Regression Model of the ‘mtcars’ data

Julian Jang

November 19th, 2015

Summary

This analysis was performed for the ‘mtcars’ dataset. By looking at a data set of a collection of cars, I was interested in exploring the relationship between a set of variables and MPG(miles per gallon) as outcome. I was particularly interested to explore.

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

In order to figure out the answers of these questions, I have performed exploratory data analysis, and used hypothesis test and linear regression as methodologies to make inference.

Explanations of variables

At first, I have looked the meaning of ‘mtcars’ dataset variables over help pages by the command ‘? mtcars’. The meanings of the variables are below.

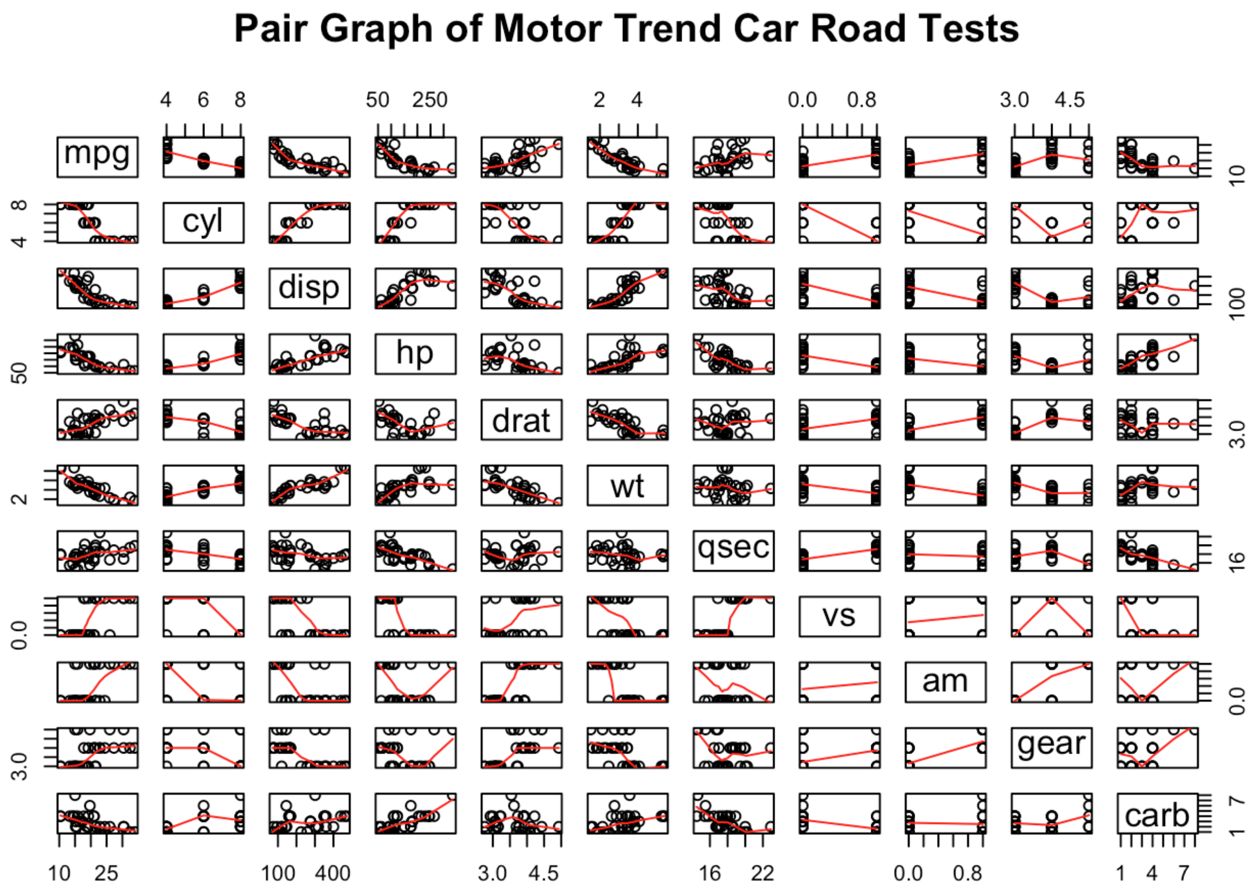
column No.	Variable Name	Explanation
[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (lb/1000)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

I found out that ‘mpg’ might be used as the response variable and the ‘am(transmission)’ is the factor variable.

