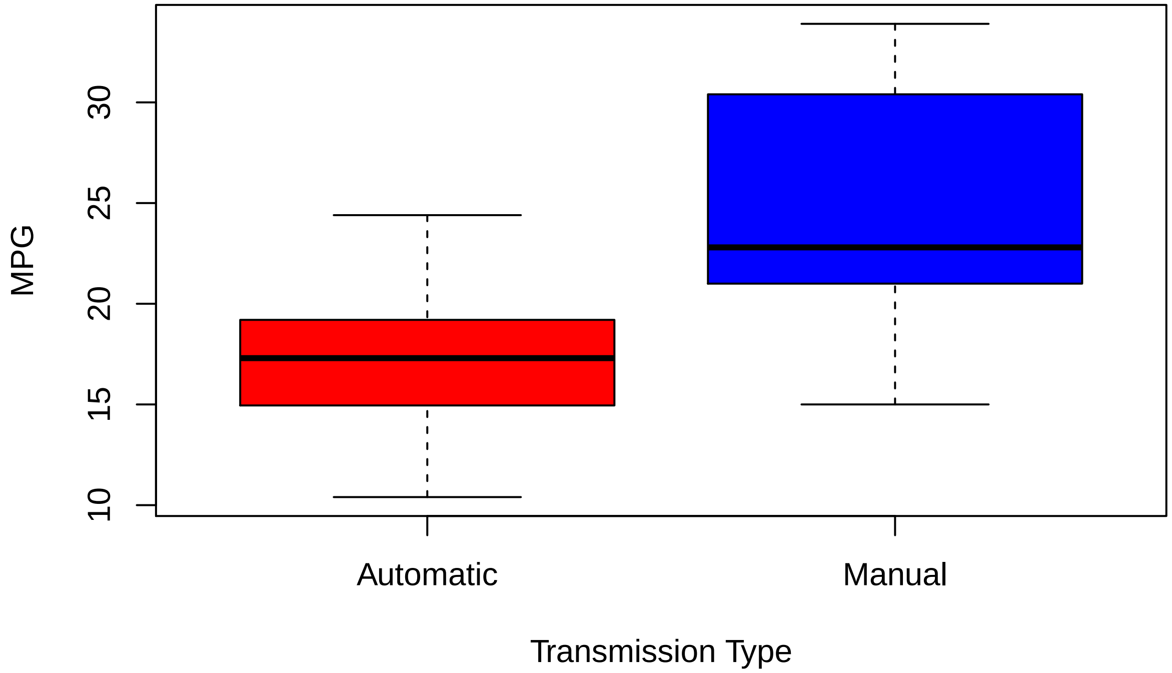
```
anova(fit,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 163.12  4    557.78 22.226 4.507e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

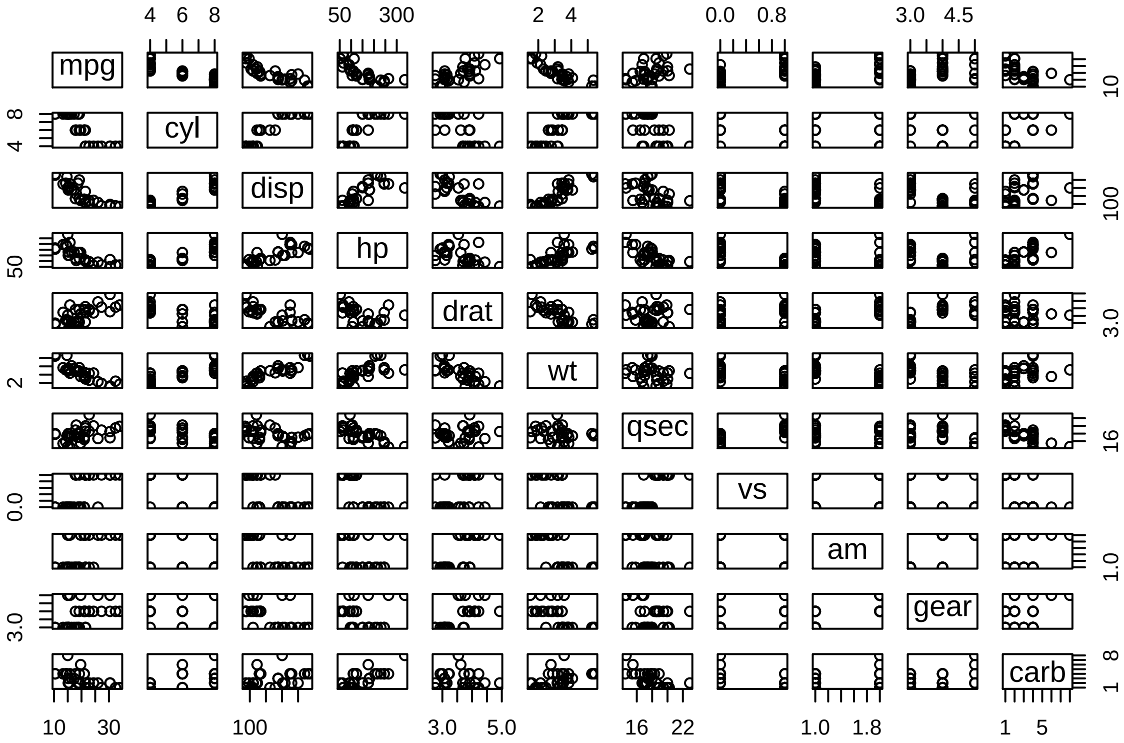
From the result we can see that the new model is significantly better than the first model, with a p-value < 0.05.

