HW4_IS457_39

39

10/8/2018

Do not remove any of the comments. These are marked by

HW 4 - Due Monday, Oct 8, 2018 in moodle and hardcopy in class

(1). Please upload R code and report to Moodle

with filename: HW4_IS457_YourClassID.

(2). Turn in hard copy of your report in class

without your name but only your class ID

Important: Make sure there is no identifying information on your printout, including name, username etc.

Only include your class ID on there.

ClassID: 39

In this assignment you will practice linear regression in R

Part 1. Linear Regression Concepts (6pts)

These questions do not require coding but you will need to explain the details.

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 is the coeeficient of variable x which impact the value of y. For example, if x increases
# by 1, then y increases correspondingly by b1.
```

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:

```
# Residuals is calculated by the distance between the point and the regression model line. In this case
# ourliers would affect the value of coefficient b0 and b1 in linear regresson fitting, which would lea
# model bias.
```

Q3. (2pts)

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

Your Answer:

```
# Two ways could be considered for dealing with outliers.
# 1st: Directly deleting the outliers
# 2nd: Modifying the outliers while fitting regression model, such as assigning reasonable
# values to detected outliers.
```

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)
summary(Rabbit)
```

```
##     BPchange          Dose             Run       Treatment  Animal
##  Min.   : 0.50   Min.   :  6.25   C1     : 6   Control:30   R1:12
##  1st Qu.: 1.65   1st Qu.: 12.50   C2     : 6   MDL    :30   R2:12
##  Median : 4.75   Median : 37.50   C3     : 6                R3:12
##  Mean   :11.22   Mean   : 65.62   C4     : 6                R4:12
##  3rd Qu.:20.50   3rd Qu.:100.00   C5     : 6                R5:12
```

```
##  Max.   :37.00   Max.   :200.00   M1     : 6
##                                   (Other):24
```

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)

Your code:

```
rabbitCon = Rabbit[Rabbit$Treatment == 'Control',]
head(rabbitCon, 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

#Your Code:
group = sample(c('1s', '2s'), size = nrow(rabbitCon), replace = T)
rabbitCon = cbind(rabbitCon, group)
print(rabbitCon)
```

```
##    BPchange   Dose Run Treatment Animal group
## 1      0.50   6.25  C1   Control     R1    1s
## 2      4.50  12.50  C1   Control     R1    1s
## 3     10.00  25.00  C1   Control     R1    1s
## 4     26.00  50.00  C1   Control     R1    2s
## 5     37.00 100.00  C1   Control     R1    2s
## 6     32.00 200.00  C1   Control     R1    1s
## 7      1.00   6.25  C2   Control     R2    1s
## 8      1.25  12.50  C2   Control     R2    2s
## 9      4.00  25.00  C2   Control     R2    1s
## 10    12.00  50.00  C2   Control     R2    2s
## 11    27.00 100.00  C2   Control     R2    2s
## 12    29.00 200.00  C2   Control     R2    2s
## 13     0.75   6.25  C3   Control     R3    2s
## 14     3.00  12.50  C3   Control     R3    1s
## 15     3.00  25.00  C3   Control     R3    1s
## 16    14.00  50.00  C3   Control     R3    1s
## 17    22.00 100.00  C3   Control     R3    2s
## 18    24.00 200.00  C3   Control     R3    2s
## 19     1.25   6.25  C4   Control     R4    1s
## 20     1.50  12.50  C4   Control     R4    1s
## 21     6.00  25.00  C4   Control     R4    1s
## 22    19.00  50.00  C4   Control     R4    1s
## 23    33.00 100.00  C4   Control     R4    2s
## 24    33.00 200.00  C4   Control     R4    1s
## 25     1.50   6.25  C5   Control     R5    1s
## 26     1.50  12.50  C5   Control     R5    1s
```

```
## 27     5.00  25.00  C5   Control     R5    2s
## 28    16.00  50.00  C5   Control     R5    1s
## 29    20.00 100.00  C5   Control     R5    1s
## 30    18.00 200.00  C5   Control     R5    2s
```

```
BP_V1 = rabbitCon$BPchange[rabbitCon$group == '1s']
BP_V2 = rabbitCon$BPchange[rabbitCon$group == '2s']
length(BP_V1)
```

