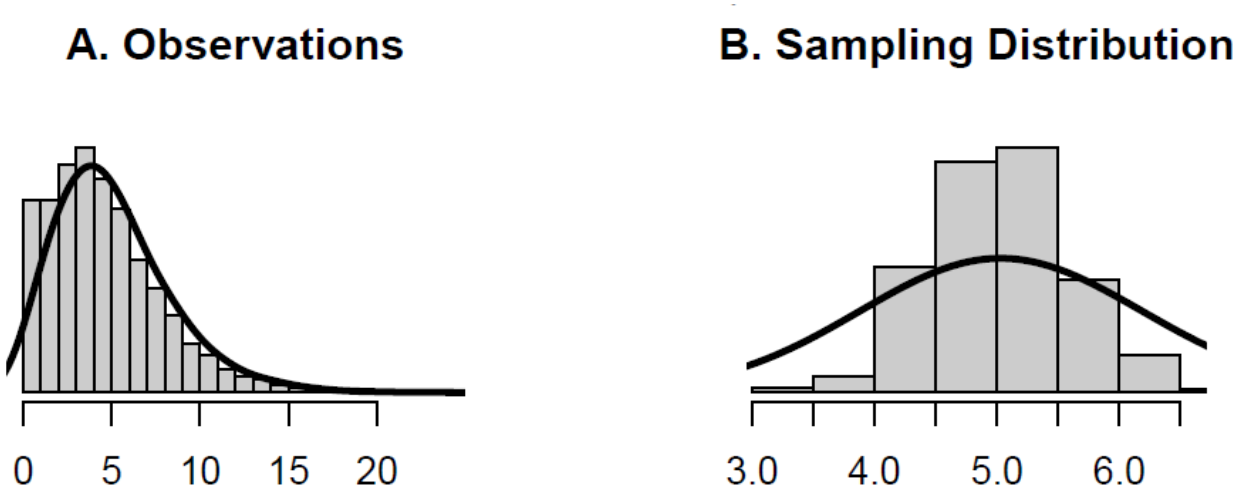
Final Exam-Lin

Bin Lin

2016-12-13

Part I Figure A below represents the distribution of an observed variable. Figure B below represents the distribution of the mean from 500 random samples of size 30 from A. The mean of A is 5.05 and the mean of B is 5.04. The standard deviations of A and B are 3.22 and 0.58, respectively.



Figure

a. Describe the two distributions (2 pts).

Figure A is unimodal, right skewed, and centered at around 4. The distribution ranges between 0 and 20, with its mean greater than its median. Figure B is also unimodal, symmetric and centered at around 5. The distribution appears to be normal and ranges between 3 to 6.5.

b. Explain why the means of these two distributions are similar but the standard deviations are not (2 pts).

Because according to central limit theorem, the distribution of the sample mean is well approximated by a normal model whose means equals to the population mean and whose standard error equals the population standard deviation divided by square root of n (sample size). That is why the means of these two distributions are similar but the standard deviations are not.

c. What is the statistical principal that describes this phenomenon (2 pts)?

Central limit theorem.

Part II Consider the four datasets, each with two columns (x and y), provided below.

```
options(digits=2)
data1 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68))
data2 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74))
data3 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73))
data4 <- data.frame(x=c(8,8,8,8,8,8,8,19,8,8,8),
                    y=c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6.89))
```

For each column, calculate (to two decimal places):

a. The mean (for x and y separately; 1 pt).

```
data_mean <- data.frame(x=c(sprintf("%.2f", mean(data1$x)), sprintf("%.2f", mean(data2$x)), spri
ntf("%.2f", mean(data3$x)), sprintf("%.2f", mean(data4$x))), y=c(sprintf("%.2f", mean(data1$y)),
 sprintf("%.2f", mean(data2$y)), sprintf("%.2f", mean(data3$y)), sprintf("%.2f",
mean(data4$y))))
data_mean
```

```
##      x    y
## 1 9.00 7.50
## 2 9.00 7.50
## 3 9.00 7.50
## 4 9.00 7.50
```

b. The median (for x and y separately; 1 pt).

```
data_median <- data.frame(x=c(sprintf("%.2f", median(data1$x)), sprintf("%.2f",
median(data2$x)), sprintf("%.2f", median(data3$x)), sprintf("%.2f", median(data4$x))), y=c(sprin
tf("%.2f", median(data1$y)), sprintf("%.2f", median(data2$y)), sprintf("%.2f", median(data3$y)),
 sprintf("%.2f", median(data4$y))))
data_median
```

```
##      x    y
## 1 9.00 7.58
## 2 9.00 8.14
## 3 9.00 7.11
## 4 8.00 7.04
```

c. The standard deviation (for x and y separately; 1 pt). For each x and y pair, calculate (also to two decimal places; 1 pt):

```
data_sd <- data.frame(x=c(sprintf("%.2f", sd(data1$x)), sprintf("%.2f", sd(data2$x)), sprintf("%.2
 sd(data3$x)), sprintf("%.2f", sd(data4$x))), y=c(sprintf("%.2f", sd(data1$y)), sprintf("%.2f", sd
data2$y)), sprintf("%.2f", sd(data3$y)), sprintf("%.2f", sd(data4$y))))
data_sd
```

```
##      x    y
## 1 3.32 2.03
## 2 3.32 2.03
## 3 3.32 2.03
## 4 3.32 2.03
```

d. The correlation (1 pt).

```
data_cor <- c(sprintf("%.2f", cor(data1$x, data1$y)), sprintf("%.2f", cor(data2$x, data2$y)), sp
rintf("%.2f", cor(data3$x, data3$y)), sprintf("%.2f", cor(data4$x, data4$y)))
data_cor
```

```
## [1] "0.82" "0.82" "0.82" "0.82"
```

e. Linear regression equation (2 pts).

Data1 linear reqression equation: y = 3.00 + 0.50 * x

Data2 linear reqression equation: y = 3.00 + 0.50 * x

Data3 linear reqression equation: y = 3.00 + 0.50 * x

Data4 linear reqression equation: y = 3.00 + 0.50 * x

```
data_lm1 <- lm(y ~ x, data1)
summary(data_lm1)
```

```
##
## Call:
## lm(formula = y ~ x, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9213 -0.4558 -0.0414  0.7094  1.8388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.000      1.125    2.67   0.0257 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00217
```

```
data_lm2 <- lm(y ~ x, data2)
summary(data_lm2)
```

```
##
## Call:
## lm(formula = y ~ x, data = data2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.901 -0.761  0.129  0.949  1.269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125    2.67   0.0258 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00218
```

```
data_lm3 <- lm(y ~ x, data3)
summary(data_lm3)
```

```
##
## Call:
## lm(formula = y ~ x, data = data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.159 -0.615 -0.230  0.154  3.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002      1.124    2.67   0.0256 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00218
```

```
data_lm4 <- lm(y ~ x, data4)
summary(data_lm4)
```

```
##
## Call:
## lm(formula = y ~ x, data = data4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002      1.124    2.67   0.0256 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 9 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.63
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00216
```

f. R-Squared (2 pts).

