CLW_R_Learning：Regression

Jason Wang

Tuesday, April 07, 2015

Regression equation: $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i$, $i = 1, 2, \cdots,$ n

- β_0, β_1, β_2,$\cdots$, β_k are parameters.
- x_{i1}, x_{i2}, x_{i3},$\cdots$, x_{ik} are constant.
- All of the ϵ_i are identically and independently distributed from $N(0, \sigma^2)$.

For matrix expression, $Y = X\beta$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

```
#Data
data(mtcars)
#Cars' names(32)
dimnames(mtcars)[[1]]
```

```
##  [1] "Mazda RX4"           "Mazda RX4 Wag"       "Datsun 710"
##  [4] "Hornet 4 Drive"      "Hornet Sportabout"   "Valiant"
##  [7] "Duster 360"          "Merc 240D"           "Merc 230"
## [10] "Merc 280"            "Merc 280C"           "Merc 450SE"
## [13] "Merc 450SL"          "Merc 450SLC"         "Cadillac Fleetwood"
## [16] "Lincoln Continental" "Chrysler Imperial"   "Fiat 128"
## [19] "Honda Civic"         "Toyota Corolla"      "Toyota Corona"
## [22] "Dodge Challenger"    "AMC Javelin"         "Camaro Z28"
## [25] "Pontiac Firebird"    "Fiat X1-9"           "Porsche 914-2"
## [28] "Lotus Europa"        "Ford Pantera L"      "Ferrari Dino"
## [31] "Maserati Bora"       "Volvo 142E"
```

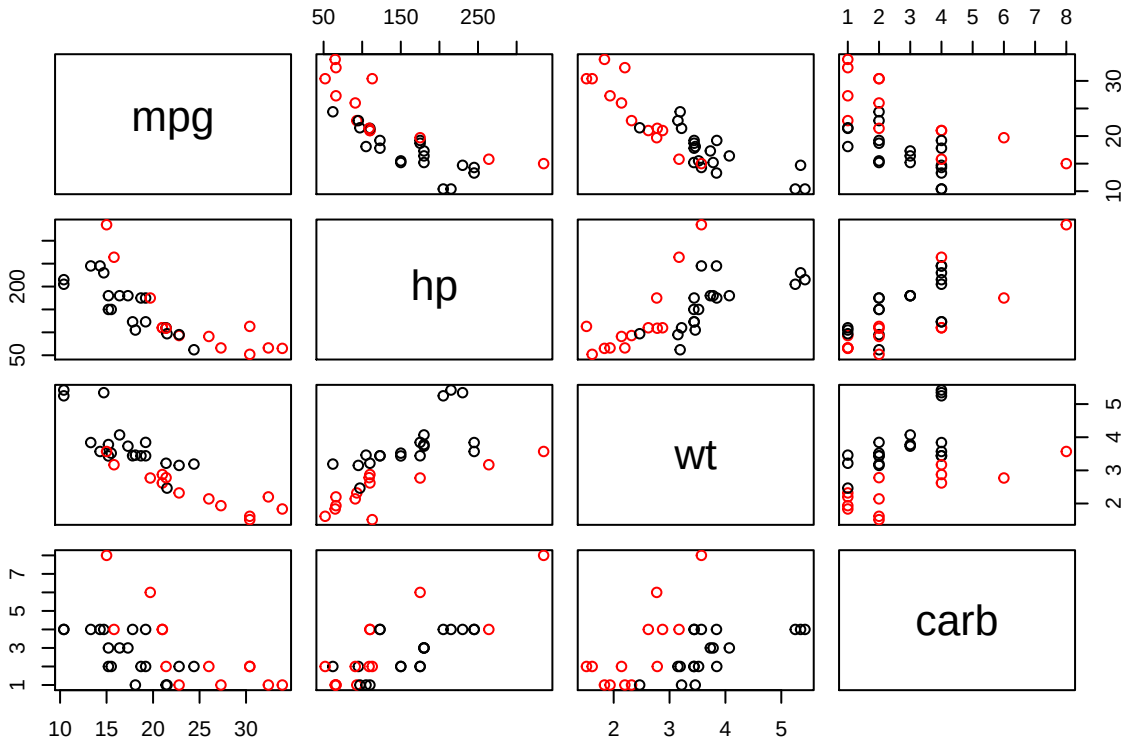
```
#Variable(11)
dimnames(mtcars)[[2]]
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

From my background perspective, I think that mpg(miles per gallon) will be related to hp(Gross horse power), wt(Weight) and am(Automatic v.s manual). After some researches, I found that vs(V-engine/Straight-engine) and carb(Carburetors) seem to be important factors.

Plot

```
#Overall view of data(Remove the variable I do not consider)
#Also change the color by am, black=auto and red=manual
pairs(mtcars[, -c(2, 3, 5, 7, 8, 9, 10)], col=mtcars$am + 1)
```



As you see above, there is a strong linear relationship between mpg and hp as well as wt. However, we also observe that it exist a relationship between hp and wt. It may happen **multicollinearity problem** when we fit model. Also, the lower number of carburetors a car have, the more efficient a car is.