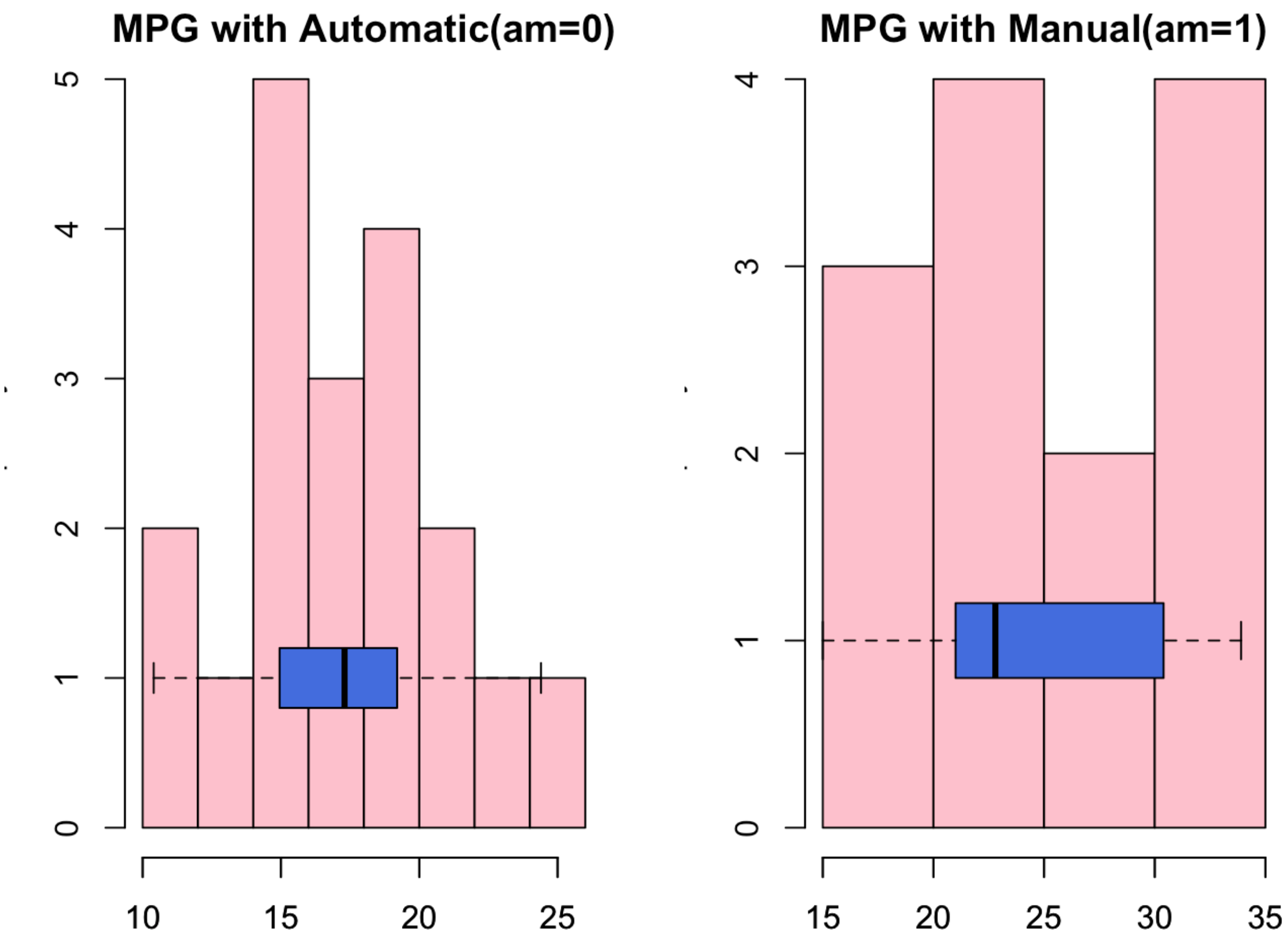
EDA for ‘mtcars’ dataset.

