Regression Models Project

Andres P

Regression Analysis Project For MTCARS Dataset

Project Executive Summary

The purpose of this project is to answer the following questions. First, is an automatic or manual transmission better for MPG?. Second, quantify the MPG difference between automatic and manual transmission.

First Question Answer: Manual Transmission is better than Automatic Transmission for MPG.

Second Question Answer: Automatic Transmission gives a car 15+ MPG. Manual Transmission give a car 20+ MPG.

Load The Dataset MTCARS

```r
library(datasets);data(mtcars)
```

Exploratory Data Analysis

See in Appendix (Plot 1) data(mtcars) str(mtcars)

mtcars data frame contains 32 observations with 11 variables. We want to examine what factors effect MPG. The correlation and coefficients between MPG and remaining 10 variables are calculated along with a scatterplot matrix of specific variables in the Appendix.

```r
require(stats)
round(cor(mtcars)[-1, 1], 2)
```

```
##   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

This shows many variables have moderate correlation with MPG because their coefficients are greater than .5. In addition, the 10 variables are correlated as well. This can be demontrated by comparing number of cylinders and other variables.

```r
require(stats)
round(cor(mtcars)[-2,2],2)
```

```
##   mpg  disp    hp  drat    wt  qsec    vs    am  gear  carb
## -0.85  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
```

For this project we are focusing on the variable transmission and the difference betwen automatic = am0 and manual = am1. To perform the analysis you plot the variable MPG against the variable AM. This can be seen below in the Appendix.

Regression Model Selection

The exploratory data analysis show that this is a multivariable regression problem and many of the variables correlate to eachother. According to the correlation coefficients, wt, cyl and disp have the strongest correlations with MPG. Now we have to compare them using models.

```r
fit1 <- lm(mpg ~ am, data = mtcars)
## Model 1: MPG vs auto or manual transmission
fit2 <- lm(mpg ~ am + wt + cyl + disp, data = mtcars)
## Model 2: MPG vs weight + number of cylinders + displacement
fit3 <- lm(mpg ~ wt + hp + cyl + disp + am, data = mtcars)
## Model 3: MPG vs weight + number of cylinders + displacement + transmission
fit4 <- lm(mpg ~ ., data = mtcars)
## Model 4: MPG vs all variables
```

For this project we care about the effect of automatic or manual transmission = AM on MPG. So we fit MPG with transmission only.

See in Appendix (Plot 2) summary(fit1)

Coefficients show manual transmission MPG increase by 7.245 MPG. In addition, the p value is $< .05$ which means the difference for manual transmission is large. However, the adjusted R-squared is .3385 which means this information could be biased without looking at other variables. Now we compate MPG with number of cylinders, weight and displacement which have corellation coefficients with MPG.

See in Appendix (Plot 3) summary(fit2)

Adjusted R-squared is now .8147. P value indicates that number of cylinders and weight have linear relationships with MPG, but displacement does not. Lets now compare models with more variables.

See in Appendix (Plot 4) anova(fit2, fit3, fit4)

The p values show adding additional variables is not necessary for this specific analysis. Instead a different model is needed where we compare weight and number of cylinders to MPG.

See in appendix (Plot 5).

Conclusion of Project

The analysis in this project shows that manual transmission is better than automatic transition for MPG. In addition, MPG is related to vehicle weight and number of cylinders.

Appendix

Plot 1

```r
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Plot 2

```r
summary(fit1)
```

```
##
## 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
```

Plot 3

```r
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + disp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## am           0.129066   1.321512   0.098  0.92292
## wt          -3.583425   1.186504  -3.020  0.00547 **
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## disp         0.007404   0.012081   0.613  0.54509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

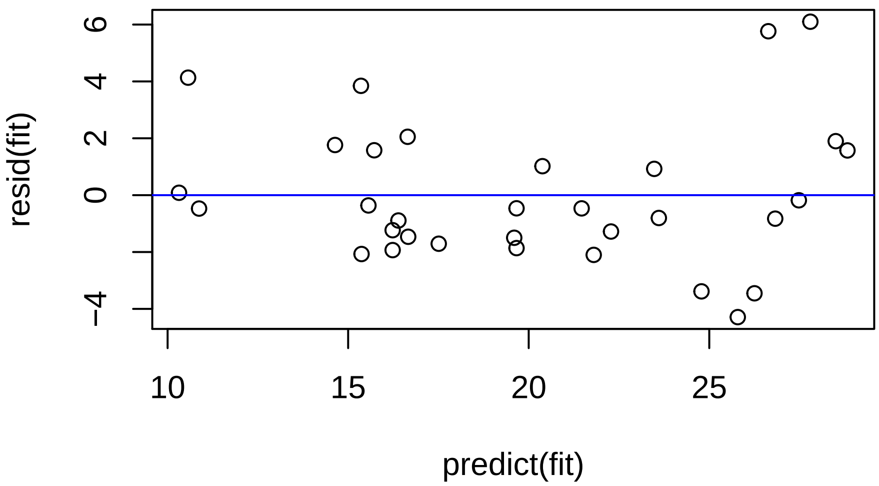
Plot 4

```r
anova(fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + wt + cyl + disp
## Model 2: mpg ~ wt + hp + cyl + disp + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 188.43
## 2     26 163.12  1    25.306 3.6030 0.07151 .
## 3     21 147.49  5    15.625 0.4449 0.81206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

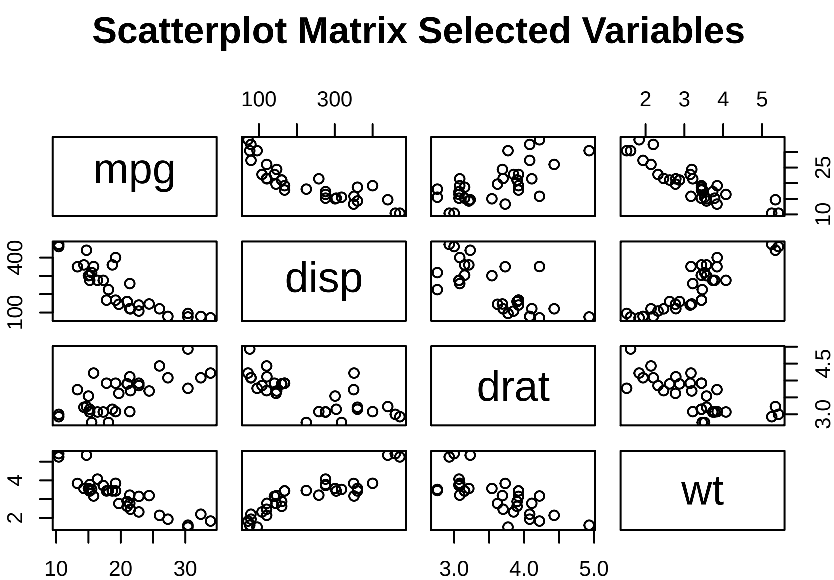
Plot 5

```r
fit <- lm(mpg ~ wt + cyl, data = mtcars)
plot(predict(fit), resid(fit))
abline(h = 0, col = "blue")
```



Scatterplot Matrix

```r
pairs(~mpg+disp+drat+wt,data=mtcars,
   main="Scatterplot Matrix Selected Variables")
```



Boxplot of MPG by AM

```
boxplot(mpg~am,data=mtcars, main="MPG Relationship To Transmission Type",
   xlab="Transmission Type(0=Automatic, 1=Manual)", ylab="Miles Per Gallon")
```