```
#Correlation
round(cor(mtcars[, -c(2, 3, 5, 7, 8, 9, 10)]), 4)
```

```
##          mpg      hp      wt    carb
## mpg   1.0000 -0.7762 -0.8677 -0.5509
## hp   -0.7762  1.0000  0.6587  0.7498
## wt   -0.8677  0.6587  1.0000  0.4276
## carb -0.5509  0.7498  0.4276  1.0000
```

For view purpose, I round the correlation to 4 digit. hp and wt, as well as hp and carb, have high correlation. Again, it manifests that we should worried about the **multicollinearity problem**. Later, we will fit a model on mpg with variables, hp, wt, vs, am and carb. So far, we will not consider any interaction term.

Fit Model

```
#Convert vs and am to factor
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
#Using lm to fit model
m1 <- lm(mpg ~ hp + wt + vs + am + carb, data=mtcars)
summary(m1)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt + vs + am + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2383 -1.6647 -0.2427  1.3095  5.0380
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.48552    3.46540   8.797 2.85e-09 ***
## hp          -0.02340    0.01322  -1.770   0.0884 .
## wt          -2.40937    0.94170  -2.559   0.0167 *
## vs1          1.77406    1.33143   1.332   0.1943
## am1          2.96877    1.51062   1.965   0.0602 .
## carb        -0.42435    0.46612  -0.910   0.3710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.51 on 26 degrees of freedom
## Multiple R-squared:  0.8546, Adjusted R-squared:  0.8266
## F-statistic: 30.56 on 5 and 26 DF,  p-value: 4.23e-10
```

We can find that the adjusted R-squared $R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}}$ is 0.8266. Our model have quite a nice explain ability on our dependent variable mpg(miles per gallon.) However, interestingly, it seems that no all the variables are highly significant in the model. This will be a problem that we should deal with later. So, first, let's come to two test, overall test and marginal test:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \beta_i \neq 0 \text{ for some } i$$

For overall test, we will examine F-statistic = $\frac{MSR}{MSE} = 30.56 < F_{0.05,5,26}$. Hence, we reject null hypothesis H_0. Some of the variables have effects in the model.

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

For marginal test, we take hp as an example. We can compute the t-statistic = $\frac{b_{hp}}{sd(b_{hp})}$ = -1.77 < $t_{0.05,30}$. Hence, we reject H_0 which implied that when other variables exist in the model, hp(Horse power) can be removed. However, in the previous graph, we indeed observed there is a linear relationship between both variables. Here come the problem, *Multicollinearity.*(Note that the t-test here can be replaced by F-test because $(t_{n-p})^2 = F_{(1,n-p)}$. However, F-test only does one-tail test, not two-tail.)

Before we deal with the **multicollinearity**, we explore more on the problem of testing. Actually, we can test multiple in one times. We generalize all the test below:

$$H_0 : \text{Reduced model (less variables)}$$
$$H_1 : \text{Full modell (more variables)}$$

$$\text{F-statistics} = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} \qquad F_{(df_R - df_F), df_F})$$

```
#Model with intercept only
m2 <- update(m1, ~ . - .)
#The same test as testing hp
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp + wt + vs + am + carb
## Model 2: mpg ~ 1
##   Res.Df     RSS Df Sum of Sq     F   Pr(>F)
## 1     26  163.74
## 2     31 1126.05 -5    -962.3 30.56 4.23e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Model with hp
m3 <- update(m1, ~ . - hp)
#The same test as testing hp
anova(m1, m3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp + wt + vs + am + carb
## Model 2: mpg ~ wt + vs + am + carb
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 163.74
## 2     27 183.47 -1   -19.732 3.1331 0.08845 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the results are the same as previous test. Besides, we also can test multiple variables.

```
#Model with hp
m4 <- update(m1, ~  . - hp - wt)
#The same test as testing hp
anova(m1, m4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp + wt + vs + am + carb
## Model 2: mpg ~ vs + am + carb
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     26 163.74
## 2     28 245.65 -2   -81.908 6.5028 0.005129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Problem

1. Multicollinearity

We use VIF(Variance Inflation Factor) to detect that whether there is multicollinearity between variable. Below situation might indicate multicollinearity in the model:

- Remove or add a variable in model and this action causes dramatically change of other variables' coefficient.
- We consider it as important variable, while it is not statistically significant.
- The sign of variables are opposited to our expectation.
- The confidence interval of important variable are very big.

