HW4_IS457_127

ClassID:127

Part 1. Linear Regression Concepts (6pts)

In this homework, "Regression" refers to the simple linear regression equation: y = b0 + b1*x

Q1. (2pts)

What is the interpretation of the coefficient b1? (What meaning does it represent?)

Your Answer:

$b1$ represents the estimated value of y will be changed by 1 unit increases in explanatory variable x1.

Q2. (2pts)

Outliers are problems for many statistical methods, but are particularly problematic for linear regression. Why is that? It may help to define what outlier means in this case.(Hint: Think of how residuals are calculated)

Your Answer:

Linear regression uses least square method to select the best regression line, which means it selects the line which has the least residual sum of squares. The residual is the square of the difference between observed response(y_hat) and response(y).Typically, points further than, say, three or four standard deviations from the mean are considered as outliers. The influence of these farthest points are amplified when choosing the best fitted line using least square methods, since the distance to the best-fit line is squared when calculating residuals. And it is very likely that the choice of the best fitted line be changed by very few outlier.

Q3. (2pts)

How could you deal with outliers in order to improve the accuracy of your model?

Your Answer:

We might do data visulization(boxplot, scatter plot) to find the potential outlier before doing linear regression. Then we can look at the residual plot after fitting linear regression model using full dataset in order to check whether there are influential points exists. If there are some influential points (possible outliers), we should check that data points and looks at whether these points involve special properties about the dataset, since removing data might be 'dangerous' and we don't want to lose information. If no such distinguishing features can be found, we can remove these outliers and then do linear regression.

Part 2. Sampling, Point Estimation, and creating functions

The following problems will use the Rabbit dataset and explore the Blood Pressure change(BPchange) for Rabbit in control group of "Treatment".
Load the data by running the following code:

```
library(MASS)
data(Rabbit)
```

Q4.

Subset the data frame to include ONLY rabbits (observations) in control group of "Treatment". (2pts)
Name it 'rabbitCon', and show the first 10 observations of your output.(2pts)

```
rabbitCon<-Rabbit[Rabbit$Treatment == 'Control',]
head(rabbitCon,n=10)
```

```
##    BPchange   Dose Run Treatment Animal
## 1      0.50   6.25  C1   Control     R1
## 2      4.50  12.50  C1   Control     R1
## 3     10.00  25.00  C1   Control     R1
## 4     26.00  50.00  C1   Control     R1
## 5     37.00 100.00  C1   Control     R1
## 6     32.00 200.00  C1   Control     R1
## 7      1.00   6.25  C2   Control     R2
## 8      1.25  12.50  C2   Control     R2
## 9      4.00  25.00  C2   Control     R2
## 10    12.00  50.00  C2   Control     R2
```

Use the sample function to generate a vector of 1s and 2s with the same length as rabbitCon, call it 'group'.(2pts)
Use this vector to split the 'BPchange' variable into two vectors, BP_V1 and BP_V2. (4pts) Print out the vectors group, BP_V1, BP_V2 and the lengths of BP_V1 and BP_V2.
IMPORTANT: Make sure to run the seed function before running the sample function to ensure the result is reproducible.

```
set.seed(457)  # DO NOT change
group<-sample(2,30,replace = T)
group
```

```
##  [1] 1 1 1 2 2 1 1 2 1 2 2 2 2 1 1 1 2 2 1 1 1 1 2 1 1 1 2 1 1 2
```

```
BP_V1<-rabbitCon$BPchange[group ==1]
BP_V1
```

```
##  [1]  0.50  4.50 10.00 32.00  1.00  4.00  3.00  3.00 14.00  1.25  1.50
## [12]  6.00 19.00 33.00  1.50  1.50 16.00 20.00
```

```
length(BP_V1)
```

```
## [1] 18
```

```
BP_V2<-rabbitCon$BPchange[group ==2]
BP_V2
```

```
##  [1] 26.00 37.00  1.25 12.00 27.00 29.00  0.75 22.00 24.00 33.00  5.00
## [12] 18.00
```

```
length(BP_V2)
```

```
## [1] 12
```

Q5(1)

