Regression Models Report

Wien Wong

Saturday, February 21, 2015

Abstract

In this report, a dataset ‘mtcars’ collected from the 1974 Motor Trend US magazine was used to explore whether an automatic or a manual transmission have different impact on miles per gallon depletion. And how much is the difference between these two transmissions? Besides, an optimal estimate of multivariate regression model was tried to established, and is finally validated by diagnostics plots.

Data Process and Analysis

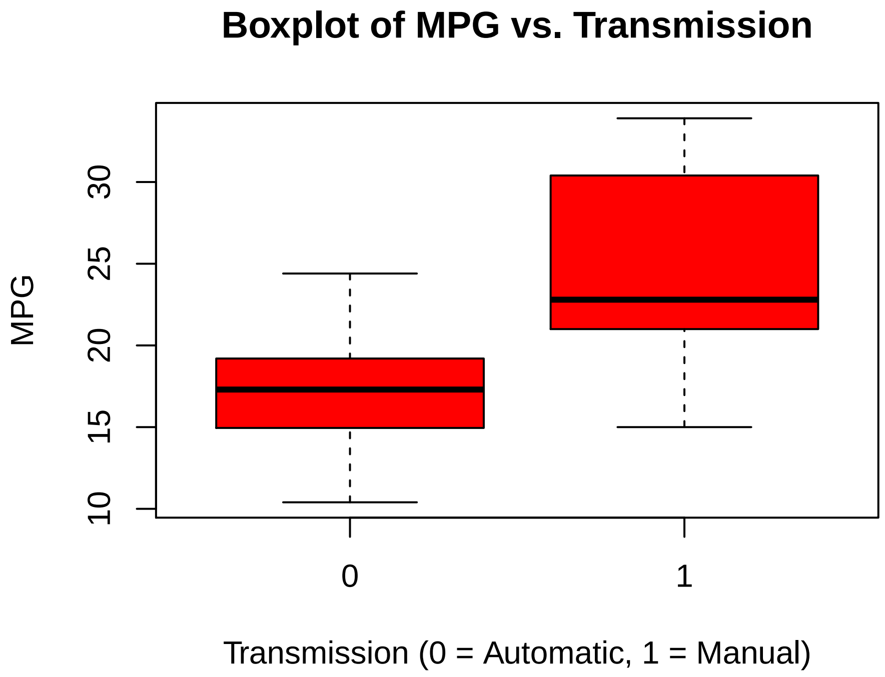
```
data(mtcars)
```

The eleven variables are explained as follows:

- `mpg`: Miles per US gallon
- `cyl`: Number of cylinders
- `disp`: Displacement (cubic inches)
- `hp`: Gross horsepower
- `drat`: Rear axle ratio
- `wt`: Weight (lb / 1000)
- `qsec`: 1 / 4 mile time
- `vs`: V/S
- `am`: Transmission (0 = automatic, 1 = manual)
- `gear`: Number of forward gears
- `carb`: Number of carburetors

Let’s try a boxplot to view if there is any difference between the two transmissions.

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



As can be seen that manual transmission shows a better performance regarding MPG in general. To confirm the difference, a t-test is performed with the null hypothesis being that there is no difference in the mean MPG for automatic and manual transmission.

```
t.test(mpg~am,data=mtcars)$p.value
```

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

The p-value is far less than 0.05 thus the null hypothesis is rejected. There indeed is difference between the two transmissions regarding MPG.