```r
library(car)
vif(m1)
```

```
##       hp       wt       vs       am     carb
## 4.043031 4.179091 2.216642 2.796825 2.790109
```

$$VIF_k = \frac{1}{1-R_k^2}, \text{ k = 1, 2, ..., p-1}$$

R_k^2 is the R-square of model that the rest of variable fit on the variable x_k.

In our example, all the variables' VIF is bigger than 1, especially hp and wt. Recall that indeed the correlation between variable is high. Probably, we should remove some variables. We start from the large correlation, hp and carb. I choose to remove carb because I think hp(horse power) will be a more important variable when it comes to evaluating mpg(miles per gallon).

If there is any variable's VIF greater than 10, it would be a strong warning that there is a multicollinearity. Also, if the average of VIF is greater that 1, it might be another warning too. However, we cannot judge which variable is linear-dependent on other variable directly by VIF. One way may be that repeat the step of adding and removing variable to make sure which variable is linear-dependent in the model.

Here, we construe the hp(horse power) as our main variable. Hence, in the following model, I will keep hp in my model

```r
m5 <- update(m1, ~ . - wt)
summary(m5)
```

```
##
## Call:
## lm(formula = mpg ~ hp + vs + am + carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3688 -1.8396  0.2073  1.8155  5.0291
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.25383    2.20127  10.564 4.32e-11 ***
## hp          -0.03239    0.01399  -2.315   0.0285 *
## vs1          2.54286    1.42408   1.786   0.0854 .
## am1          5.85636    1.10240   5.312 1.32e-05 ***
## carb        -0.67701    0.50014  -1.354   0.1871
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.755 on 27 degrees of freedom
## Multiple R-squared:  0.818,  Adjusted R-squared:  0.791
## F-statistic: 30.33 on 4 and 27 DF,  p-value: 1.237e-09
```

We check the vif again.

```
vif(m5)
```

```
##       hp       vs       am     carb
## 3.757361 2.103740 1.235668 2.664877
```

We removed the carb(carburetor) this time

```
m6 <- update(m1, ~ . - wt - carb)
summary(m6)
```

```
##
## Call:
## lm(formula = mpg ~ hp + vs + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0175 -1.7433  0.1203  1.4900  5.5150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.33420    2.23293  10.450 3.61e-11 ***
## hp          -0.04472    0.01078  -4.150 0.000281 ***
## vs1          2.65885    1.44247   1.843 0.075901 .
## am1          5.29854    1.03757   5.107 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.796 on 28 degrees of freedom
## Multiple R-squared:  0.8056, Adjusted R-squared:  0.7848
## F-statistic: 38.68 on 3 and 28 DF,  p-value: 4.31e-10
```

The model seems to be quite good while it seems that we can remove the vs.

```r
m7 <- update(m1, ~ . - wt - carb - vs)
summary(m7)
```

```
##
## Call:
## lm(formula = mpg ~ hp + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3843 -2.2642  0.1366  1.6968  5.8657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.584914   1.425094  18.655  < 2e-16 ***
## hp          -0.058888   0.007857  -7.495 2.92e-08 ***
## am1          5.277085   1.079541   4.888 3.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 29 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.767
## F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

After fixing the multicollinearity, we turn to outliers.

2. Outliers

Outliers from x perspective

We use leverage(h_{ii}) to detect x-outliers.

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$$

$h_{ii} > 0.5$ means that the effect of leverage is large.

$0.2 \leq h_{ii} \leq 0.5$ means that the effect of leverage is small.

```r
#No hat value is bigger than 0.5
sum(hatvalues(m7) > 0.5)
```