Data1 R-Squared:0.6665

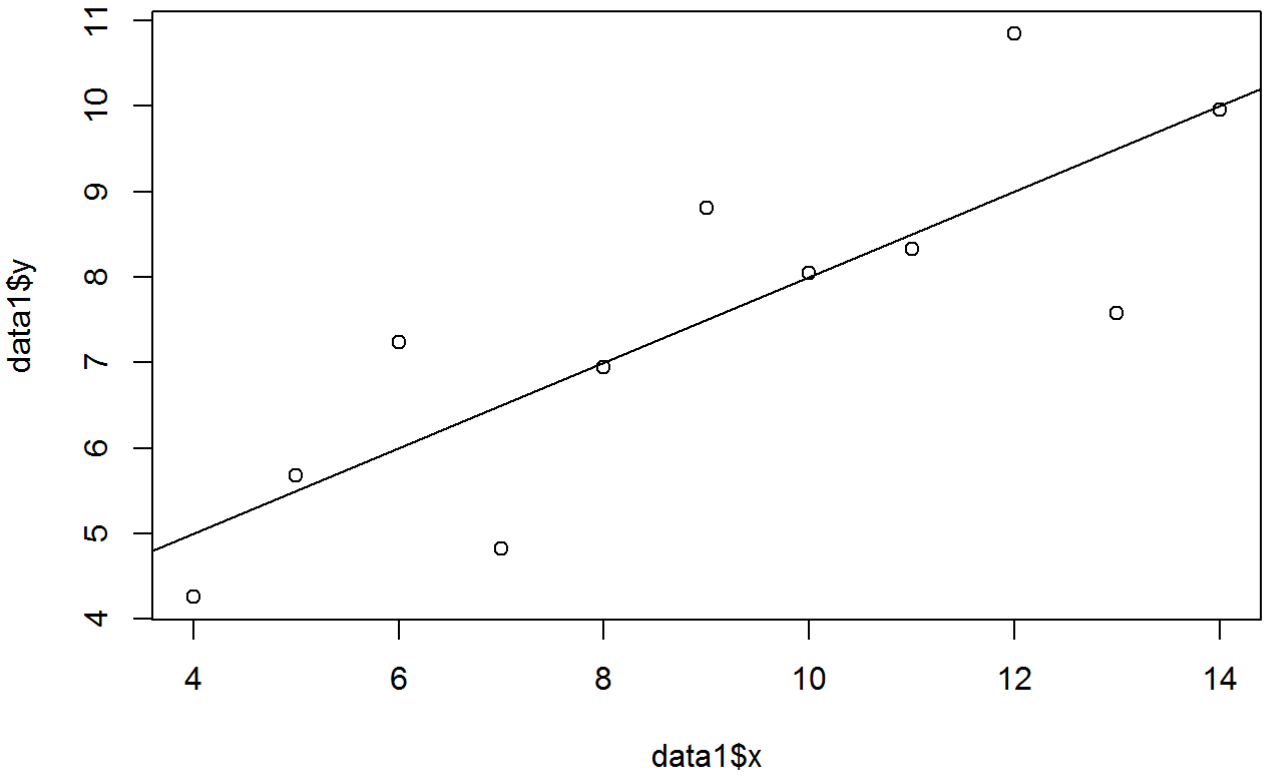
Data2 R-Squared:0.6662

Data3 R-Squared:0.6663

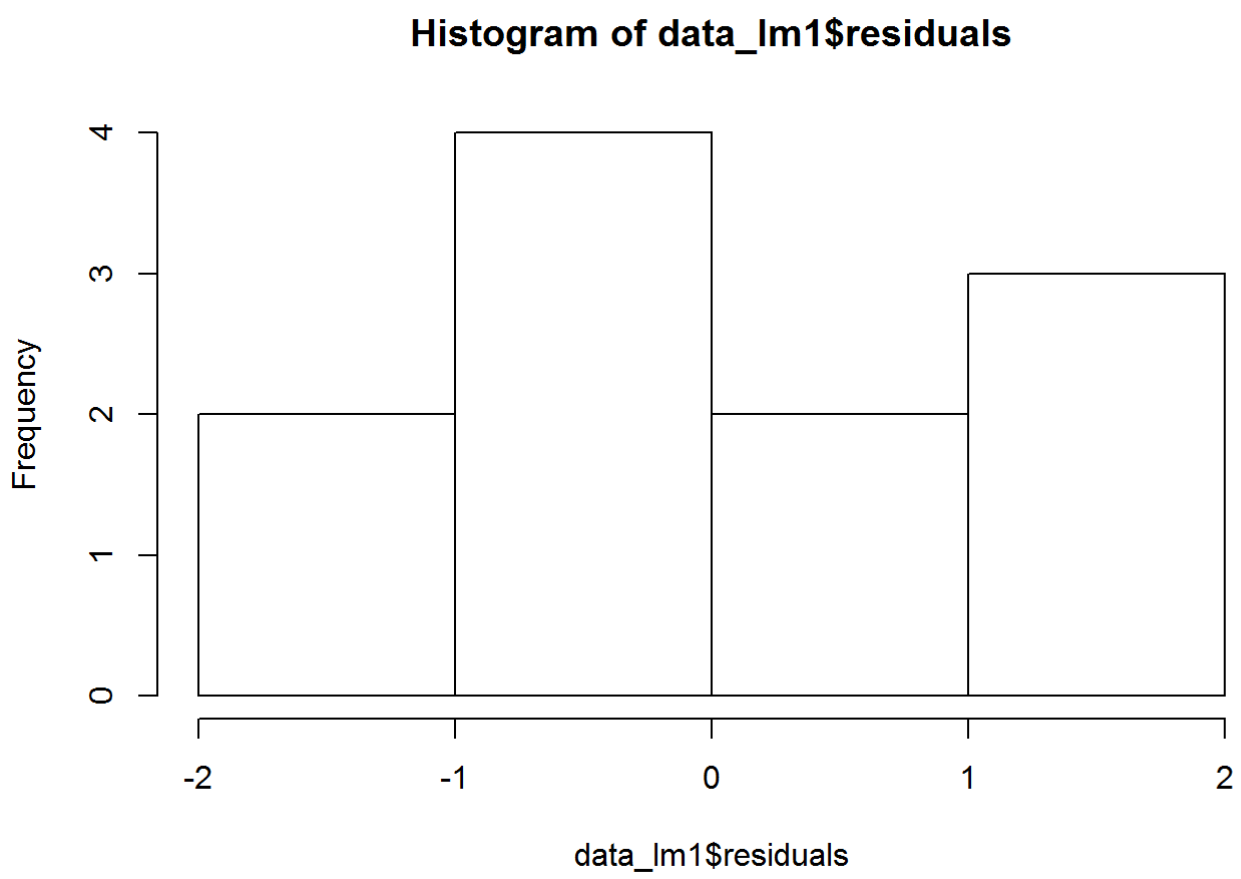
Data4 R-Squared:0.6667

For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

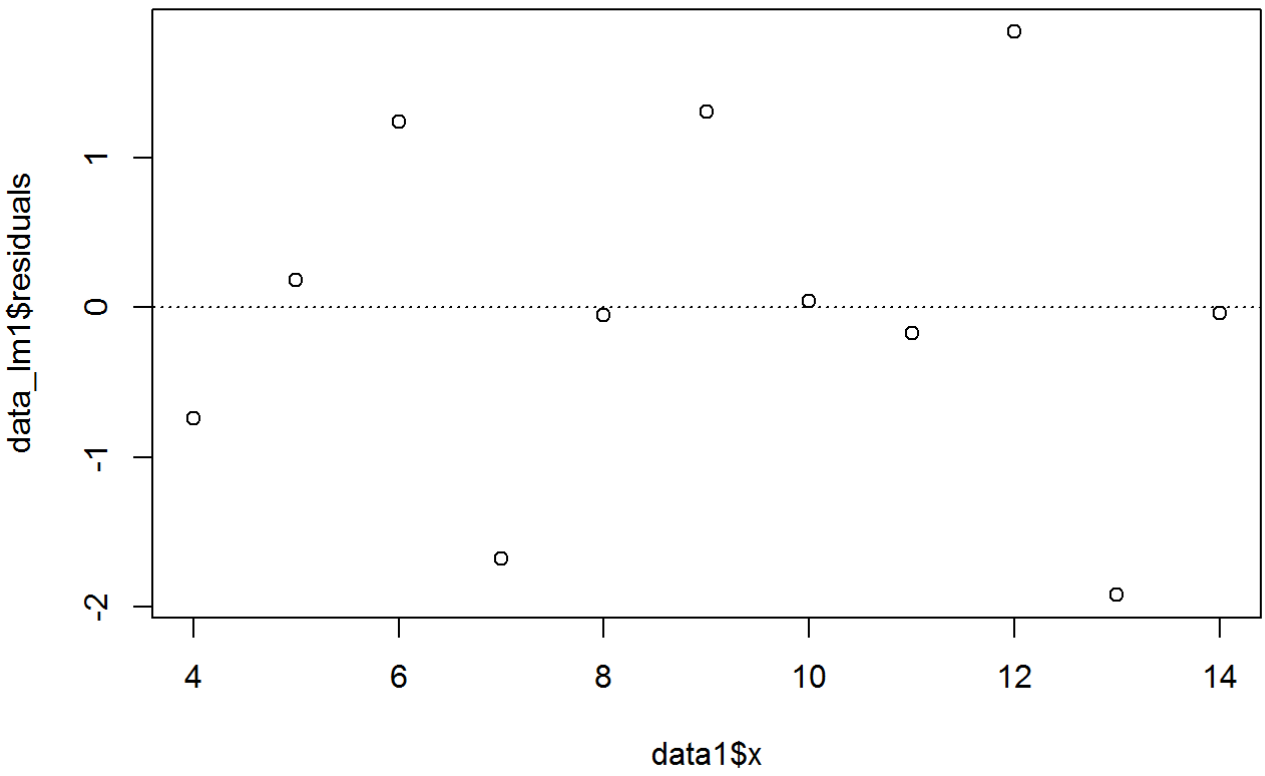
```
plot(data1$y ~ data1$x)
abline(data_lm1)
```