I've started the analysis by performing initial exploratory data analysis(EDA) to get a better idea of the existing patterns between variables in the data set. Below is that I created pairwise scatter plots. This is a nice way to investigate the relationship between all the variables in this data set. And, I also created histograms of 'mpg' according to am type.

```
pairs(mtcars, panel = panel.smooth, main="Pair Graph of Motor Trend Car Road Tests")
```




```
par(mfrow = c(1, 2), mar=c(3.1, 3.1, 1.1, 2.1))
hist(mtcars$mpg[mtcars$am == 0], col = "pink", main = "MPG with Automatic(am=0)")
boxplot(mtcars$mpg[mtcars$am == 0], horizontal=T,  outline=T, frame=F, col = "royalblue", add =
T)
hist(mtcars$mpg[mtcars$am == 1], col = "pink", main = "MPG with Manual(am=1)")
boxplot(mtcars$mpg[mtcars$am == 1], horizontal=T,  outline=T, frame=F, col = "royalblue", add =
T)
```



Also, I have compared the average of 'mpg' between 'Automatic' and 'Manual'. The result of it showed me They have difference of average. So, I needed to analyze by t.test.

```
with(mtcars, tapply(mpg, am, mean))
```

```
##        0        1
## 17.14737 24.39231
```

Statistical Inference(t test)

I have performed the t.test with my hypothesises below.

- Null hypothesis(HO): The 'mpg' is not different according to type of 'transmission'.
- Alternative hypothesis(H1): The 'mpg' is different according to type of 'transmission'.

```
with(mtcars, t.test(mpg ~ am))
```

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

The result of ‘t.test’ show me that the ‘p-value’ is ‘0.0014’ in 95% confidence interval. So the two groups are not same each other.

Regression Model

1. Regression modeling with only the ‘am’ variable

```
mtcars$am <- factor(mtcars$am)
summary(lm(mpg ~ am, data = mtcars))
```

```
##
## 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 ***
## am1            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 coefficents of the result is ‘0.0003’ of ‘am’ factor and ‘1.13e-15’ of a intercept. And, ‘p-value’ is ‘0.0003’ but the adjusted R^2 is ‘0.3385’. So, I can say that the realation of the variables between ‘mpg’ and ‘am’ are meaningful. But, this model is not enough to explain the relations between ‘mpg’ and other variables.

2-1. Regression modeling with the stepwise method

I have performed regression modelling with all variales by stepwise method.

```
summary(step(lm(mpg ~ ., data = mtcars), trace = 0, steps = 10000))
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

