Regression Model Analysis on mtcars

Author: Byron.Bian

Date: Date: June 20, 2015

Overview: In this report, I will explore the relationships of multi-variates within the dataset of mtcars by virtue of technique of regression modeling and then answer two questions in which magzine of Motor Trend is particularly interested. ***

Phase I: Construct Multi-variates Linear Regression Model

In order to fit a parsimonious linear model, firstly I need to explore the diagnostic information of a full-variates model(Regress outcome-MPG to all predictor variables)

```
fit.lm<-lm(mpg~.,data=mtcars)
lm.coef<-summary(fit.lm)$coef
lm.coef[,4]
```

```
## (Intercept)         cyl        disp          hp        drat 
    wt 
##  0.51812440  0.91608738  0.46348865  0.33495531  0.63527790  0.06
325215 
##        qsec          vs          am        gear        carb 
##  0.27394127  0.88142347  0.23398971  0.66520643  0.81217871
```

We can see that all P-Values are greater than .05, none of variables are significant, so this model is not ideal. And I can remove those redudant variates by using R function called Step.

```
fit.lm<-step(fit.lm,direction="backward")
```

Here, step function chose the most parsimonious model in terms of minimum AIC value (61.31), and we can check the regression statistics

```
summary(fit.lm)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

Now, we can see all three predictor variates are significant and R-square is ideal. Furthermore, variate of [am] is of category, so I need to change into factor and re-fit the model

```
mycars<-mtcars[,c(1,6,7,9)]
mycars$am<-as.factor(mycars$am)
fit.lm<-lm(mpg~.,data=mycars)
summary(fit.lm)
```

```
## 
## Call:
## lm(formula = mpg ~ ., data = mycars)
## 
## 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
```

Now, the linear muti-variates regression model can be expressed like as below

mpg=9.6178 - 3.9165wt + 1.2259qsec + 2.9358(am="manual")

Phase II: Interpretation

According to the model I constructed, the two questions magzine of Motor Trend is interested in can be answered.

*Is an automatic or manual transmission better for MPG?
Because the coefficient of am1 is 2.9358>0, and mpg means Miles/(US) gallon (Miles per gallon could drive). Then we know that Manual Transmission is better for MPG

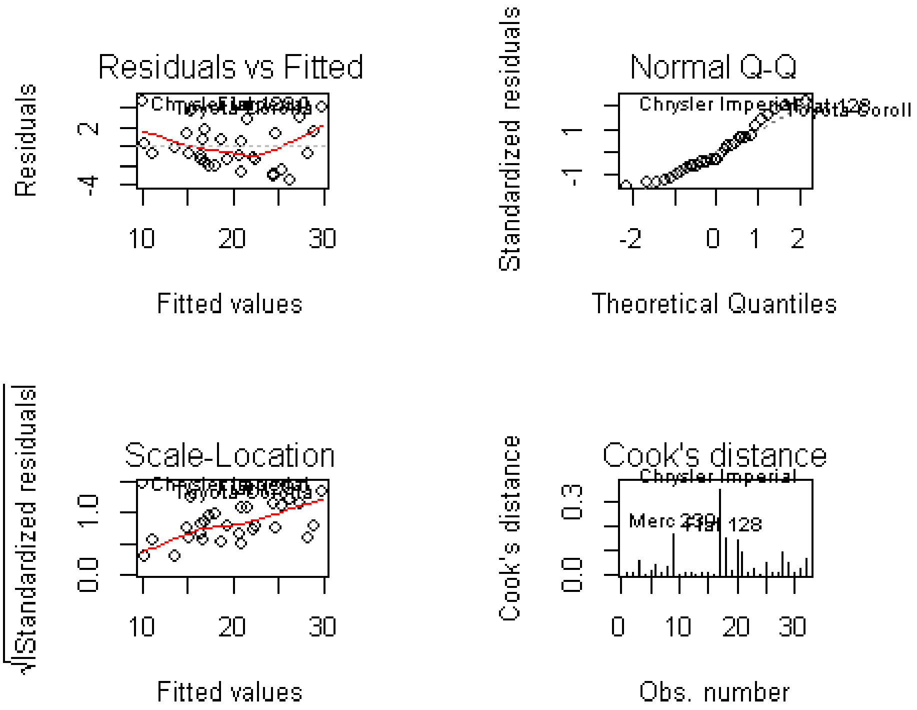
*Quantify the MPG difference between automatic and manual transmissions
Given other variables are keeping un-changed, the Miles/gallon of manual transmission are about 2.9358 more than automatic transmission.

By the way, qsec has also positive coefficient. So given other variates are constant, each increase of qsec will lead to 1.2559 accruement of mpg. As for wt, under qsec and am are unchanged,each increase of 1t will lead to 3.9165 decrease of mpg.

Appendix

Figures of Regression Model Diagnostics

```
par(oma=c(0,0,0,0),mai=c(1,1,.5,.5),mfrow=c(2,2));
plot(fit.lm,which=1)
plot(fit.lm,which=2)
plot(fit.lm,which=3)
plot(fit.lm,which=4)
```



```
#normality of residuals by shapiro test
fit.res<-residuals(fit.lm)
shapiro.test(fit.res)

##
##  Shapiro-Wilk normality test
##
## data:  fit.res
## W = 0.9411, p-value = 0.08043
```

1 From fitted-value vs residuals, phenomenon of heteroscedasticity isn't obvious
2 From Normal Q-Q plot & shapiro test, we can check the normality of residuals
3 From standardized residuals & cook-distance plot, Chrysler Imperial may be strong influential point.