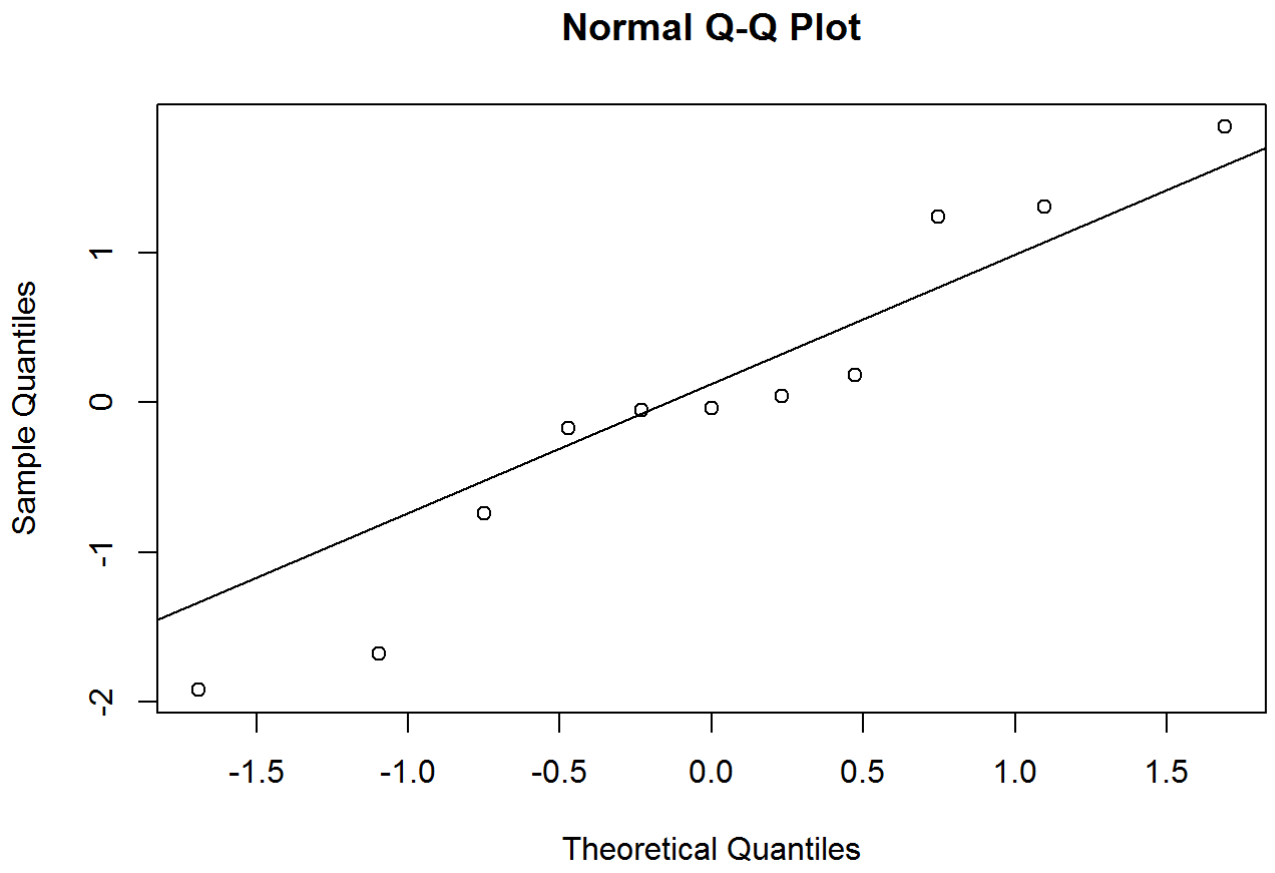

```
hist(data_lm1$residuals)
```




```
plot(data_lm1$residuals ~ data1$x)
abline(h = 0, lty = 3)
```




```
qqnorm(data_lm1$residuals)
qqline(data_lm1$residuals)
```



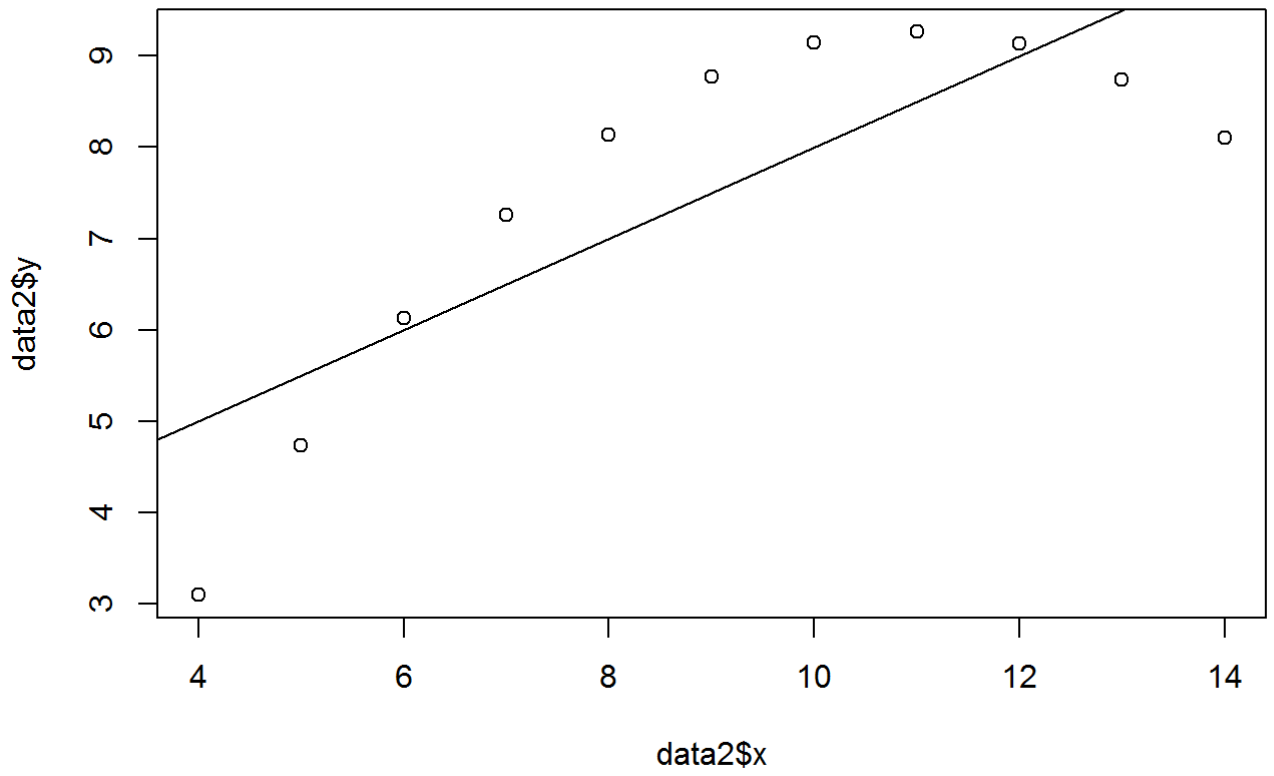
Data1:

Linearity: The scatterplot shows linear relationship, and the residual plot shows no patten, and all the data points are evenly distributed on both side of the 0 line.