Calculate the mean and the standard deviation for each of the two vectors, BP_V1 and BP_V2. (4pts)
Create a 95% confidence interval for your sample means using Z score.(4pts) (you can use the following formula for the Confidence Interval: mean +/- 1.96 * standard deviation).
Compare the confidence intervals, do they seem to agree or disagree, explain (their ranges? differences?).(2pts)
Note: the z score for 95% confidence interval is 1.96.

```
mean(BP_V1)
```

```
## [1] 9.541667
```

```
sd(BP_V1)
```

```
## [1] 10.53574
```

```
mean(BP_V2)
```

```
## [1] 19.58333
```

```
sd(BP_V2)
```

```
## [1] 12.27355
```

```
CI_V1<-c(mean(BP_V1)-1.96*sd(BP_V1),mean(BP_V1)+1.96*sd(BP_V1))
CI_V1
```

```
## [1] -11.10839  30.19172
```

```
CI_V2<-c(mean(BP_V2)-1.96*sd(BP_V2),mean(BP_V2)+1.96*sd(BP_V2))
CI_V2
```

```
## [1] -4.472834 43.639501
```

n1 <- length(BP_V1)
n2 <- length(BP_V2)
CI1 <- c(mu1-(z*std1)/sqrt(n1),mu1+(z*std1)/sqrt(n1))
CI2 <- c(mu2-(z*std2)/sqrt(n2),mu2+(z*std2)/sqrt(n2))

The 95% confidence interval of BP_V1 is [-11.10839,30.19172] and the mean of BP_V1 is 9.541667; The 95% confidence interval of BP_V2 is [-4.472834,43.639501] and the mean of BP_V1 is 19.58333. Both of the mean of BP_V1 and BP_V2 lie in the middle of their CI.
The 95% confidence interval is the interval has a 0.95 probability of containing the mean of that population. BP_V1 and BP_V2 are randomly draw from the same dataset, so their 95% confidence intervals should be compatible. According to the result, the mean of BP_V1(9.541667) lies in the 95% confidence interval of BP_V2 [-4.472834,43.639501]. And the mean of BP_V2(19.58333) lies in the 95% confidence interval of BP_V1 [-11.10839 30.19172]. Therefore, they agree with each other.

Q5(2) From what you practice in 5 (1), let's generalize the calculation process. (5pts)

Write a function to calculate the 95% confidence intervals of any input vector (numerical) x, according to the formula given in the previous question.

```
find_CI<-function(x){
c(mean(x)-1.96*sd(x),mean(x)+1.96*sd(x))
}
find_CI(BP_V1)#test! first value is the lower bound, second value is the upper bound.
```

```
## [1] -11.10839  30.19172
```

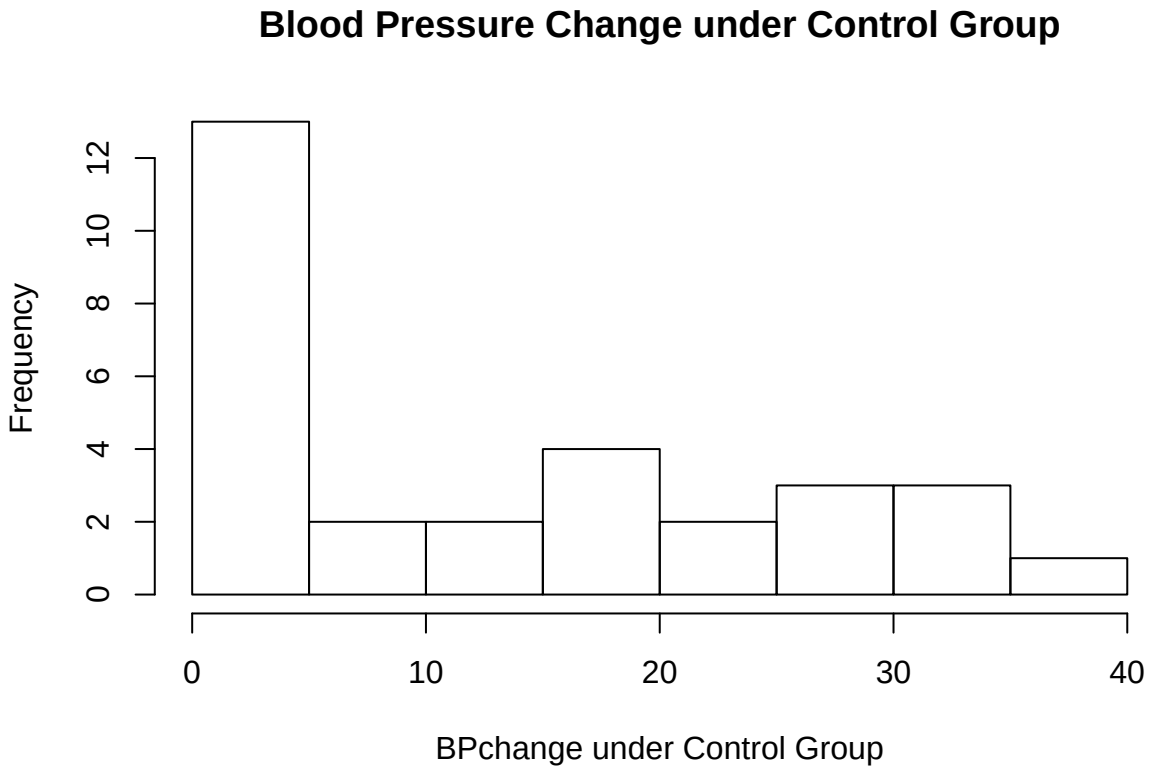
Q6.