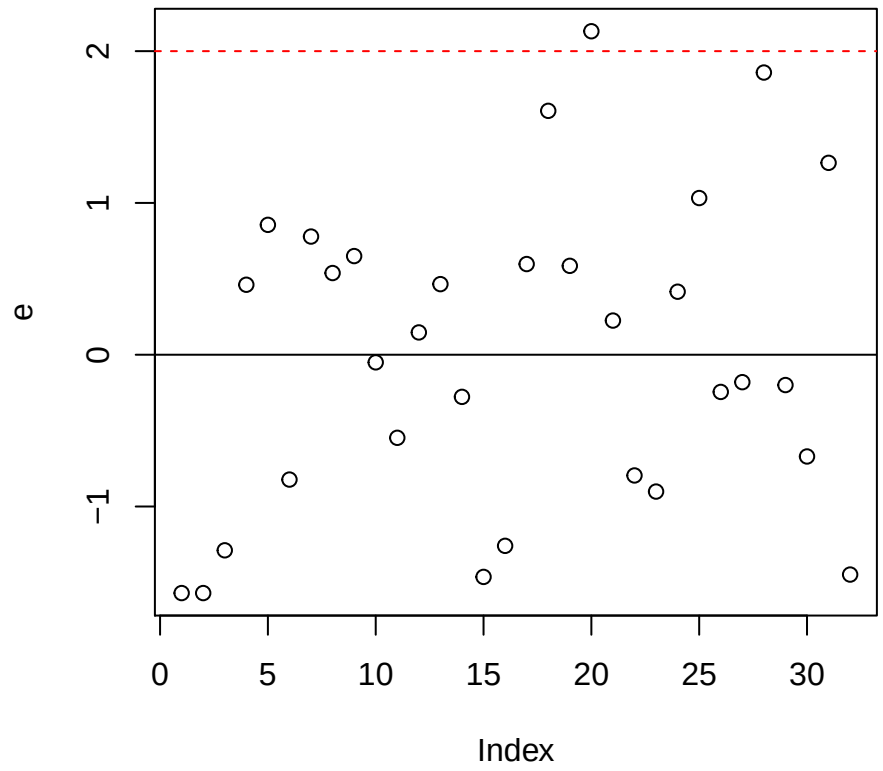
```
## [1] 0
```

Outliers from y perspective

We can use residuals to detect y outlier. Here we use standardized residuals:

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

```
#Stand
e <- rstandard(m7)
plot(e); abline(h=0); abline(h=c(-2, 2), col="red", lty=2)
```



The y outlier seems to be not so serious. Hence, we do not need to do transformation or other procedure to correct our model. After checking outliers, we turn to examine whether these outliers or other points in the data are extremely influential.

3. Influential points

Here, we introduce three influential measures: DFFITS, DFBETAS and Cook's Distance.

- DFFITS: The difference between fitted value(whole data) and fitted value(remove one data point). If DFFITS is big, it indicates that the data is influential to model. Here are two empirical standard:

$|(DFFITS)_i| > 1$ (for small and moderate data) or $|(DFFITS)_i| > 2\sqrt{\frac{p}{n}}$ (for large data).

- DFBETAS: The difference between coefficients with whole data and the coefficients(removing one data). Same as DFFITS, if DFBETAS is large, then the data is influential. Here are two empirical standard:

$|(DFBETAS)_i| > 1$ (for small and moderate data) or $|(DFBETAS)_i| > \frac{2}{\sqrt{n}}$ (for large data).

- Cook's Distance: It is a comprehensive indicator that measures whether certain data point is influential.