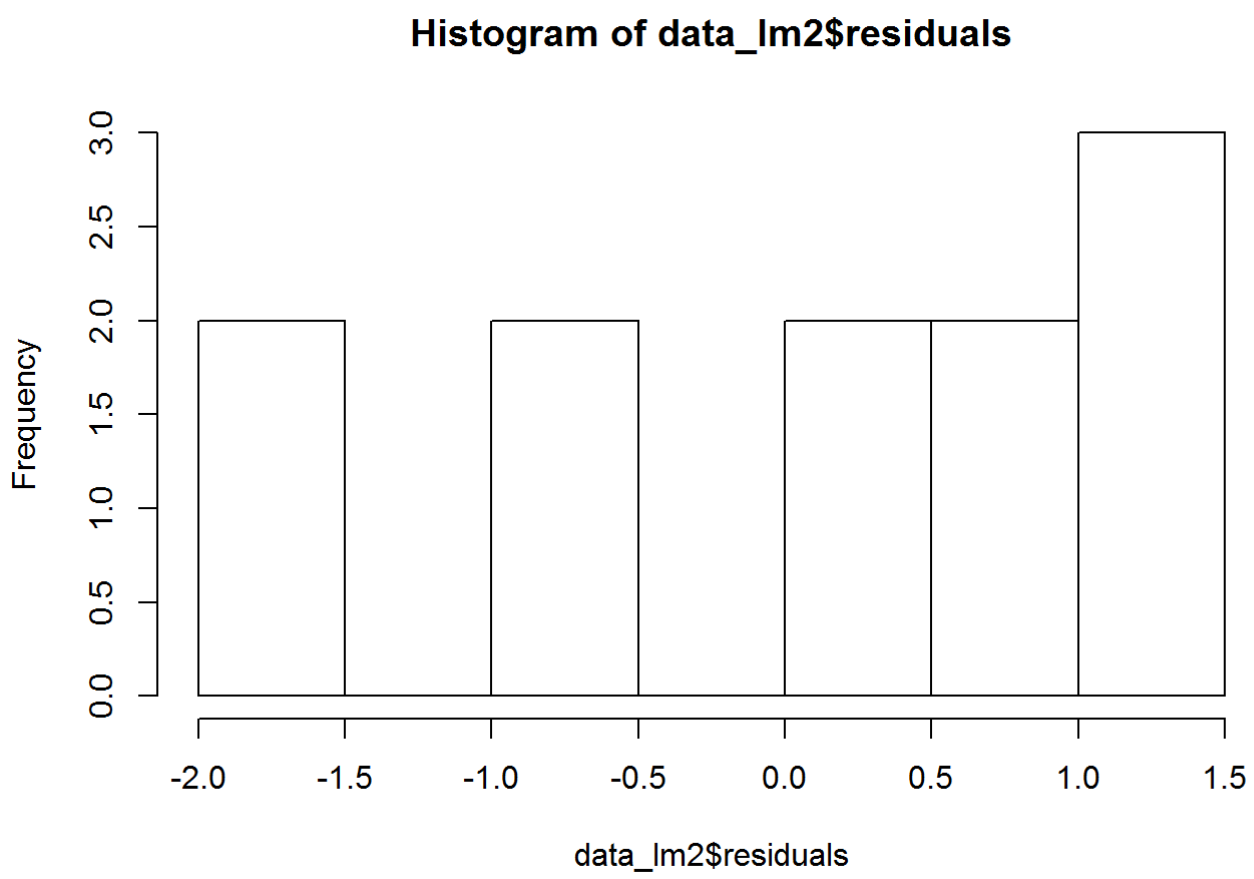
Nearly normal residuals: The histogram of the residual shows almost unimodel and bell shaped distribution that is quite symmetric. The normal probability plot also indicate normal distribution of residuals.

Constant variability: The variability of residuals around the 0 line is roughly constant. No pattern or fan shape observed.

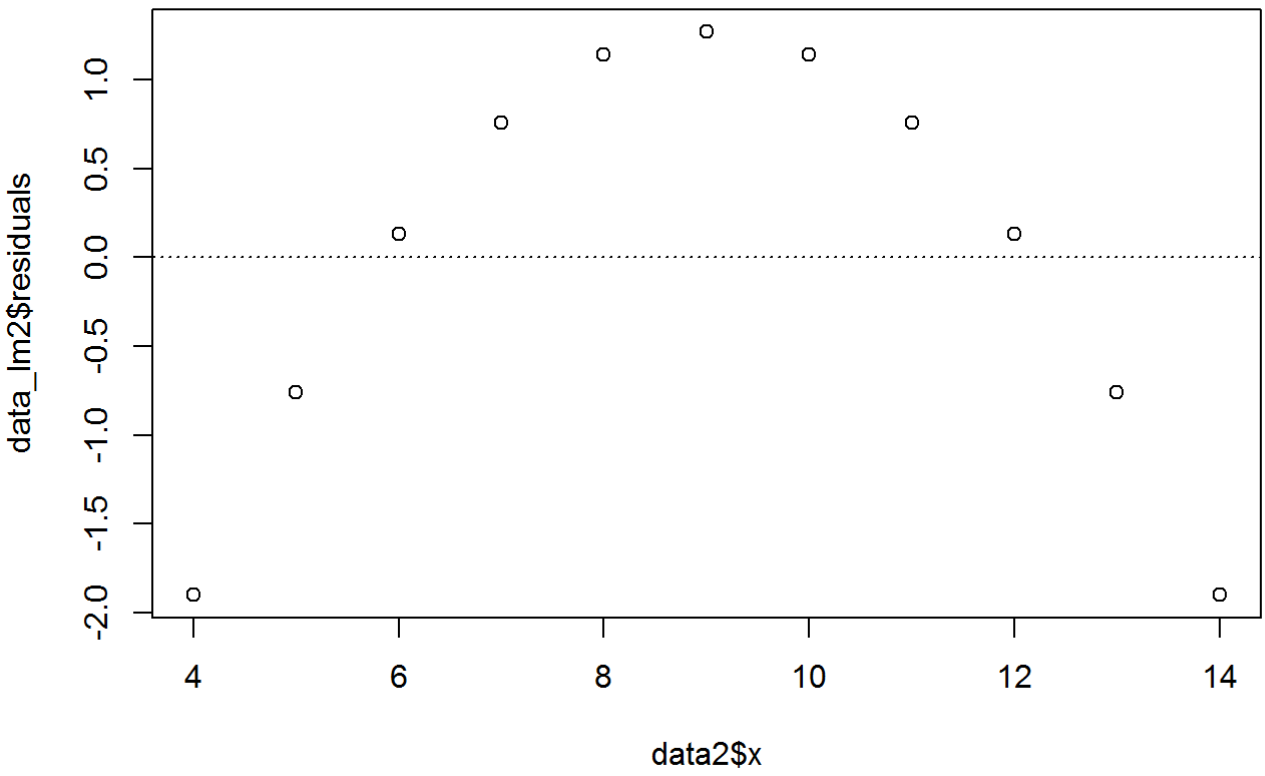
```
plot(data2$y ~ data2$x)
abline(data_lm2)
```



