Final Exam

Vivian Kong

12/14/2016

Data 606 Final Exam

Part I

a. Describe the two distributions

Graph A is unimodal and highly Skewed right. In the skewed right graph, mode < median < mean. We'll have outliers is greater than 17.93. (4 standard deviation above the mean)
Graph B is approximately normal, so it is symmetric. Mean, median and mode should be all aound 5.04.

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

Distribution A approximate the mean from the population, and Distribution B has a relatively large sample size, so the mean is approximately same as the Distribution A's mean.

Distribution A is the sample standard deviaiton Distribution B is the standard deviation of the the sampling distribution mean, also we can call it standard error. To Find the standard error of the distribution B, we need to use the standard deviation of A divide by the square root of the sample size. SE= 3.22/sqrt(30)=0.58. Also, We can apply the central limit theorem for the distribution B since its sample size is relatively large (n > = 30). That's why we can find the distribution B is an approximately normal distribution.

c. What is the statistical principal that describes this phenomenon.

Distribution B can apply the Central Limit Theorem. It has a sample of at least 30 independent observations. No matter what shape of the Distribution A is, we will have an approximately normal curve.

Part II

```r
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))
```

a. The mean

Data 1

```
round(mean(data1$x), digits=2)

## [1] 9

round(mean(data1$y), digits =2)

## [1] 7.5
```

Data 2

```
round(mean(data2$x), digits=2)

## [1] 9

round(mean(data2$y), digits =2)

## [1] 7.5
```

Data 3

```
round(mean(data3$x), digits=2)

## [1] 9

round(mean(data3$y), digits =2)

## [1] 7.5
```

Data 4

```
round(mean(data4$x), digits=2)

## [1] 9

round(mean(data4$y), digits =2)

## [1] 7.5
```

b. The median

data 1

```
round(median(data1$x), digits=2)

## [1] 9

round(median(data1$y), digits=2)

## [1] 7.58
```

data 2

```
round(median(data2$x), digits=2)
```

```
## [1] 9

round(median(data2$y), digits=2)

## [1] 8.14
```

data 3

```
round(median(data3$x), digits=2)

## [1] 9

round(median(data3$y), digits=2)

## [1] 7.11
```

data 4

```
round(median(data4$x), digits=2)

## [1] 8

round(median(data4$y), digits=2)

## [1] 7.04
```

c. The standard deviation

Data 1

```
round(sd(data1$x), digits =2)

## [1] 3.32

round(sd(data1$y), digits =2)

## [1] 2.03
```

Data 2

```
round(sd(data2$x), digits =2)

## [1] 3.32

round(sd(data2$y), digits =2)

## [1] 2.03
```

Data 3

```
round(sd(data3$x), digits =2)

## [1] 3.32

round(sd(data3$y), digits =2)

## [1] 2.03
```

Data 4

```
round(sd(data4$x), digits =2)

## [1] 3.32

round(sd(data4$y), digits =2)

## [1] 2.03
```

d. The correlation

Data 1

```
round(cor(data1$x, data1$y), digits = 2)

## [1] 0.82
```

Data 2

```
round(cor(data2$x, data2$y), digits = 2)

## [1] 0.82
```

Data 3

```
round(cor(data3$x, data3$y), digits = 2)

## [1] 0.82
```

Data 4

```
round(cor(data4$x, data4$y), digits = 2)

## [1] 0.82
```

e. Linear Regression Equation

Data 1

```
equation.lm1<- lm(y ~ x, data=data1)
coefficients <- coefficients(equation.lm1)
coefficients

## (Intercept)           x
##   3.0000909   0.5000909
```

$$y' = 0.50x + 3.00$$

Data 2

```
equation.lm2 <- lm(y ~ x, data=data2)
coefficients <- coefficients(equation.lm2)
coefficients

## (Intercept)           x
##    3.000909    0.500000
```

$$y' = 0.50x + 3.00$$

Data 3

```
equation.lm3 <- lm(y ~ x, data=data3)
coefficients <- coefficients(equation.lm3)
coefficients

## (Intercept)           x
##   3.0024545   0.4997273
```

$$y' = 0.50x + 3.00$$

Data 4

```
equation.lm4 <- lm(y ~ x, data=data4)
coefficients <- coefficients(equation.lm4)
coefficients

## (Intercept)           x
##   3.0017273   0.4999091
```

$$y' = 0.50x + 3.00$$

f. R-Squared

data 1

```
equation.lm <- lm(y ~ x, data=data1)
round(summary(equation.lm)$r.squared, digits = 2)

## [1] 0.67
```

data 2

```
equation.lm <- lm(y ~ x, data=data2)
round(summary(equation.lm)$r.squared, digits = 2)

## [1] 0.67
```

data 3

```
equation.lm <- lm(y ~ x, data=data3)
round(summary(equation.lm)$r.squared, digits = 2)

## [1] 0.67
```

data 4

```
equation.lm <- lm(y ~ x, data=data4)
round(summary(equation.lm)$r.squared, digits =2)

## [1] 0.67
```