The following code can help us find out all the measures.

```
#DFFITS
DFFITS <- dffits(m7)
#DFBETAS
DFBETAS <- dfbetas(m7)
#Cook's Distance
Cook_D <- cooks.distance(m7)

round(data.frame(DFFITS=DFFITS, DFBETAS=DFBETAS, Cook.Distance=Cook_D), 4)
```

```
##                     DFFITS DFBETAS..Intercept. DFBETAS.hp DFBETAS.am1
## Mazda RX4           -0.4724             -0.0676     0.0765     -0.3298
## Mazda RX4 Wag       -0.4724             -0.0676     0.0765     -0.3298
## Datsun 710          -0.3982             -0.1101     0.1246     -0.2523
## Hornet 4 Drive       0.1258              0.1073    -0.0640     -0.0825
## Hornet Sportabout    0.2039              0.0633     0.0348     -0.1157
## Valiant             -0.2327             -0.2035     0.1269      0.1515
## Duster 360           0.2648             -0.0774     0.1870     -0.0704
## Merc 240D            0.1991              0.1940    -0.1506     -0.1171
## Merc 230             0.1944              0.1768    -0.1184     -0.1241
## Merc 280            -0.0128             -0.0100     0.0051      0.0085
## Merc 280C           -0.1399             -0.1097     0.0562      0.0929
## Merc 450SE           0.0350              0.0090     0.0079     -0.0191
## Merc 450SL           0.1112              0.0285     0.0252     -0.0608
## Merc 450SLC         -0.0662             -0.0170    -0.0150      0.0362
## Cadillac Fleetwood  -0.4013             -0.0011    -0.1870      0.1740
## Lincoln Continental -0.3610              0.0306    -0.1955      0.1401
## Chrysler Imperial    0.1836             -0.0365     0.1165     -0.0594
## Fiat 128             0.5633              0.2537    -0.2871      0.2924
## Honda Civic          0.2115              0.1101    -0.1246      0.0975
## Toyota Corolla       0.7802              0.3556    -0.4025      0.4017
## Toyota Corona        0.0661              0.0597    -0.0395     -0.0423
## Dodge Challenger    -0.1877             -0.1072     0.0225      0.1207
```

```
## AMC Javelin         -0.2134             -0.1218     0.0256      0.1372
## Camaro Z28           0.1401             -0.0410     0.0990     -0.0373
## Pontiac Firebird     0.2474              0.0768     0.0423     -0.1404
## Fiat X1-9           -0.0822             -0.0370     0.0419     -0.0426
## Porsche 914-2       -0.0547             -0.0159     0.0180     -0.0342
## Lotus Europa         0.5674              0.0670    -0.0758      0.4018
## Ford Pantera L      -0.1026              0.0726    -0.0821     -0.0659
## Ferrari Dino        -0.2136              0.0801    -0.0907     -0.1666
## Maserati Bora        1.0281             -0.8146     0.9220      0.5642
## Volvo 142E          -0.4335             -0.0656     0.0742     -0.3012
##                    Cook.Distance
## Mazda RX4                 0.0705
## Mazda RX4 Wag             0.0705
## Datsun 710                0.0516
## Hornet 4 Drive            0.0054
## Hornet Sportabout         0.0140
## Valiant                   0.0183
## Duster 360                0.0237
## Merc 240D                 0.0135
## Merc 230                  0.0129
## Merc 280                  0.0001
## Merc 280C                 0.0067
## Merc 450SE                0.0004
## Merc 450SL                0.0042
## Merc 450SLC               0.0015
## Cadillac Fleetwood        0.0515
## Lincoln Continental       0.0425
## Chrysler Imperial         0.0115
## Fiat 128                  0.0998
## Honda Civic               0.0153
## Toyota Corolla            0.1773
## Toyota Corona             0.0015
## Dodge Challenger          0.0119
## AMC Javelin               0.0153
## Camaro Z28                0.0067
## Pontiac Firebird          0.0204
## Fiat X1-9                 0.0023
## Porsche 914-2             0.0010
## Lotus Europa              0.0979
## Ford Pantera L            0.0036
## Ferrari Dino              0.0155
## Maserati Bora             0.3448
## Volvo 142E                0.0602
```

There are some influential points in mtcars data. However, we do not have the sufficient

knowledge to determine that whether those data points are irregular or those points are important that we should keep each of them in the model. We will not remove any data.(We also think that it should be.)

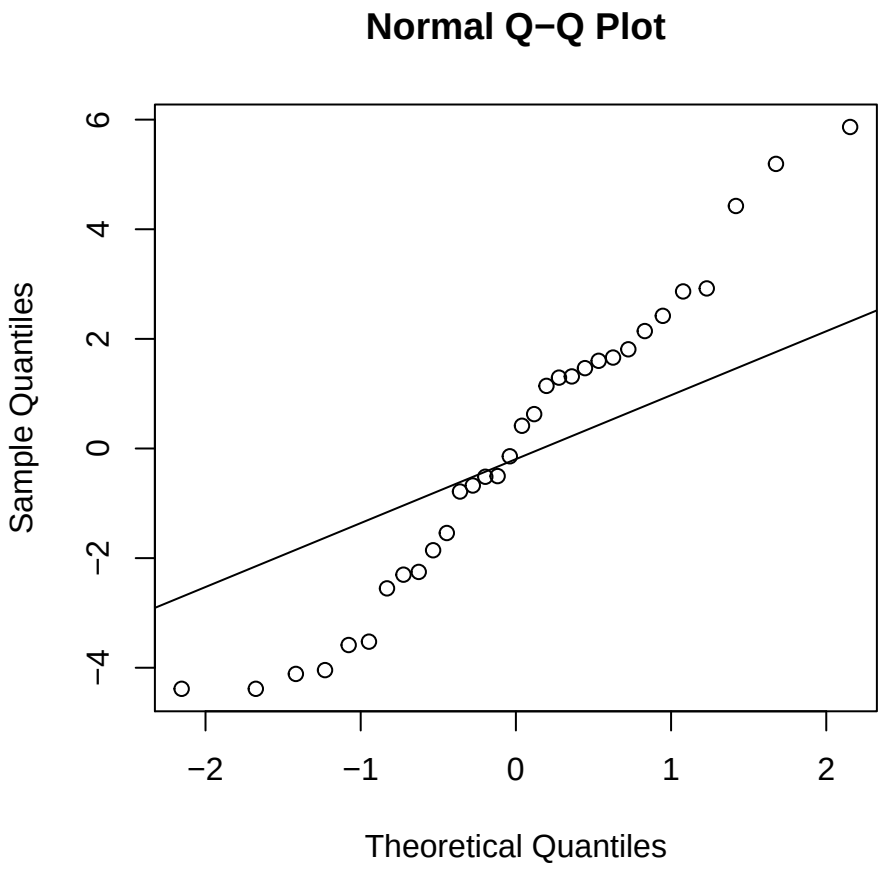
In R, there is a function that provide all the influential measures. We can utilize it to determine the influential points.(We do not show the result here)

```
#Summary
influence.measures(m1)
```