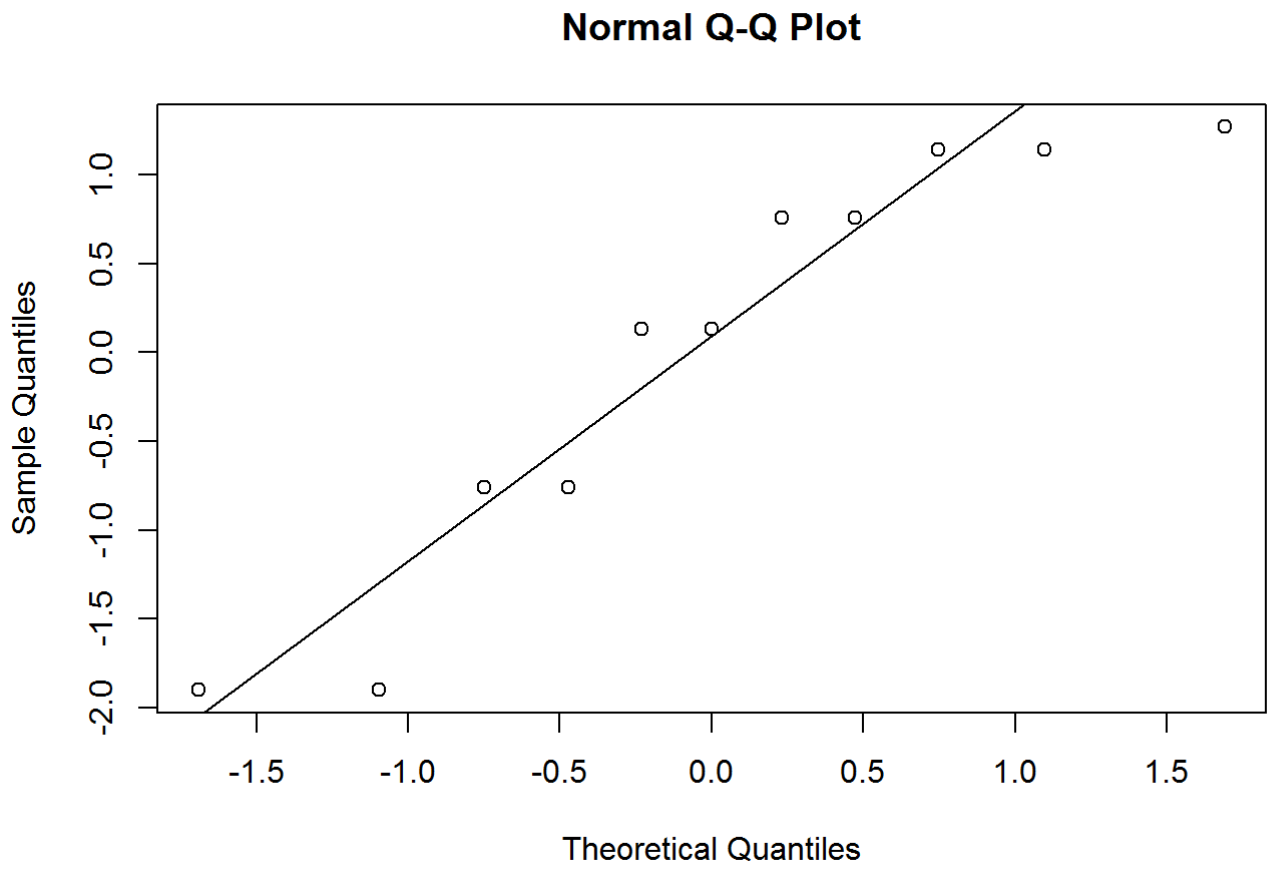

```
hist(data_lm2$residuals)
```




```
plot(data_lm2$residuals ~ data2$x)
abline(h = 0, lty = 3)
```




```
qqnorm(data_lm2$residuals)
qqline(data_lm2$residuals)
```



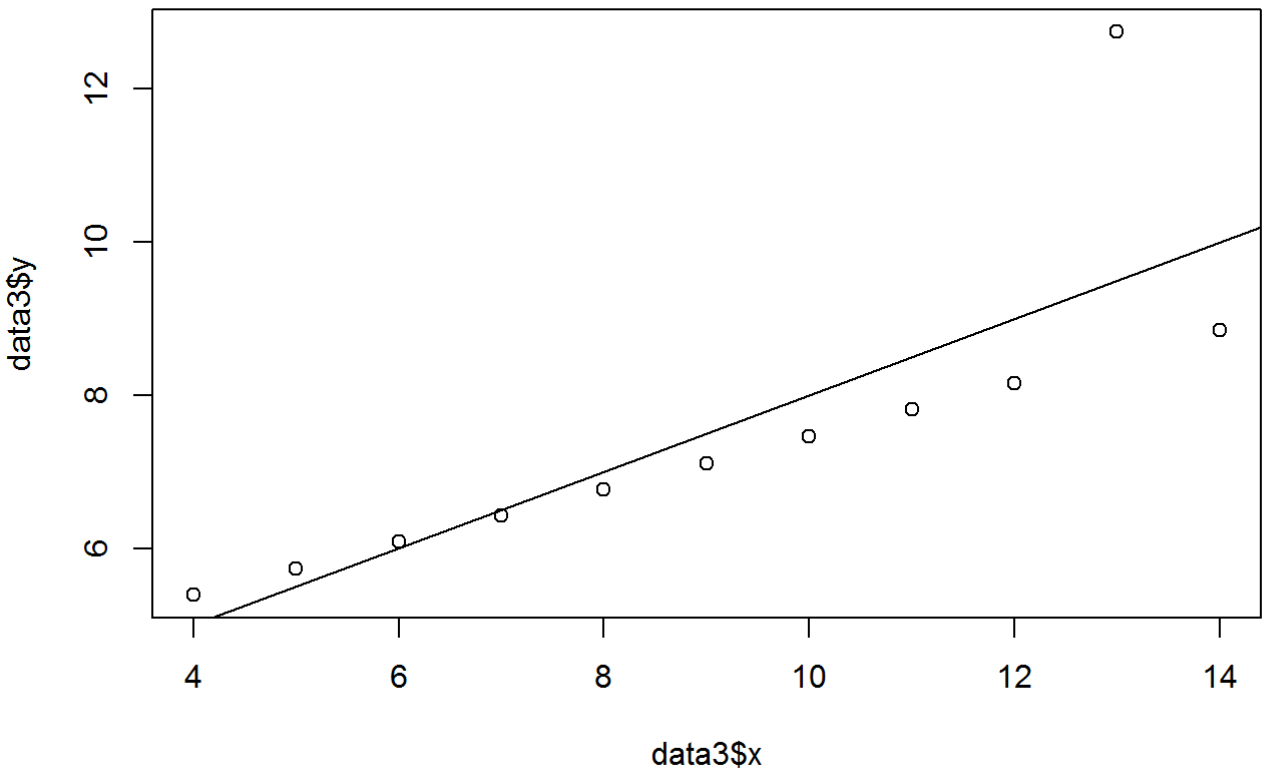
Data2:

Linearity: The scatterplot shows non-linear relationship, and the residual plot shows a parabola patten,

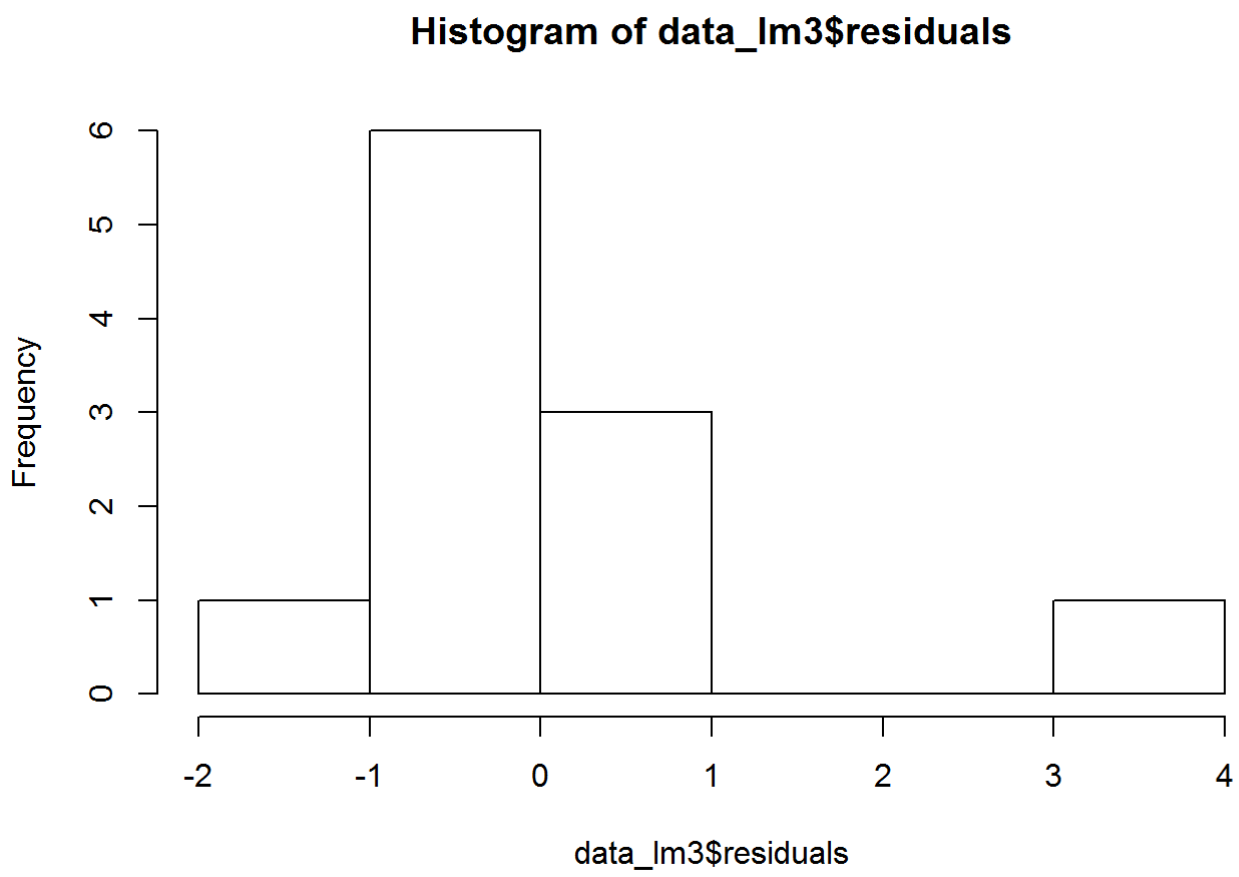
Nearly normal residuals: The histogram of the residual is uniformly distributed. Many points are far away from the normal probability plot.