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

```
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 +/- 2 * standard deviation).

Compare the confidence intervals, do they seem to agree or disagree, explain (their ranges? differences?). (2pts)

Your code:

```
BP_V1_mean = mean(BP_V1)
BP_V2_mean = mean(BP_V2)
BP_V1_sd = sd(BP_V1)
BP_V2_sd = sd(BP_V2)
confint_V1 = c(BP_V1_mean - 2*BP_V1_sd, BP_V1_mean + 2*BP_V1_sd)
confint_V2 = c(BP_V2_mean - 2*BP_V2_sd, BP_V2_mean + 2*BP_V2_sd)
print(confint_V1)
```

```
## [1] -11.52982  30.61315
```

```
print(confint_V2)
```

```
## [1] -4.963777 44.130443
```

```
# According to the result, the two confidence intervals seem to agree because the size of
# their ranges are similar but the BP_V2 confidence interval is a littel wider and the
# distribution of the BP_V2 is located at the right side.
```

Note: the z score for 95% confidence interval is 1.96.

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.

```
confint95 = function(x){
  conf_intv = c(mean(x) - 2*sd(x), mean(x) + 2*sd(x))
  return(conf_intv)
}
```

Q6.

Using the hist() fucntion, plot a histogram of BPchange of rabbits under control group as well as for

the MDL group (separately). (2pts)

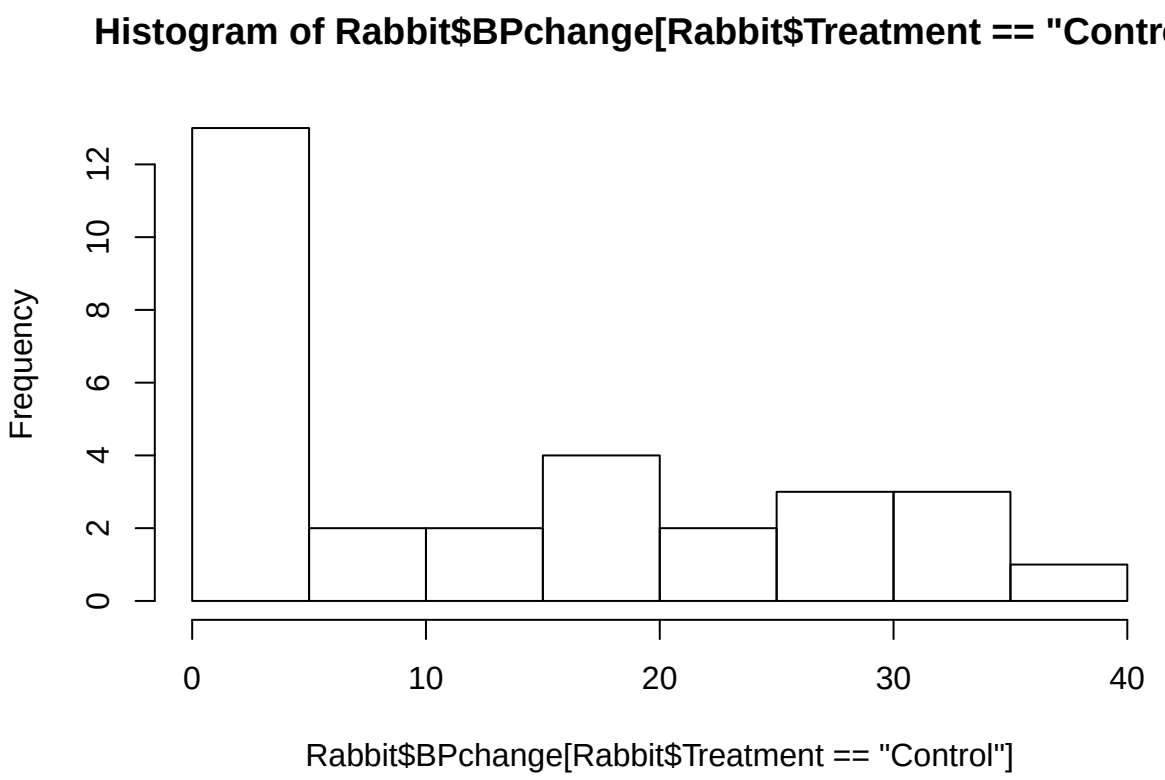
Do the histograms resemble a normal distribution? why or why not? (2pts)

Comment on the shape of the distributions you see in the histgrams. What does the shape indicate in the

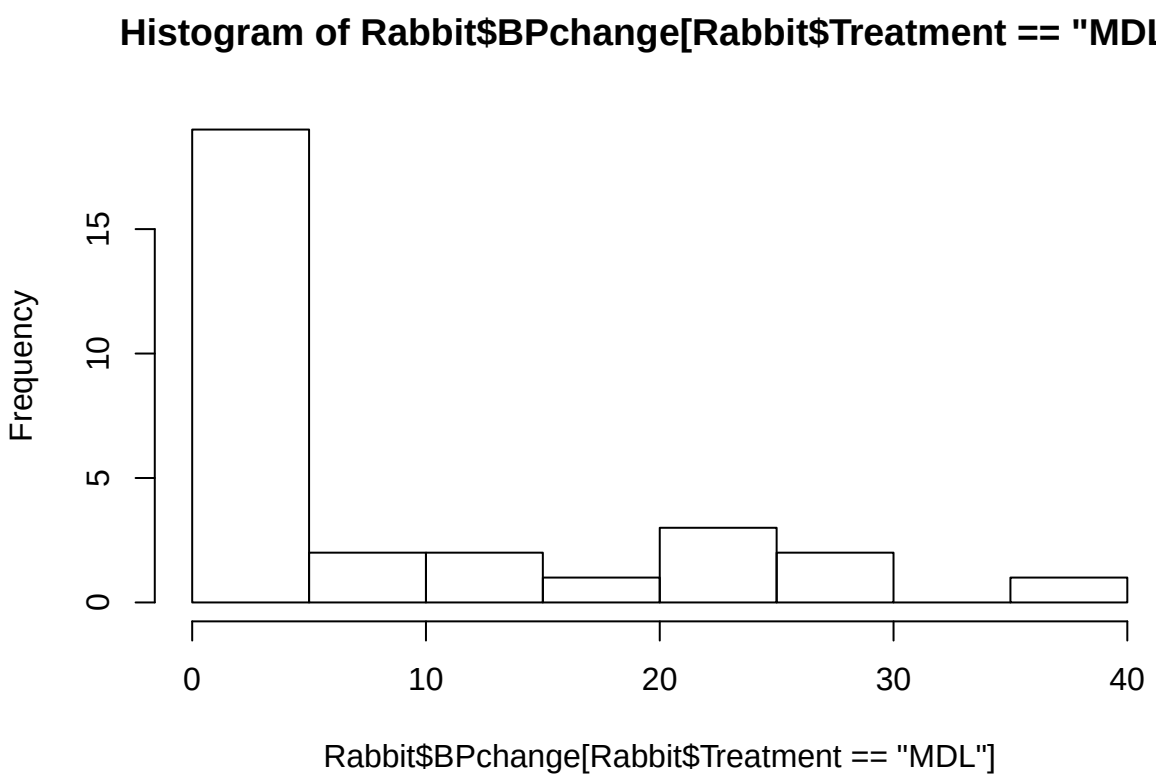
context of this dataset?(4pts)

Your Code:

```
hist(Rabbit$BPchange[Rabbit$Treatment == 'Control'])
```




```
hist(Rabbit$BPchange[Rabbit$Treatment == 'MDL'])
```




```
# The histograms do not resemble a normal distribution because the majority distribution of the
# histograms converge at the most left side instead of the central part.
# The histgram shape of both groups are similar, in which the majority is located at the range of
```