Using the hist() fucntion, plot a histogram of BPchange of rabbits under control group as well as for the MDL group (separately). (2pts)
Do the histograms resemble a normal distribution? why or why not? (2pts)
Comment on the shape of the distributions you see in the histgrams. What does the shape indicate in the context of this dataset?(4pts)

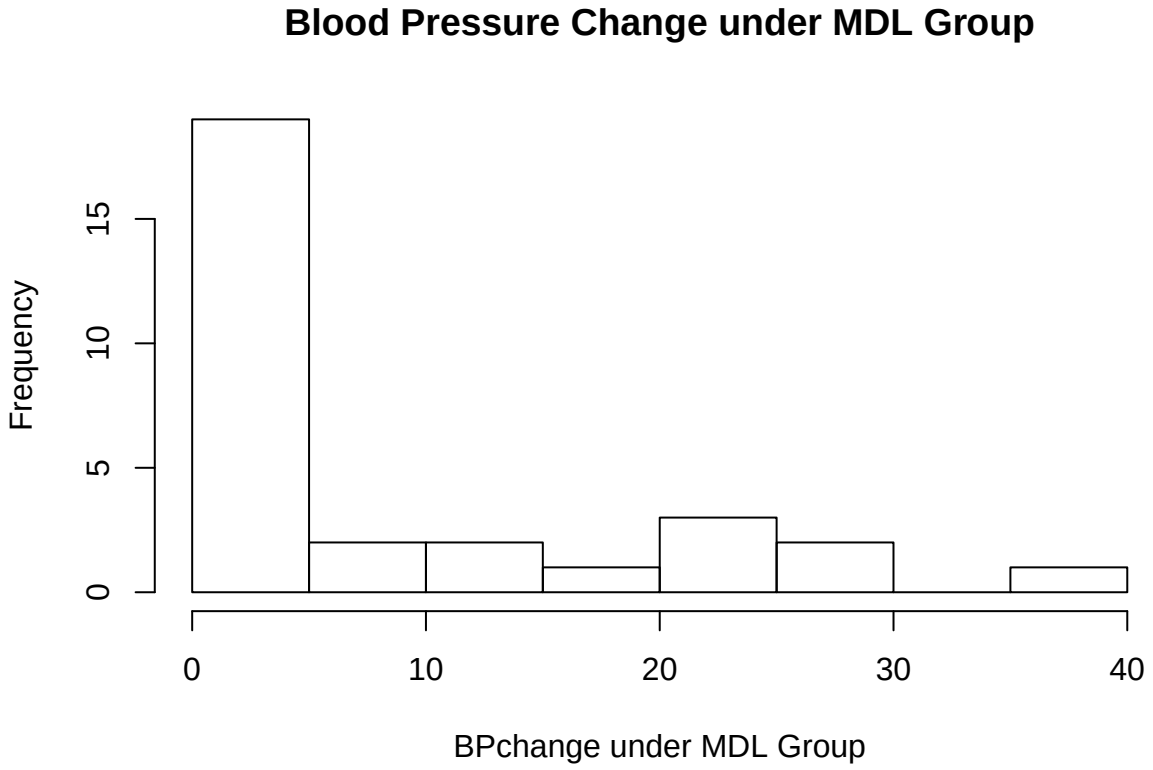
```
rabbitMDL<-Rabbit[Rabbit$Treatment == 'MDL',]
head(rabbitMDL)
```

```
##    BPchange   Dose Run Treatment Animal
## 31     1.25   6.25  M1       MDL     R1
## 32     0.75  12.50  M1       MDL     R1
## 33     4.00  25.00  M1       MDL     R1
## 34     9.00  50.00  M1       MDL     R1
## 35    25.00 100.00  M1       MDL     R1
## 36    37.00 200.00  M1       MDL     R1
```

```
hist(rabbitCon$BPchange, xlab = "BPchange under Control Group",
     main="Blood Pressure Change under Control Group")
```




```
hist(rabbitMDL$BPchange,xlab = "BPchange under MDL Group",
     main="Blood Pressure Change under MDL Group")
```



According to the graph above, neither of them seem to resemble a normal distribution, because their shapes are not symmetric and bell-shaped. And according to the shape of histgrams, both of them are positive skew which means both of them have a long positive tail. Besides, for rabbit under control group and MDL group,the number of rabbits' BPchange which is less than 5 is the largest. And for the rest of rabbits in each group, BPchange tend to uniformly distributed in range[5,40].