Constant variability: The variability of residuals is not constant. And the residual plot shows there is a pattern exist.

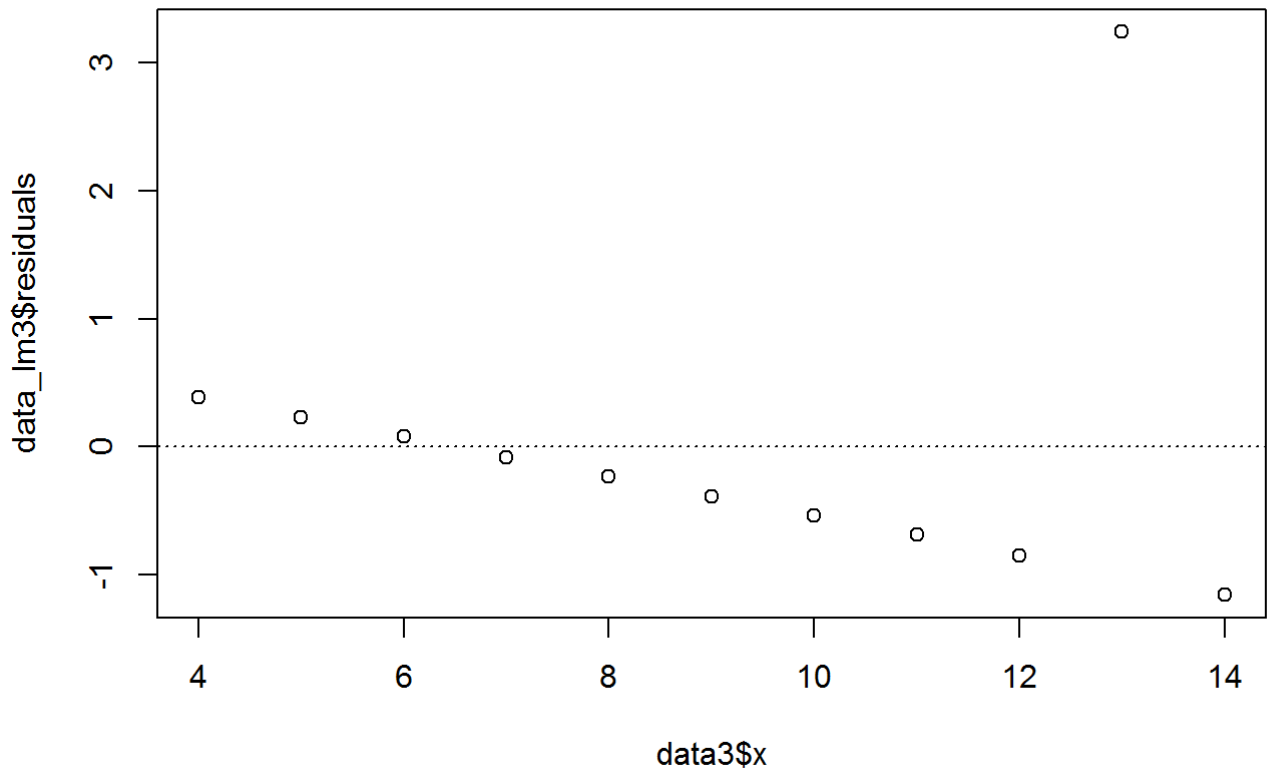
```
plot(data3$y ~ data3$x)
abline(data_lm3)
```



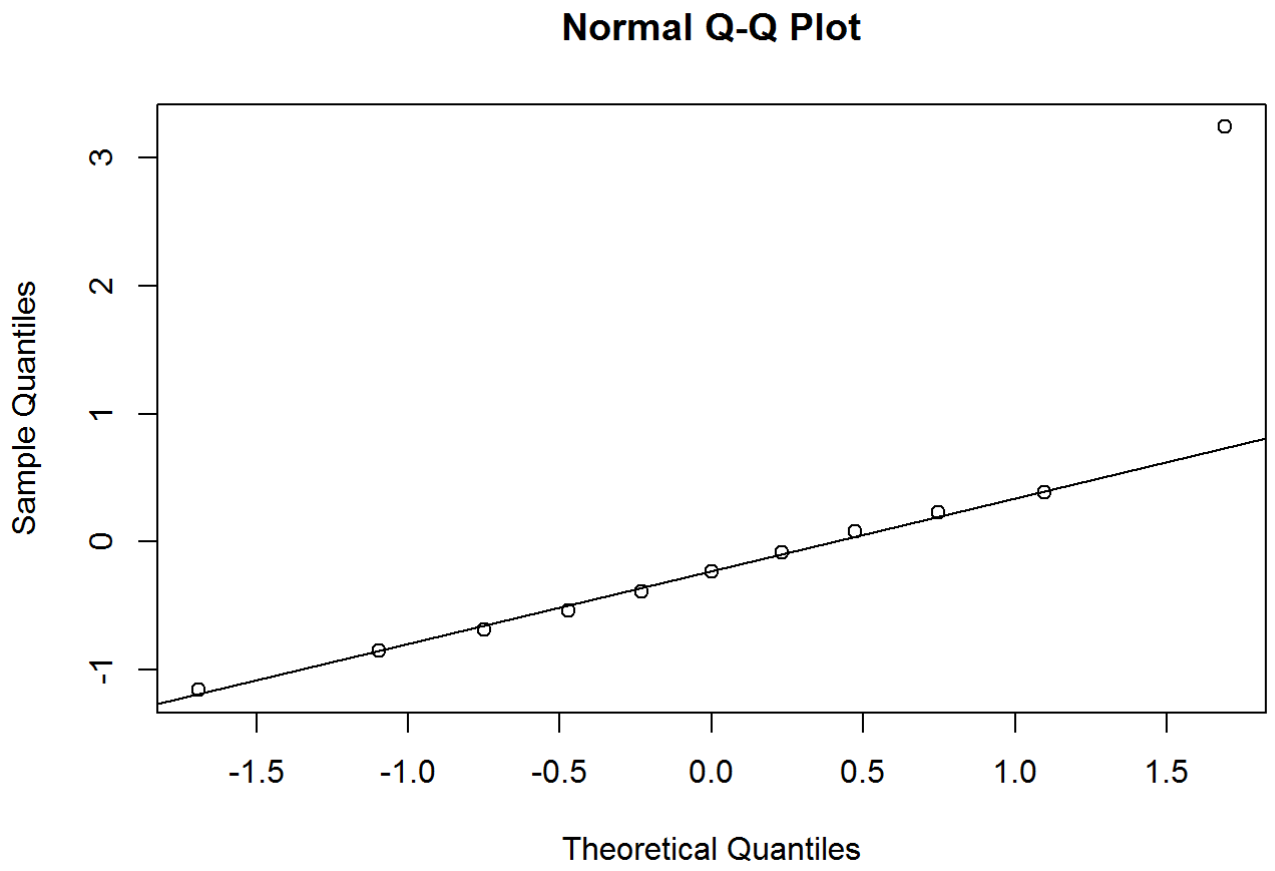

```
hist(data_lm3$residuals)
```




```
plot(data_lm3$residuals ~ data3$x)
abline(h = 0, lty = 3)
```




```
qqnorm(data_lm3$residuals)
qqline(data_lm3$residuals)
```