```
# 0~5, which means the situation of slight blood pressure change is the most frequent.
```

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)

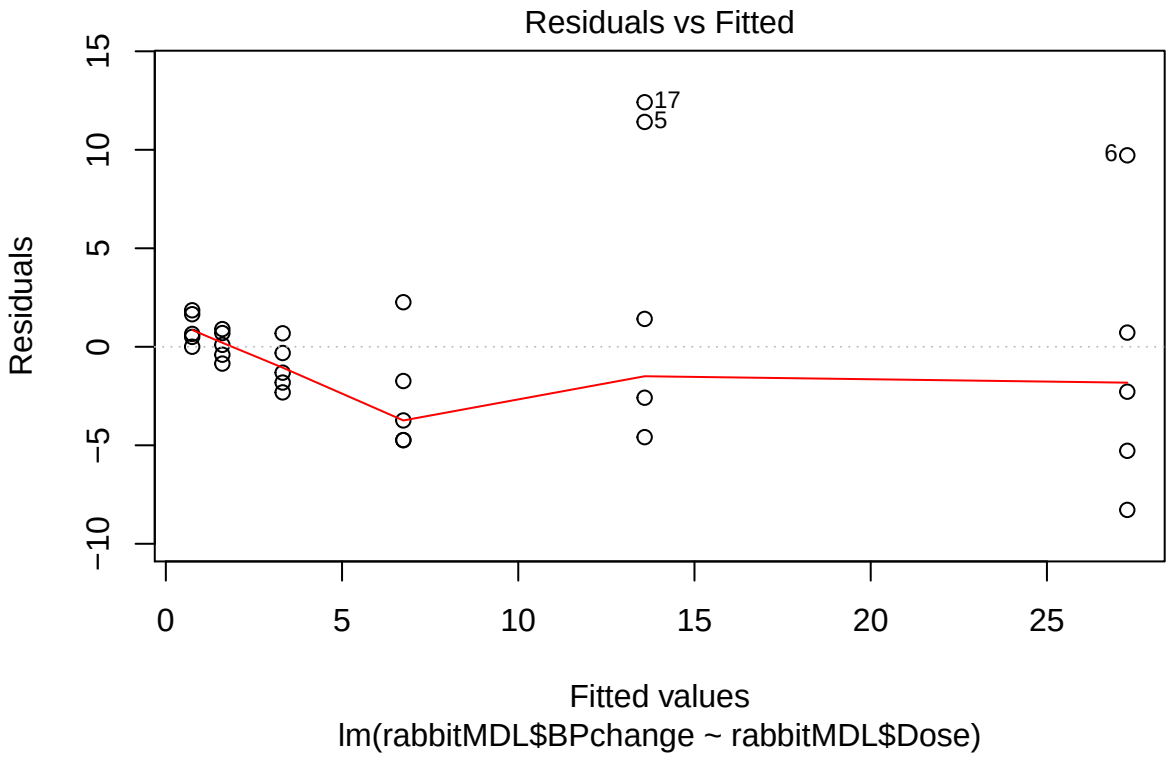
Would you consider this a good model or not? Please explain. (2pts)