Part 3 Linear Regression

This problem will use the same dataset as Part 2. We will focus on two variables:
BP change: change in blood pressure relative to the start of the experiment.
Dose: dose of Phenylbiguanide in micrograms.
To start with, let us define a null hypthosesis. If we want to test the effect of dosage on BPchange, the null hypthosesis is:

H0: Dosage has no effect on BPchange. H0: B1 = 0
HA: B1 != 0

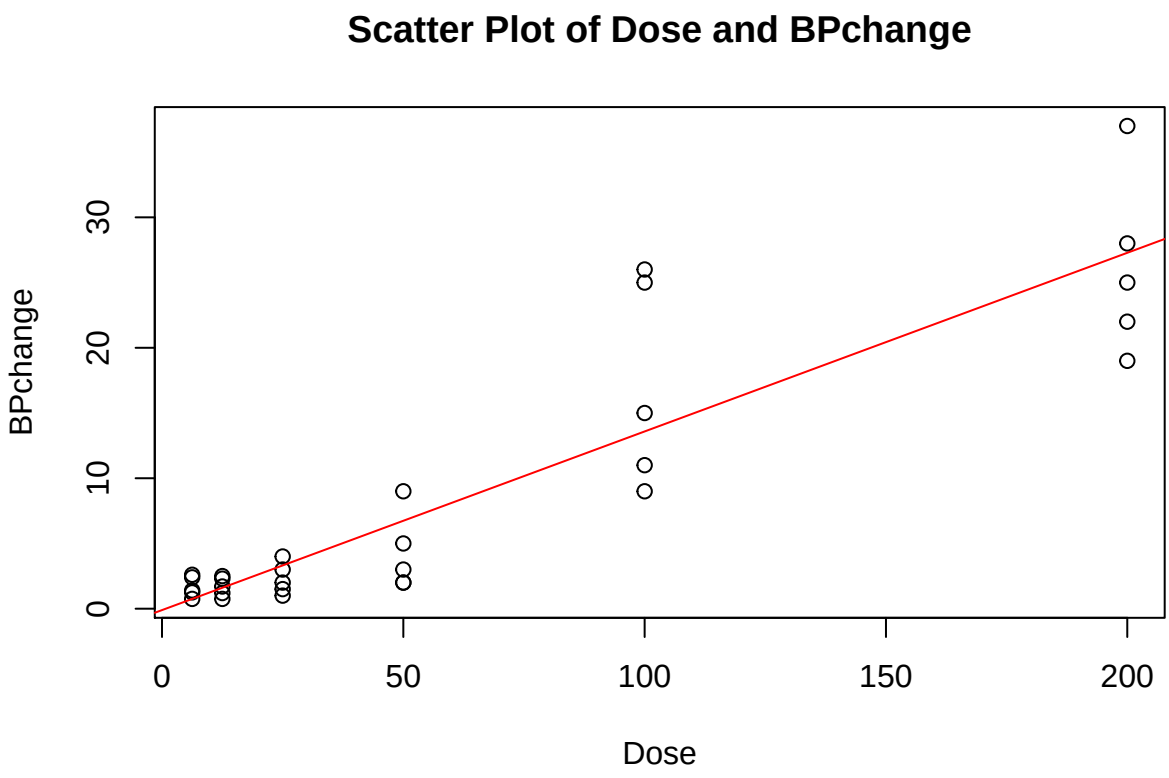
Q7.