There are 10 predicted variables and some must play minor roles to MPG consumption. Thus analysis of variances is necessary.

```
summary(aov(mpg ~ ., data = mtcars))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
## disp         1   37.6    37.6   5.353  0.03091 *  
## hp           1    9.4     9.4   1.334  0.26103    
## drat         1   16.5    16.5   2.345  0.14064    
## wt           1   77.5    77.5  11.031  0.00324 ** 
## qsec         1    3.9     3.9   0.562  0.46166    
## vs           1    0.1     0.1   0.018  0.89317    
## am           1   14.5    14.5   2.061  0.16586    
## gear         1    1.0     1.0   0.138  0.71365    
## carb         1    0.4     0.4   0.058  0.81218    
## Residuals   21  147.5     7.0                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

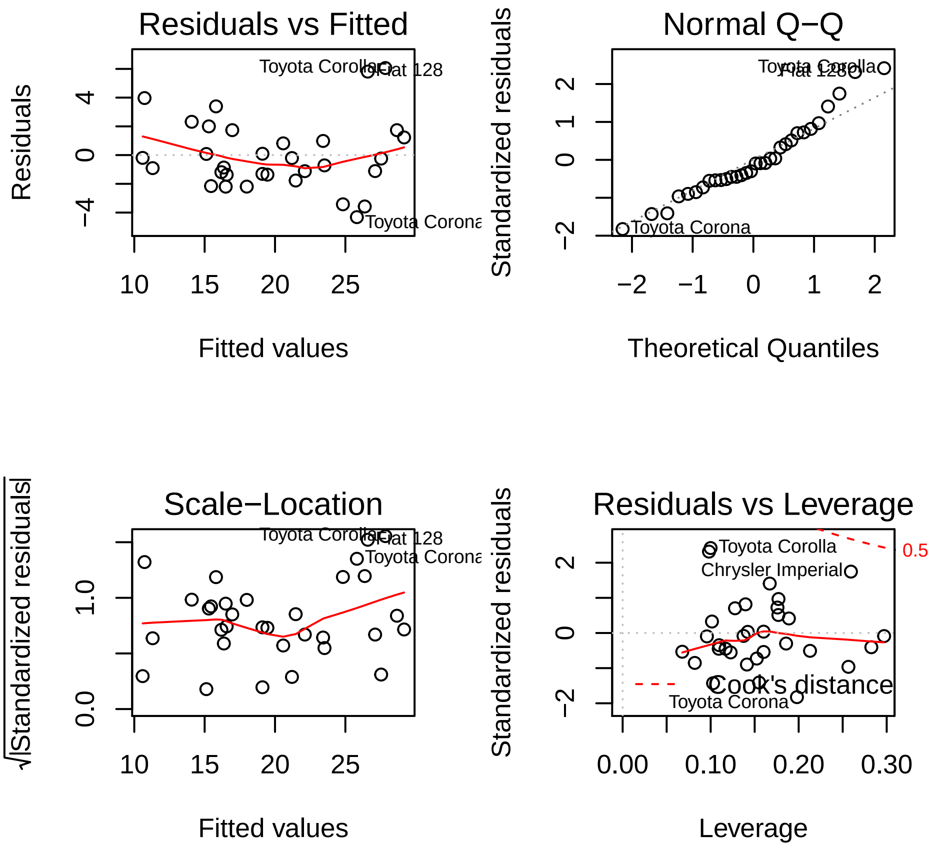
Variables with p-value less than 0.5 are more important. Thereby, 4 fit models are proposed.

```
fit1 <- lm(mpg ~ cyl + wt, data = mtcars)
fit2 <- lm(mpg ~ cyl + wt + disp, data = mtcars)
fit3 <- lm(mpg ~ cyl + wt + am, data = mtcars)
fit4 <- lm(mpg ~ cyl + wt + disp + am, data = mtcars)
```

Summary of the 4 fit models are attached in the appendix. The largest adjusted r-squared value among the four models is 0.8327, corresponding to the `fit4` model. It indicates that the `fit4` model can explain up to 83.27% of the total variation. Therefore, `fit4` is selected as the final multivariate model.

Last, residual diagnostic plot is performed.

```
par(mfrow = c(2,2))
plot(fit4)
```



The Residuals vs Fitted plot shows residuals against fitted values. If any pattern is apparent in the points on this plot, then the linear regression model may not be suitable in this case. The Normal Q-Q plot indicates the residuals are normally distributed. The Residuals vs Leverage plot refers to the standardized residuals against leverage. The standardized residuals are centered around zero. On this plot, the red smoothed line stays close to the horizontal gray dashed line and that no points have too much leverage (a large Cook's distance).

It is concluded that weight, displacement and number of cylinders play important role on miles per US gallon.

Appendix

```
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## cyl          -1.5078     0.4147  -3.636 0.001064 **
## wt           -3.1910     0.7569  -4.216 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + disp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4035 -1.4028 -0.4955  1.3387  6.0722
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.107678   2.842426  14.462 1.62e-14 ***
## cyl         -1.784944   0.607110  -2.940  0.00651 **
## wt          -3.635677   1.040138  -3.495  0.00160 **
## disp         0.007473   0.011845   0.631  0.53322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.595 on 28 degrees of freedom
## Multiple R-squared:  0.8326, Adjusted R-squared:  0.8147
## F-statistic: 46.42 on 3 and 28 DF,  p-value: 5.399e-11
```

```
summary(fit3)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ cyl + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## wt           -3.1251     0.9109  -3.431  0.00189 **
## am            0.1765     1.3045   0.135  0.89334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

```r
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + disp + am, 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 ***
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## wt          -3.583425   1.186504  -3.020  0.00547 **
## disp         0.007404   0.012081   0.613  0.54509
## am           0.129066   1.321512   0.098  0.92292
## ---
## 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
```