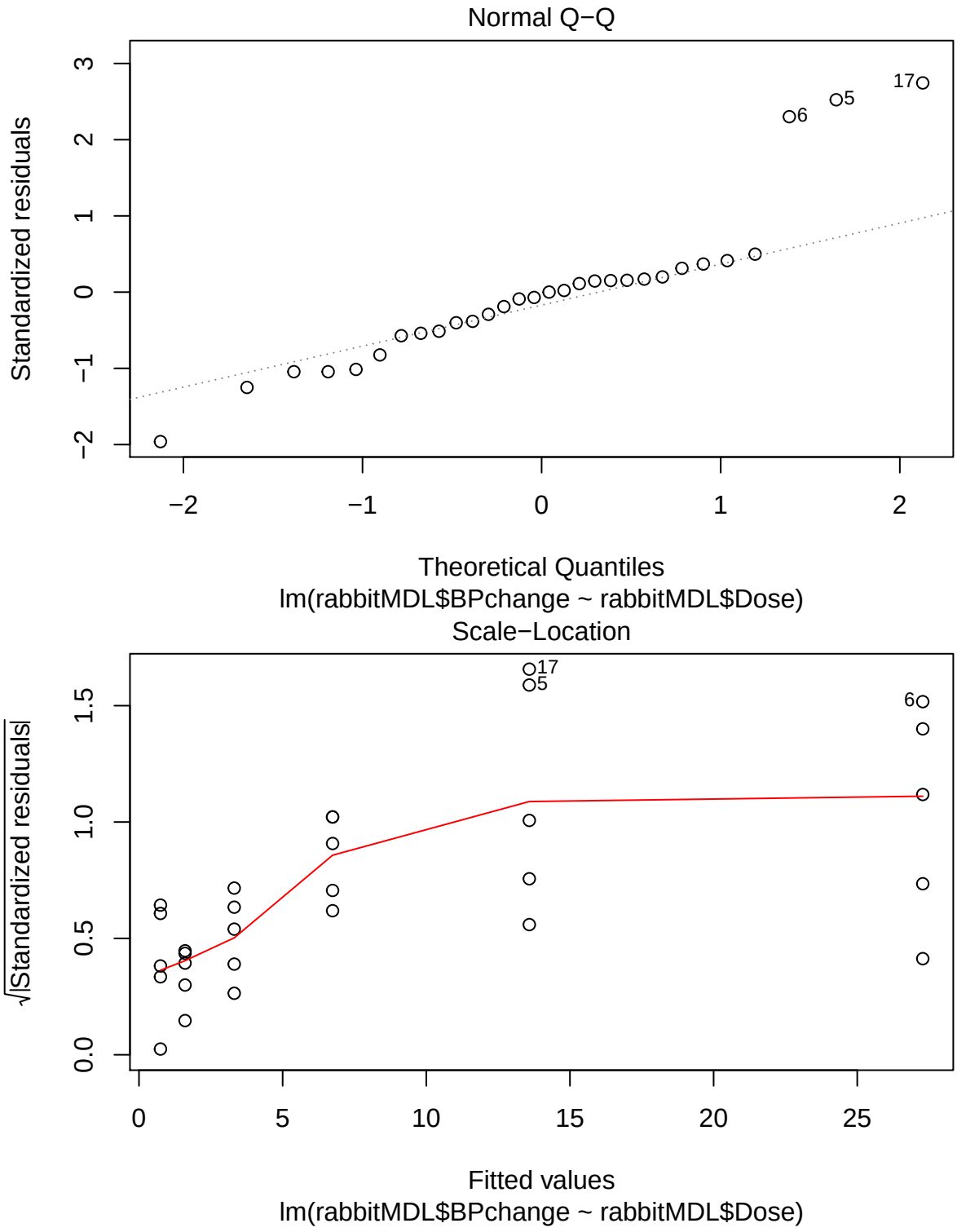
Your code:

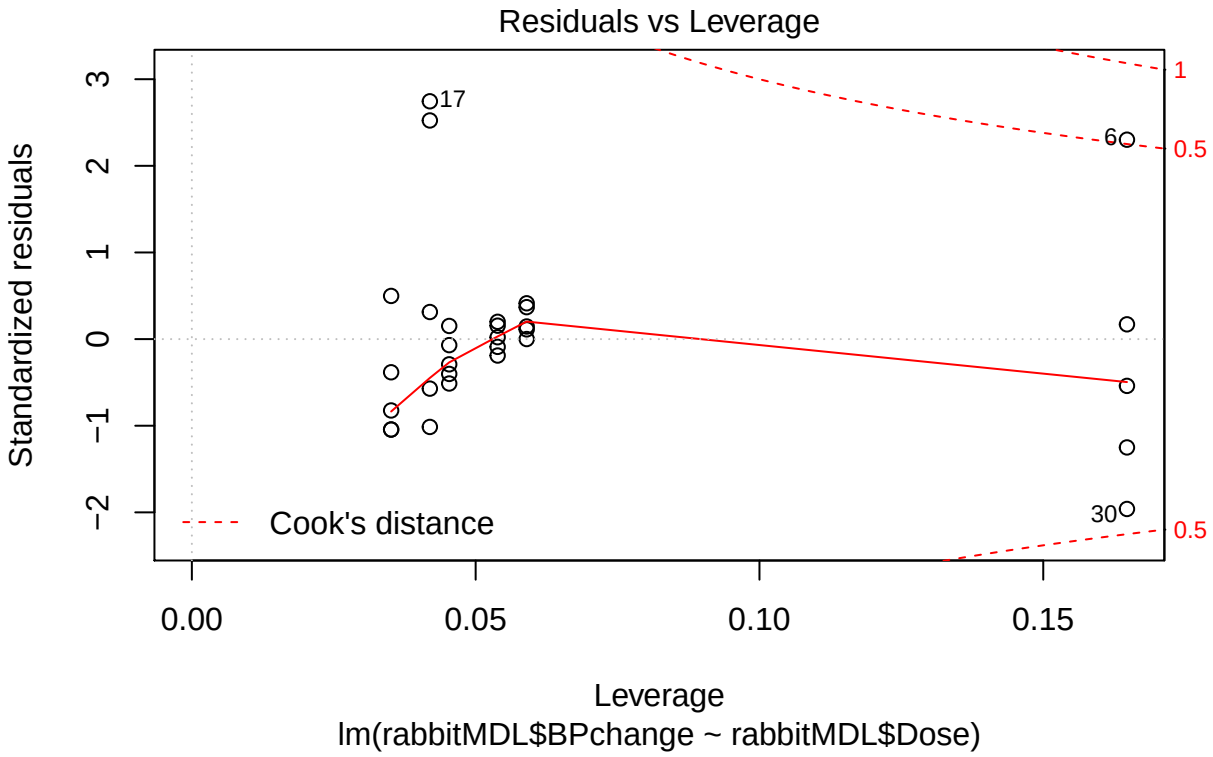
```r
rabbitMDL = Rabbit[Rabbit$Treatment == 'MDL',]
model_BP = lm(rabbitMDL$BPchange ~ rabbitMDL$Dose)
summary(model_BP)
```