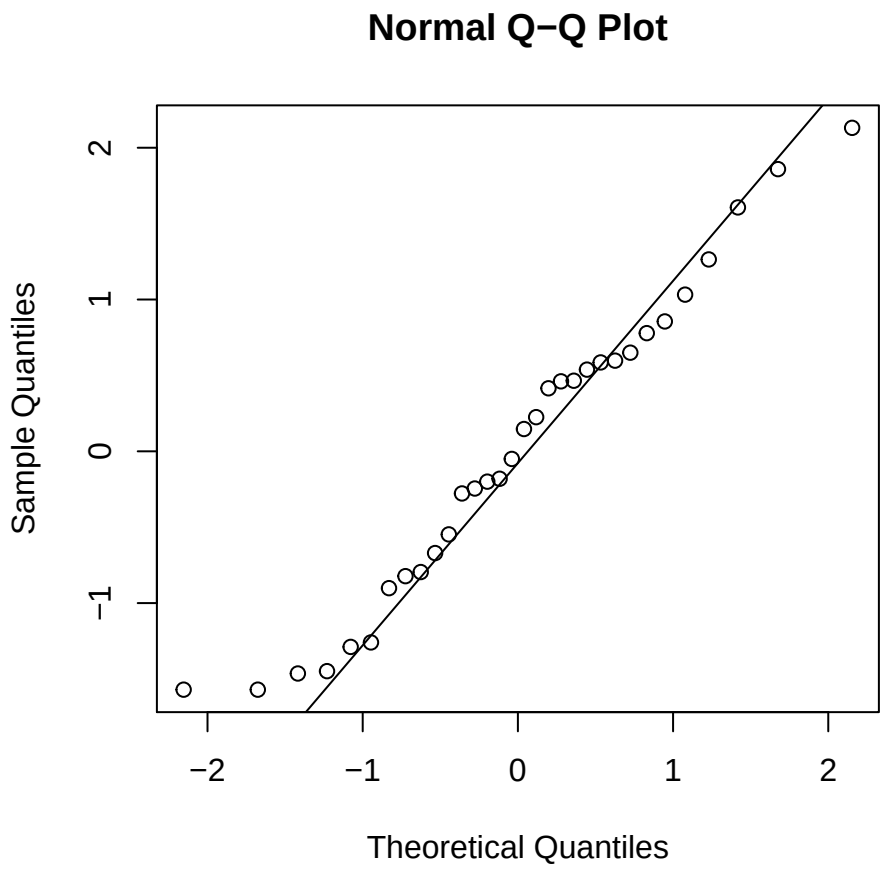
Diagnostic Checking

1. Normality

```
qqnorm(resid(m7))
qqline(rnorm(100))
```



```r
qqnorm(rstandard(m7))
qqline(rnorm(100))
```



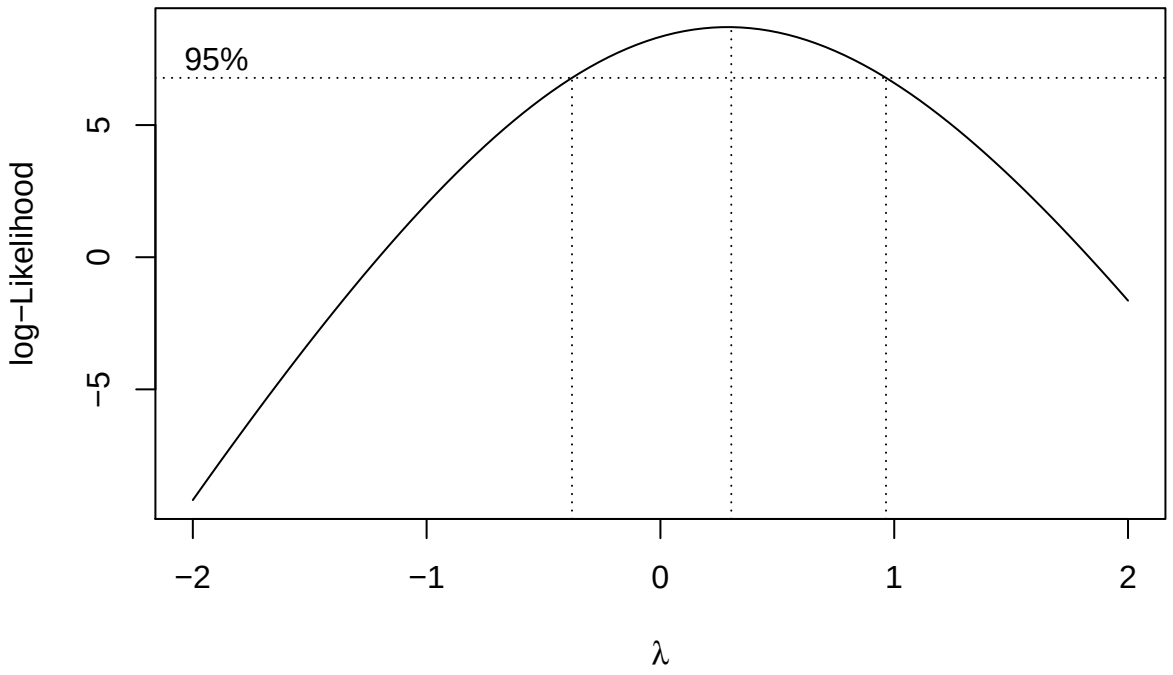
H_0 : residuals are from Normal Distribution
H_1 : residual are not from Normal Distribution

```r
ks.test(resid(m7), "pnorm")
```

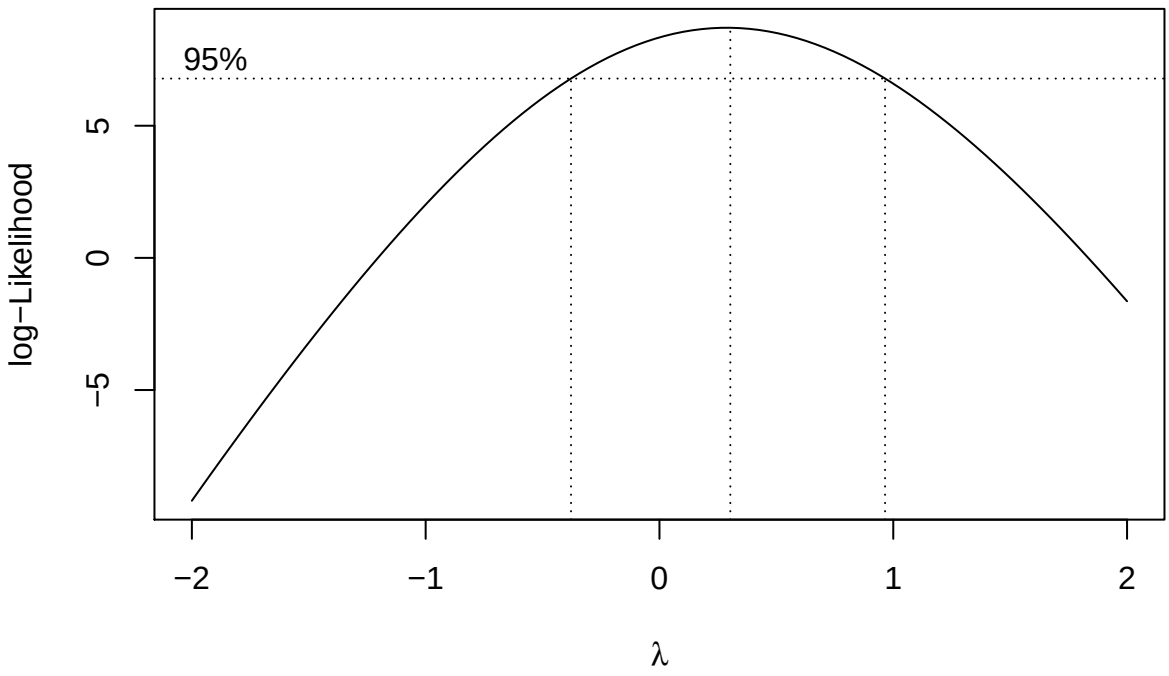
```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  resid(m7)
## D = 0.3109, p-value = 0.003026
## alternative hypothesis: two-sided
```

The p-value is 0.0030259 is less that $\alpha(0.05)$. The residual is not from the normal distribution. Hence, we can use the Box-Cox transformation to tranforms the data.

```
library(MASS)
boxcox(m7)
```