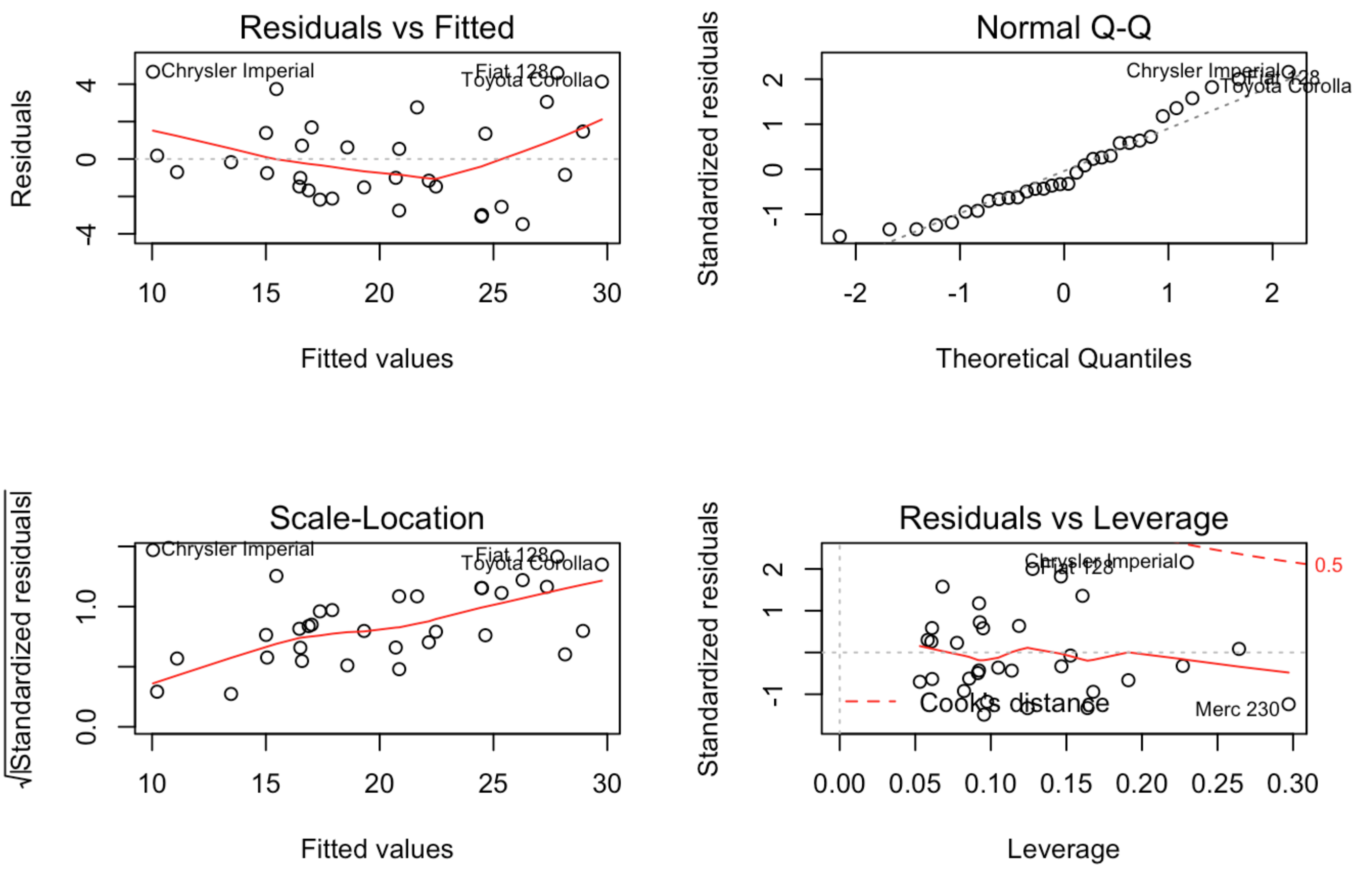
The 'p-value' and the adjusted R^2 was each '1.21e-11' and '0.8336'. I think this model is nice. And, I have looked over the conefficients of the variables. Among five variables, only 'wt' and 'qsec' have significant meaning in 99% confidence interval. And, the coefficients of 'wt' and 'qsec' are '-3.9165' and '1.2259'.

Conclusion

This model is "mpg = wt + qsec + am". The Adjusted R-squared value is 0.8336. And, all of the coefficients are significant at 0.05 significant level. Then, I have tried the 'Residual Diagnostics', the result is that residuals are met with the normal distribution, the independence, the random distribution and no outliers.

Appendix

```
par(mfrow = c(2, 2))
plot(step(lm(mpg ~ ., data = mtcars), trace = 0, steps = 10000))
```



Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5