Fit a linear regresssion using Dose to predict BPchange, using lm() for rabbits under MDL treatment. (2pts)
Name it 'model_BP'. What function would you use to get the summary statistics from lm models? Go ahead and use it.(2pts)
Examine the model diagnostics using plot(). Comment on the plots, what do the fitted values, noise, outliers look like?(8pts)
Would you consider this a good model or not? Please explain. (2pts)

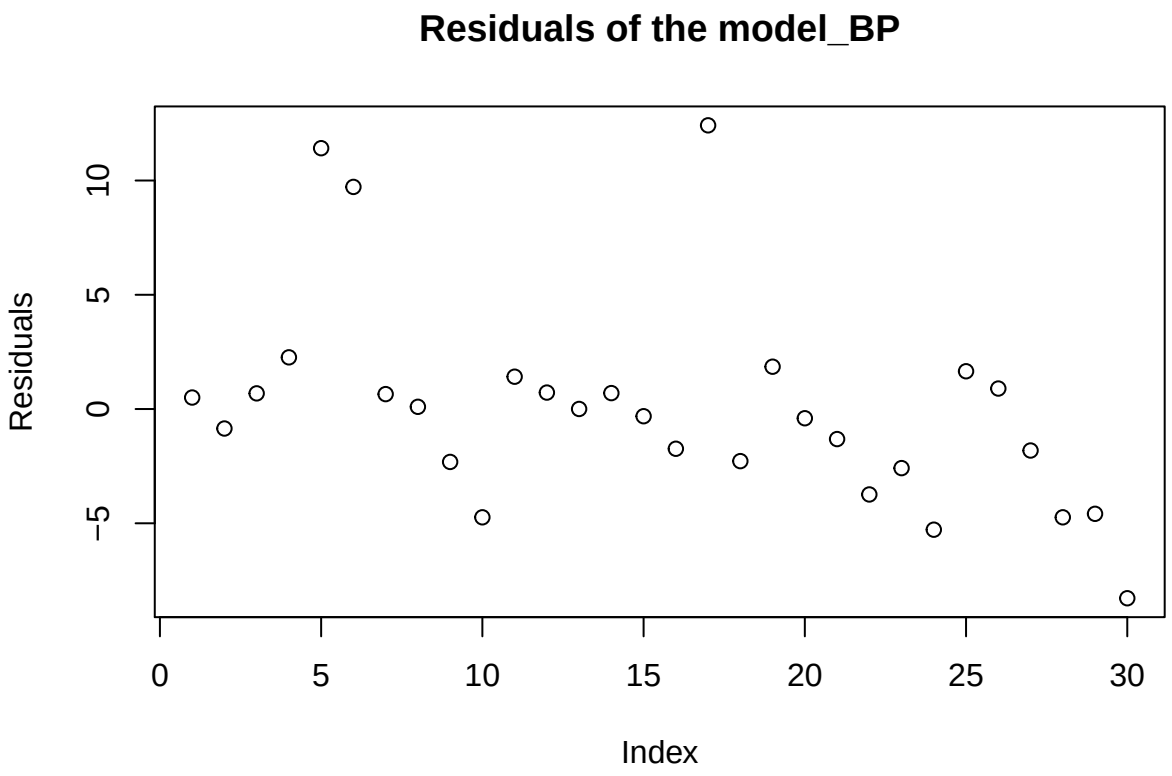
```
model_BP<-lm(BPchange~Dose, data = rabbitMDL)
summary(model_BP)
```

```
##
## Call:
## lm(formula = BPchange ~ Dose, data = rabbitMDL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2802 -2.3063 -0.1561  0.8526 12.4142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.10861    1.17505  -0.092    0.927
## Dose         0.13694    0.01246  10.986 1.16e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.62 on 28 degrees of freedom
## Multiple R-squared:  0.8117, Adjusted R-squared:  0.805
## F-statistic: 120.7 on 1 and 28 DF,  p-value: 1.159e-11
```

