Regression Models project - Motor Trend Data Analysis Report

Hsin-Wen Chang

Saturday, July 25, 2015

Executive Summary

In this report, I analyze `mtcars` data set which was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles .By Using regression models explain how **automatic** (am = 0) and **manual** (am = 1) transmissions features affect the **mpg** feature. The t-test shows the performance difference between cars with automatic and manual transmission. By fitting several linear regression models and select the one with highest adjusted R-squared value. Given that weight and a quarter mile time are hold constant, manual transmitted cars are 14.079 + (-4.141)*weight more mpg (miles per gallon) on average better than automatic transmitted cars. Thus, cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher MPG values.

Exploratory Data Analysis

First,loading the data set `mtcars` and change some variables from `numeric` class to `factor` class.

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.0.3
```

```r
data(mtcars)
mtcars[1:3, ] # Sample Data
```

```
##                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
```

```r
dim(mtcars)
```

```
## [1] 32 11
```

```r
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Inference

By making a hypothesis assuming the MPG has a normal distribution. Using the two sample T-test.

```
result <- t.test(mpg ~ am)
result$p.value
```

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

```
result$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Regression Analysis

Fit the full model as the following:

```
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel)
```

The Residual standard error as 2.833 on 15 degrees of freedom.The Adjusted R-squared value is 0.779 means that the model can explain about 78% of the variance of the mpg variable, none of the coefficients are significant at 0.05 significant level.

```
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel)
```

Scatter Plot of mpg vs. Weight by Transmission

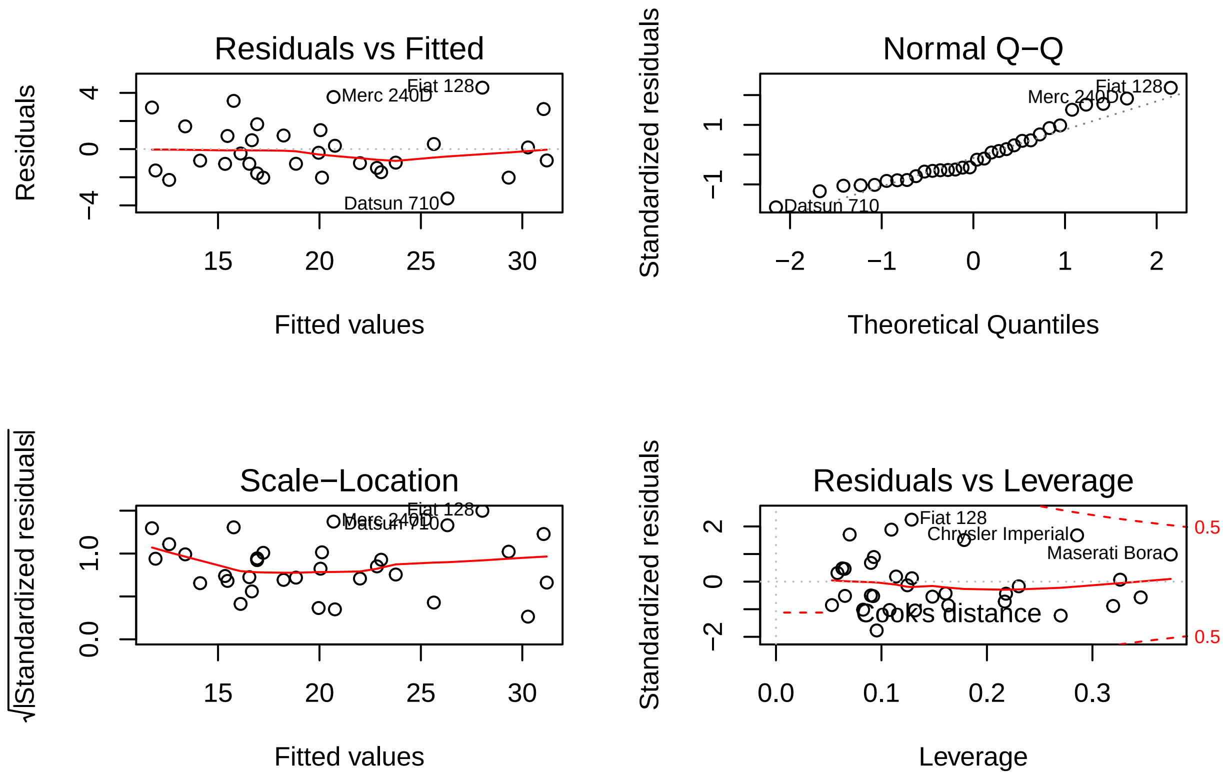
```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("Scatter Plot of mpg vs. Weight by Transmission")
```

Step model is "mpg ~ wt + qsec + am". It has the Residual standard error as 2.459 on 28 degrees of freedom. And the Adjusted R-squared value is 0.8336 means the model can explain about 83% of the variance of the mpg variable. All of the coefficients are significant at 0.05 significant level.The scatter plot indicates there appears to be an interaction term between "wt" variable and "am" variable. since automatic cars tend to weigh heavier than manual cars. We have the following model including the interaction term:

```
amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(amIntWtModel) # results hidden
```

Residual Plots

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

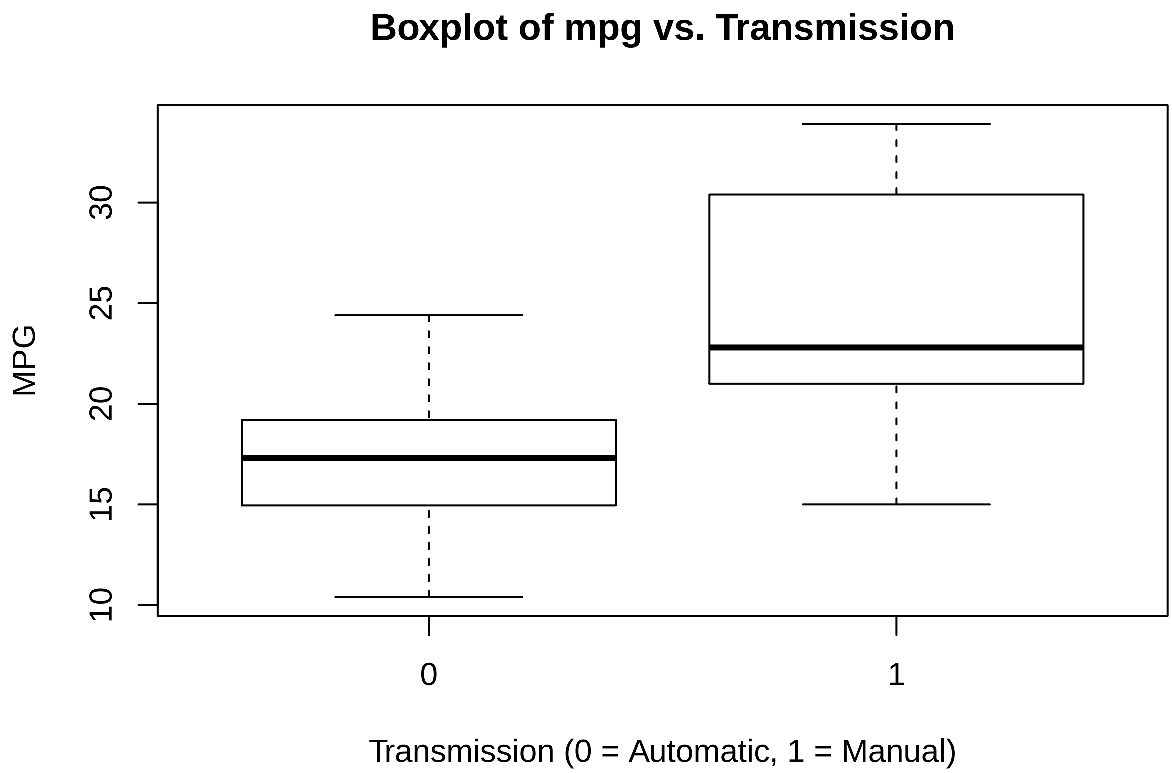


The Residual standard error as 2.084 on 27 degrees of freedom. And the Adjusted R-squared value is 0.8804 means the model can explain about 88% of the variance of the mpg variable. All of the coefficients are significant at 0.05 significant level.

```r
amModel<-lm(mpg ~ am, data=mtcars)
summary(amModel) # results hidden
```

Boxplot of mpg vs. Transmission

```r
boxplot(mpg ~ am, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="MPG",
        main="Boxplot of mpg vs. Transmission")
```



It shows on average, a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased.The Residual standard error as 4.902 on 30 degrees of freedom and the Adjusted R-squared value is 0.3385 means the model can explain about 34% of the variance of the mpg variable. The low Adjusted R-squared value also indicates the needs to add other variables to the model.
The final model:

```
anova(amModel, stepModel, fullModel, amIntWtModel)
confint(amIntWtModel)
```

Selecting the model with the highest Adjusted R-squared value, "mpg ~ wt + qsec + am + wt:am".

```
summary(amIntWtModel)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

The result shows that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant. Cars with manual transmission add 14.079 + (-4.141)*wt more mpg (miles per gallon) on average than cars with automatic transmission.

Residual Analysis and Diagnostics

According to the residual plots, we can verify the following underlying assumptions:

1. The Scale-Location plot confirms the points are randomly distributed.
2. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.
3. The Residuals vs. Fitted plot shows no consistent pattern supporting the accuracy of the independence assumption.

4. As for the Dfbetas, the measure of how much an observation has effected the estimate of a regression coefficient as follow:

```
sum((abs(dfbetas(amIntWtModel)))>1)
```

```
## [1] 0
```

The above analyses meet basic assumptions of linear regression and answer the questions.