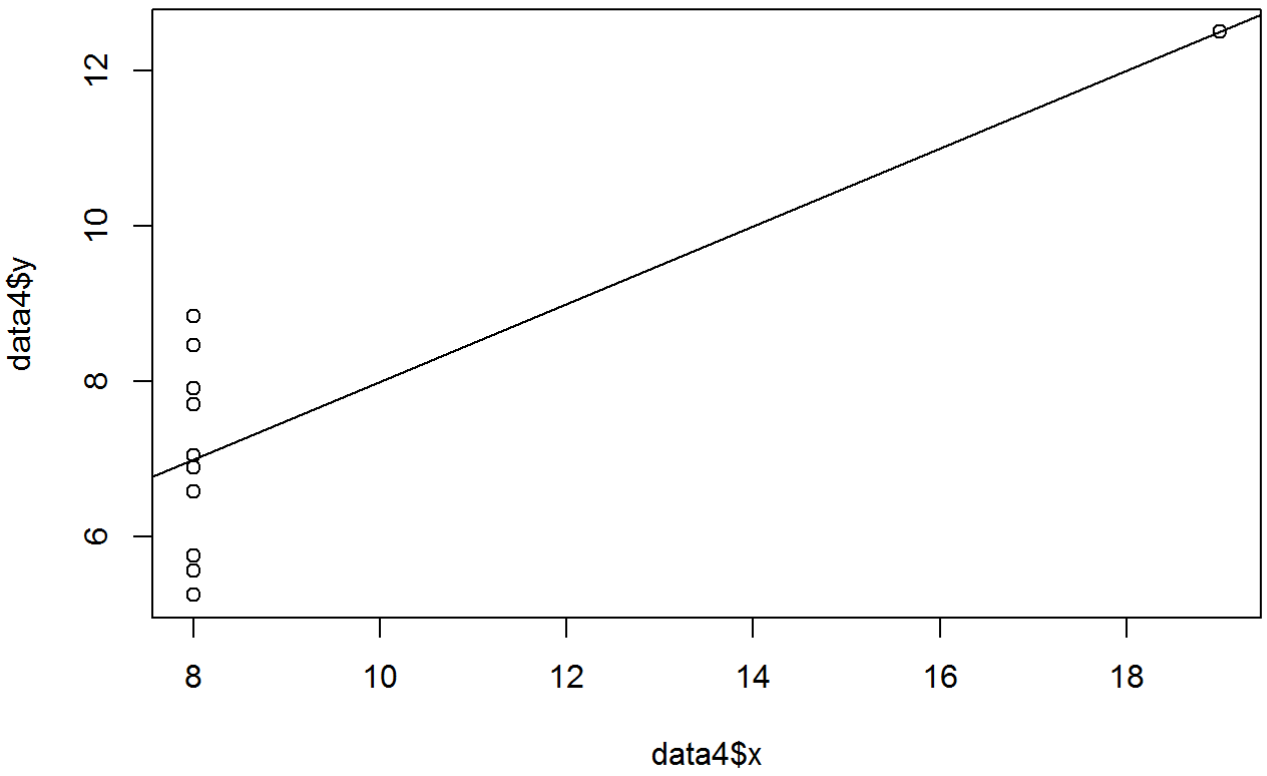
Data3:

Linearity: The scatterplot shows linear relationship, however, there is an outlier that significantly affect the slope of the regression line. In addition, the residual plot shows patten.

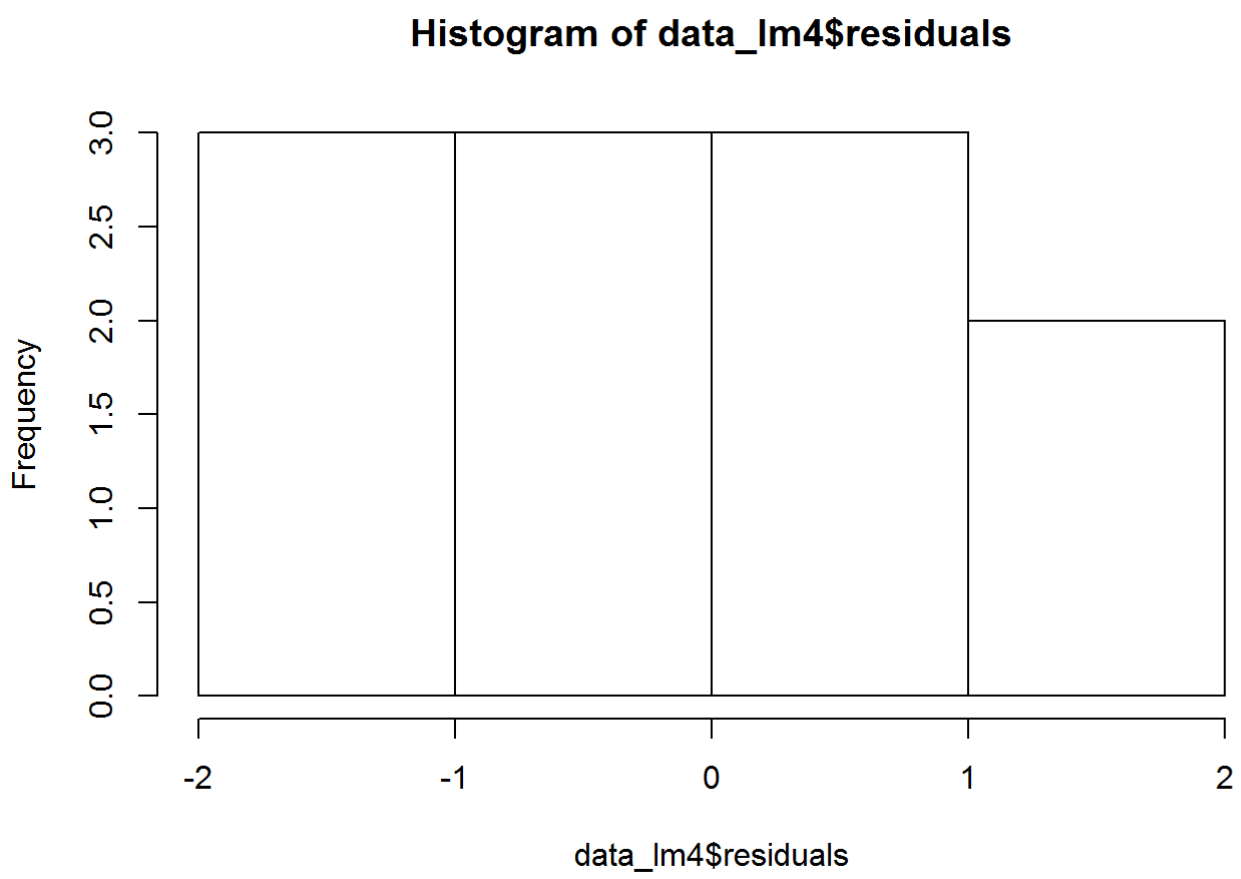
Nearly normal residuals: The histogram of the residual shows almost unimodel and bell shaped distribution. However, there is clear outlier exist in the graph. That outlier can also be observed in the normal probability plot.

Constant variability: The variability of residuals around the 0 line is constantly changing and appear to have pattern.

