Motor Trend Exploratory Data Analysis

Xinyu W

Overview

This project analyze the data set "mtcars", which was extracted from *Motor Trend* US magazine for motor trend car road tests and contains 32 observations on 11 variables, and explore the relationship between a set of variables and miles per gallon (MPG) (outcome). The project works around two questions:

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

Executive Summary

With t-test, we can see a significant lead of manual transmission over auto type by a difference *7.245*. But that difference is adjusted to *1.80* including confounding variables ('cyl', 'disp', 'wt', and 'hp'). These comparison can also be found in plot appendix at last.

Data Analysis

Load Data

```
data("mtcars"); head(mtcars, n=2)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4       21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag   21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

In this project, we will mainly work on "mpg" and "am" variables. Using **?mtcars**, we can find "am" stands for Transmission (0 = automatic, 1 = manual). Therefore, we better factor this variable and some other variables in order to find the difference between/among them later.

```
mtcars$am <- factor(mtcars$am, labels = c("automatic", "manual"))
mtcars$vs <- factor(mtcars$vs); mtcars$cyl <- factor(mtcars$cyl)
mtcars$gear <- factor(mtcars$gear); mtcars$carb <- factor(mtcars$carb)
```

1. "Is an automatic or manual transmission better for MPG"
Referred to *Appendix*, we find there's a seemingly obvious increase in the distribution of MPG for manual transmission. In addition to the **Plot 1**, here we attempt to do a comparison numerically. The mean of automatic transmission (17.15) is lower than that of manual data (24.39), **which means that manual transmission cars preform better than automatic ones with regard to mpg.**

```r
aggregate(mpg~am, mtcars, mean)
```

```
##          am      mpg
## 1 automatic 17.14737
## 2    manual 24.39231
```

2. "Quantify the MPG difference between automatic and manual transmissions"
From above, we can calculate that **the difference is 7.245 in favor of manual transmission cars.** However, to see whether the difference is significant, we should do a *hypothetical test*, and here we choose t-test.
H0: The difference between transmission types (0 and 1) is 0.
H1: The difference between transmission types (0 and 1) is not equal to 0.

```r
hypoTest <- t.test(mpg~am, mtcars); hypoTest$p.value
```

```
## [1] 0.001373638
```

From above, we see **p-value** is far less than .05, which allows us to *reject H0.* The difference (Again, **7.245**) is significant.

```r
regression1 <- lm(mpg ~ am, data=mtcars);summary(regression1)
```

```
##
## 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 ***
## ammanual       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
```

By running the regression model between 'mpg' and 'am', we see the *R-squared* is 0.36, which means 'am' alone only explains 36% of the variance. **Confounding variable** exist.

3. Regression Analysis – Confounding Variables
We first need to decide which one/ones of them has/have significant correlation.

```r
confoundTest <- aov(mpg~., data=mtcars); summary(confoundTest)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
```

```
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By doing the regression analysis, we find **'cyl', 'disp', 'wt', and 'hp'** stand out for their correlation besides 'am', thus we can start adjusting our regression model.

```r
regression2 <- lm(mpg~am+cyl+disp+wt+hp, data=mtcars)
anova(regression1, regression2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     25 150.41  5    570.49 18.965 8.637e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(regression2)
```