```
##
## Call:
## lm(formula = rabbitMDL$BPchange ~ rabbitMDL$Dose)
##
## 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
## rabbitMDL$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
```

```r
plot(model_BP)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(rabbitMDL$BPchange ~ rabbitMDL$Dose)
Scale−Location
√|Standardized residuals|
Fitted values
lm(rabbitMDL$BPchange ~ rabbitMDL$Dose)




```
# According to corresponding fitted values, The residuals of this model distribute around 0, which is r
# The distribution of Normal Q-Q fits the trend of a line, which indicates the residuals are normally d
# There are 4 significant outliers: point 5, 6, 17, 30, which are supposed to be removed or adjusted
# while improving model performance.
# This model would be considered as good. In terms of p-value of Dose, it is fairly closed to 0, which
# the correlation between the two variables is strong. In addition, the adjusted R-squred value of
# this model is 80.5%, which indicates 80.5% of data are able to be predicted by this regression 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)

Your Code:

```
tvalue = qt(1-0.05/2,nrow(rabbitMDL)-2)
print(tvalue)
```

```
## [1] 2.048407
```

```
b = coef(summary(model_BP))[[2]]
seb = coef(summary(model_BP))[2,2]
```

```r
confintDose = c(b - tvalue*seb, b + tvalue*seb)
print(confintDose)
```

```
## [1] 0.1114109 0.1624768
```

Q9.

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

Your Answer:

```r
# The null hypothesis would be rejected because the p-value of the model is 1.159e-11, less than 0.001.
```