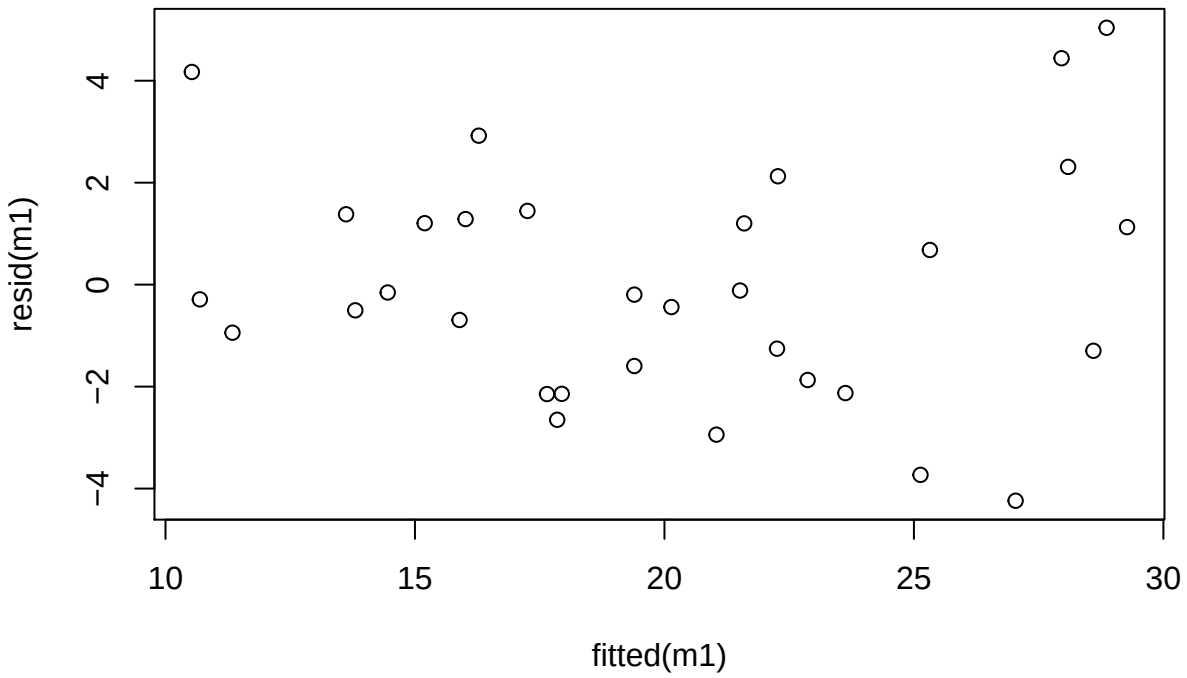

```
(lambda <- boxcox(m7)$x[which.max(boxcox(m7)$y)])
```




```
## [1] 0.3030303
```

2. Constant Variance

```
plot(fitted(m1), resid(m1))
```



From the graph above, we judge that the variance is contant.

3. Independent

$$H_0 : \text{residuals are independent}$$
$$H_1 : \text{residuals are not independent}$$

```
library(car)
durbinWatsonTest(m7)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        0.198867      1.457342   0.066
##  Alternative hypothesis: rho != 0
```

The p-value is greater that $\alpha(0.05)$. Hence, we do not reject the independence assumption.

Model Identification

We can also do stepwise regression to selection our model while it really need more understanding about the data and its field. By that, we can judge our model more correctly and accurately. Heve, we just show the code in R that can do the step wise regression.

```r
library(leaps)
```

```r
X <- mtcars[, c(4, 6, 8, 9 , 11)]
y <- mtcars$mpg
```

```r
out <- summary(regsubsets(X, y, nbest=2, nvmax=ncol(X)))
report <- cbind(out$which, out$adjr2, out$rss, out$bic, out$cp)
report
```

Interpretation

$$\text{mpg} = 26.5849 + \text{-}0.0589\ \text{hp} + 5.2771\ \text{am}$$

This is our final model. For fixed am(automatic or manual), per unit hp rise, the average mpg will be go down -0.0589. It also tells us that the automatic car will be more efficient than the manaul car.

To conclude, when fitting a regression model, there is no a standard procedure to follow such as multicollinearity → outliers → influential points in this report. Instead, many step is interchangeable as long as the analysis is reasonable. Also, we do not consider interaction term which will be important when we fit a regression model, which is the future task we will discuss.

References

*Multiple Regression http://www.statmethods.net/stats/regression.html

- Regression Diagnostics http://www.statmethods.net/stats/rdiagnostics.html