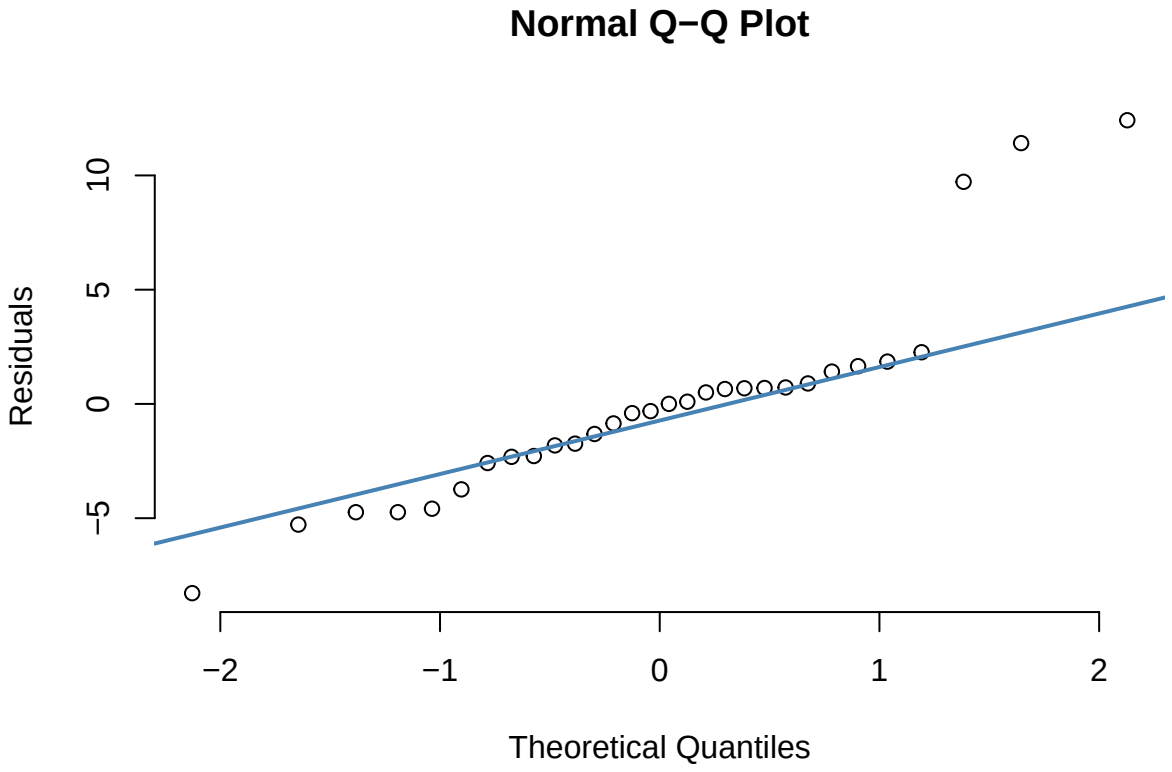
```
plot(x=rabbitMDL$Dose, y = rabbitMDL$BPchange,xlab='Dose',ylab = 'BPchange',
     main = 'Scatter Plot of Dose and BPchange')
abline(model_BP$coefficients[1],model_BP$coefficients[2],col = 'red')
```




```
plot(model_BP$residuals, ylab = 'Residuals', main = 'Residuals of the model_BP' )
```




```
#Draw the quantile-quantile plot to see whether the residuals follow normal distribution
qqnorm(model_BP$residuals, ylab= 'Residuals',pch = 1, frame = FALSE)
qqline(model_BP$residuals, col = "steelblue", lwd = 2)
```



I will use `summary(model_BP)` to get the summary statistics from lm models.
As we can see from the graph above, in general, the fitted values fit the majority of actually value, but there

are four data points (two of them at Dose = 100;two of them at Dose = 200) away from the fitted value. So they might be the outliers. As for the noise(residuals), they tend to be randomly distributed around the regression line.
As we can see from the residuals plot, the residuals seem randomly distributed around 0, but there are four "suspicious" points which are more far away from the mean than any other points. So they might be the outliers. Then, according to the quantile-quantile plot of residuals, we can see that except four points, the rest residual points seem to be normally distributed. Thus, there seems exist three potential outliers in the original dataset.
In general, I would like to say it is a OK model. The p-value is smaller than .05, so the model is significant and the R-square is 0.8117 which means this model can explain 81.17% variation of the BPchange, so it is good. However, there exist some "suspicious" points which might be outliers, and we need to do further analysis to decide whether keep them in the model.

Q8.

With the summary statistics from above, calculate the 95% confidence interval for Dose using t score (2pts)
Note: use this code to find the t score: tvalue <- qt(1-0.05/2,nrow(rabbitMDL)-2)

```
tvalue <- qt(1-0.05/2,nrow(rabbitMDL)-2)
CI_Dose<-c(0.13694-tvalue* 0.01246,0.13694+tvalue* 0.01246)#Estimate Std = 0.13694; Std.Error = 0.01246
CI_Dose#first value is the lower bound of 95% CI, second value is the upper bound.
```

```
## [1] 0.1114168 0.1624632
```

Q9.

Based on the result from Q7& Q8 (p-value and CI), would you reject the null hypothesis or not? Explain. (2pts)

According to the result, P-value is 1.159e-11<0.05, we should reject null hypothesis. And the 95% confidence intervals is [0.1114168, 0.1624632] which is not include 0. Therefore, we can reject the null hypothesis.