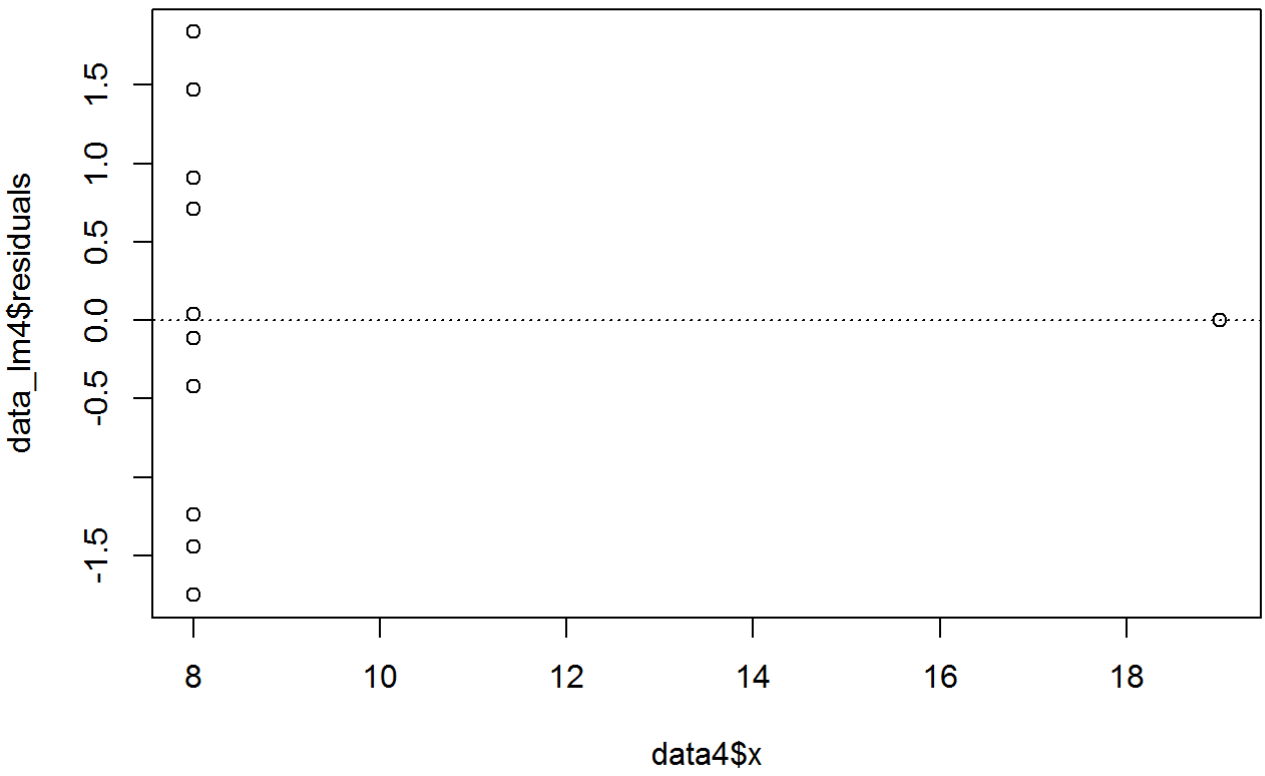
```
plot(data4$y ~ data4$x)
abline(data_lm4)
```



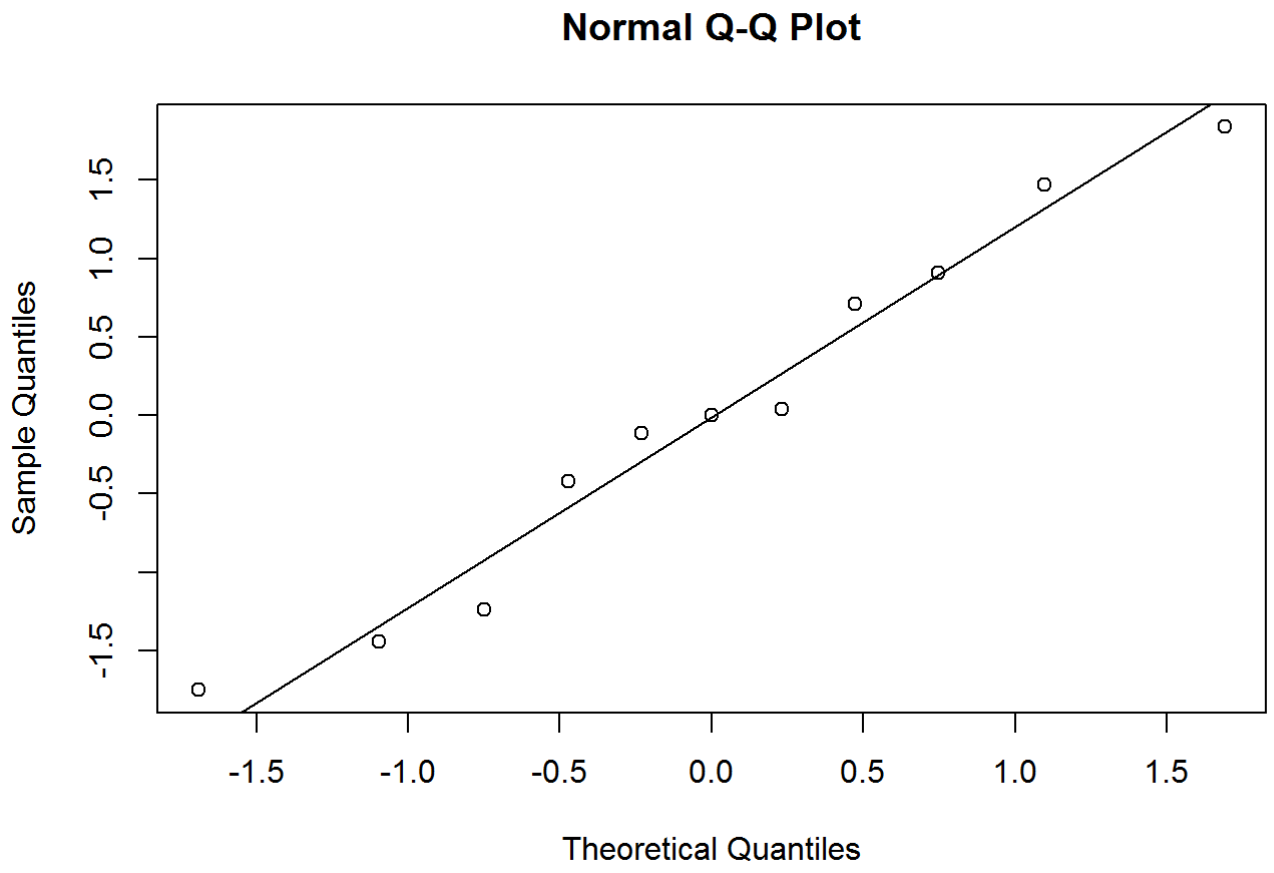

```
hist(data_lm4$residuals)
```




```
plot(data_lm4$residuals ~ data4$x)
abline(h = 0, lty = 3)
```




```
qqnorm(data_lm4$residuals)
qqline(data_lm4$residuals)
```



Linearity: The scatterplot shows vertical linear relationship with an outlier, and the residual plot shows patten, and all the data points except the outlier line up at x=8.

Nearly normal residuals: The histogram of the residual shows uniform distribution. The normal probability plot indicates normal distribution of residuals.

Constant variability: The variability of residuals around the 0 line is constantly changing and appear to have pattern.

Conclusions: Only Dataset 1 is appropriate to estimate a linear regression model. All the others are not appropriate for a linear regression model.

Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (2 pts)

It is important to include appropriate visualizations when analyzing data, because in this case, all 4 datasets will generate same formulas for linear regression model. Not only the regression line, the mean, standard deviation, correlation except median are all the same across the board. However, after we create some graphs such as scatterplot, histogram, residual plot, and qqplot, we are able to tell the distribution and relationship of the data. For example, the scatterplot will clear shows if the explanatory and response variables have linear relationship or not. In addition, from the visualization, it is easier for us to tell if the models we create (could be linear or non-linear model) fit the datasets and if any particular data points are outliers which can be removed from dataset.