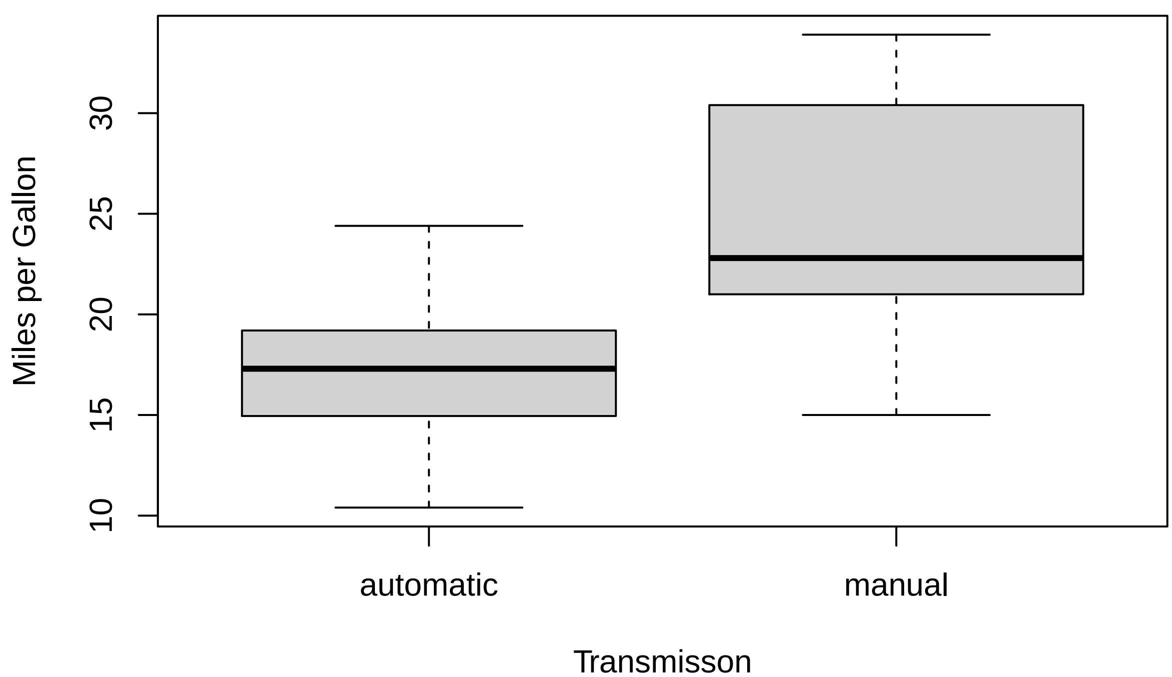
```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## ammanual     1.806099   1.421079   1.271   0.2155
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## wt          -2.738695   1.175978  -2.329   0.0282 *
## hp          -0.032480   0.013983  -2.323   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

Regression model 2 here clearly explains the relationship better than the previous one through our variance table. With **Plot 2**, we can see the "Residual vs Fitted" plot shows little variance. It explains **86% (R-squared)** of the variance. With these four variables included, the manual transmission advantage falls to **a 1.80 difference** compared to auto transmission.

Appendix

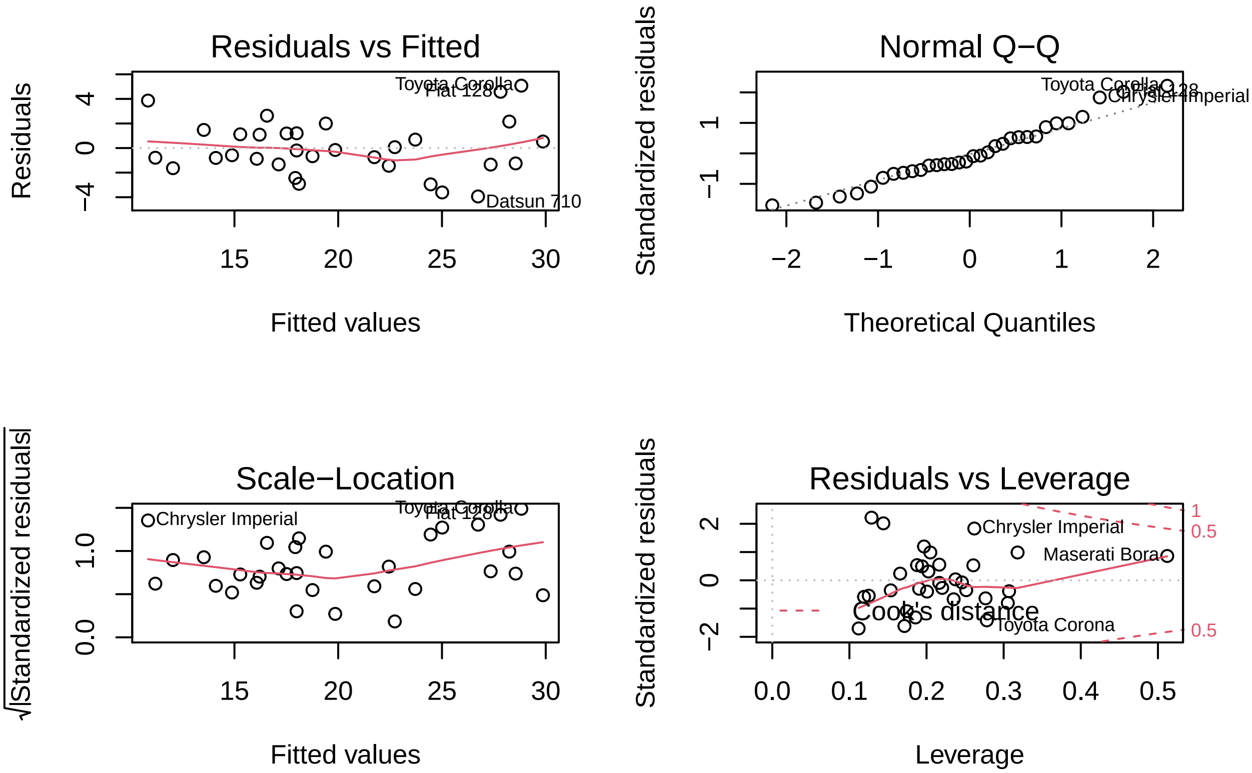
Plot 1 - "Boxplot of MPG by Transmission Types"

```
boxplot(mpg~am, data = mtcars, xlab = "Transmisson", ylab = "Miles per Gallon")
```



plot2 - Regression Model Analysis with Confouding Data

```
par(mfrow = c(2,2))
plot(regression2)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Maserati Bora
Toyota Corona
Cook's distance
1
0.5