Appendix

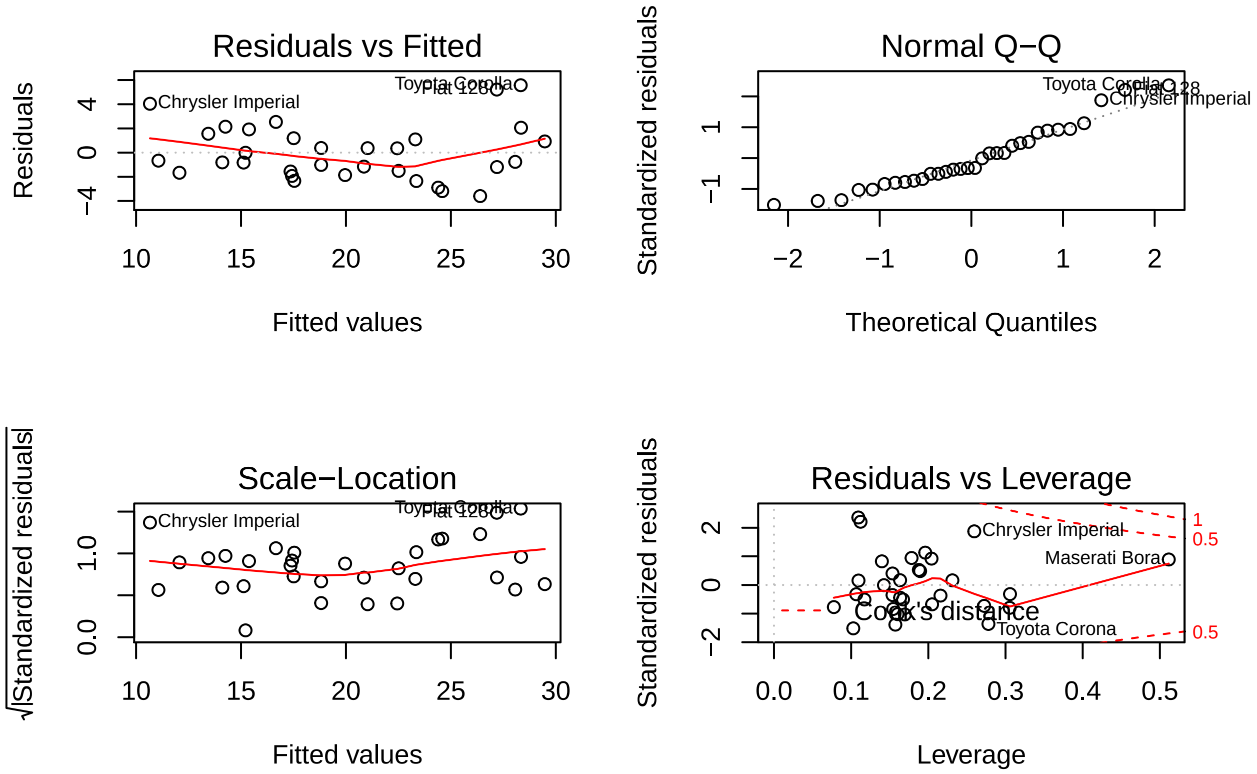
```
# Plot 1: Boxplot of MPG by Transmission tyep
boxplot(mpg ~ am, data = mtcars, col = (c("red","blue")), ylab = "MPG", xlab = "Transmission Type")
```



```
# Plot 2: Pairs plot of the variables
pairs(mpg~., mtcars)
```



```
# Plot 3: residual plots
par(mfrow = c(2,2)); plot(fit2)
```


Residuals vs Fitted
Residuals
Fitted values
−4
0
4
10
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
10
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
0.0
0.1
0.2
0.3
0.4
0.5
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5