G. For each pair, is it appropriate to estimate a linear regression model? Why or why not?

1. Linearity: The data should show a linear trend

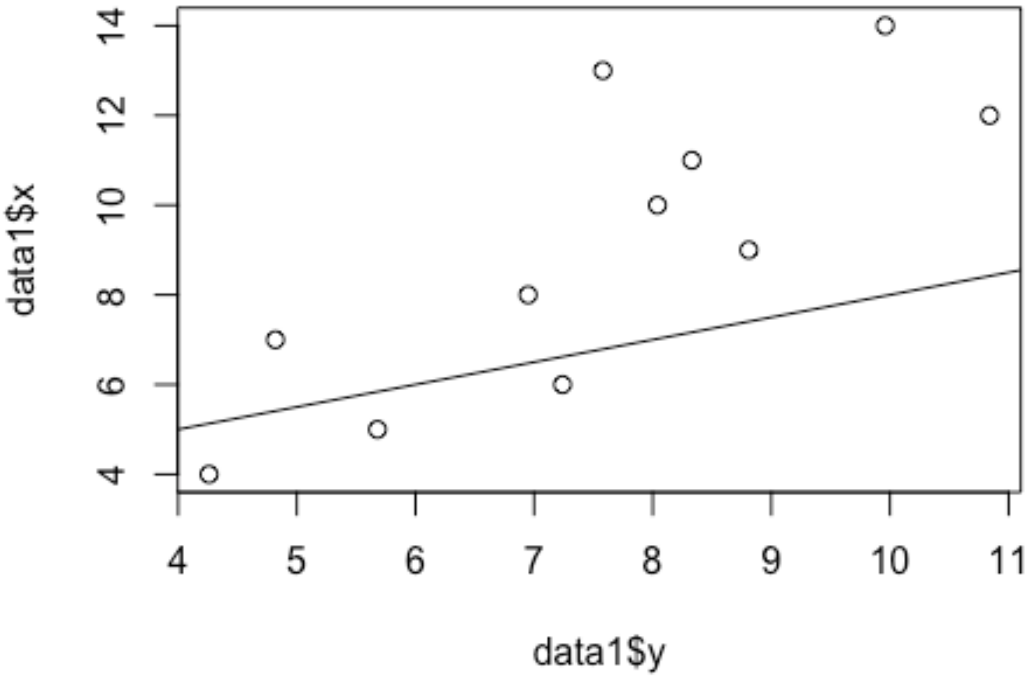
2. Nearly Normal Residuals: Generally the residuals must be nearly normal

3. Constant variability: The variability of points around the least square line remains roughly constant.

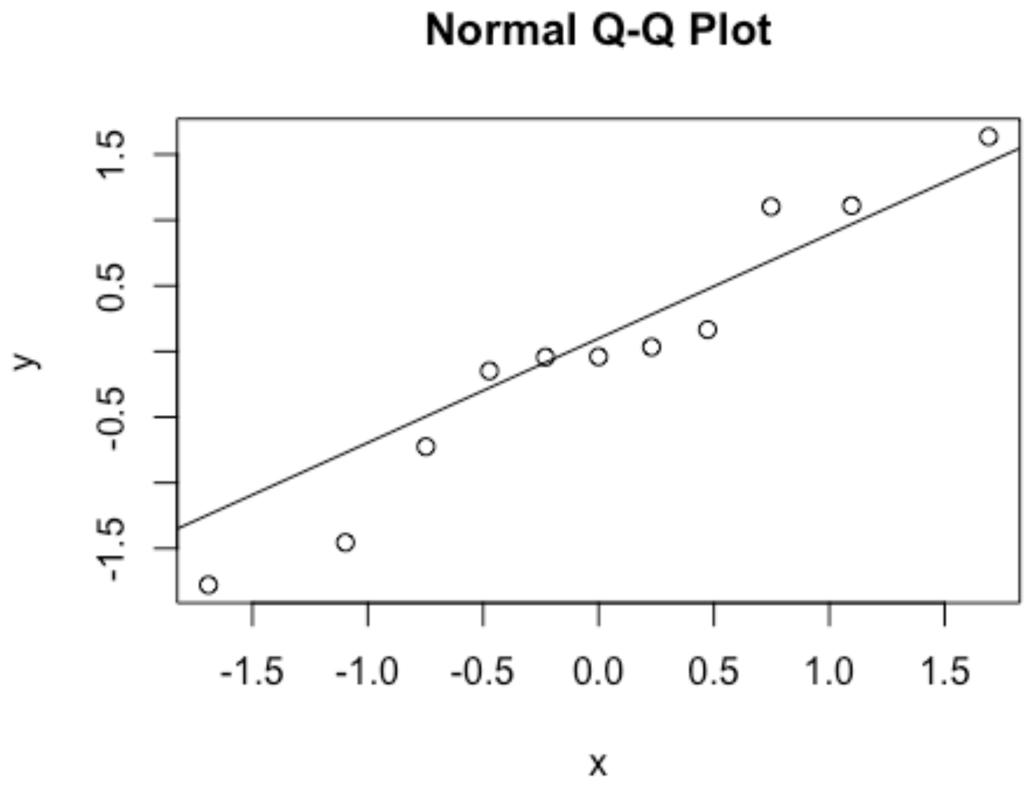
4. Independent Obervations The observations of the data set must be independent.

For Data 1

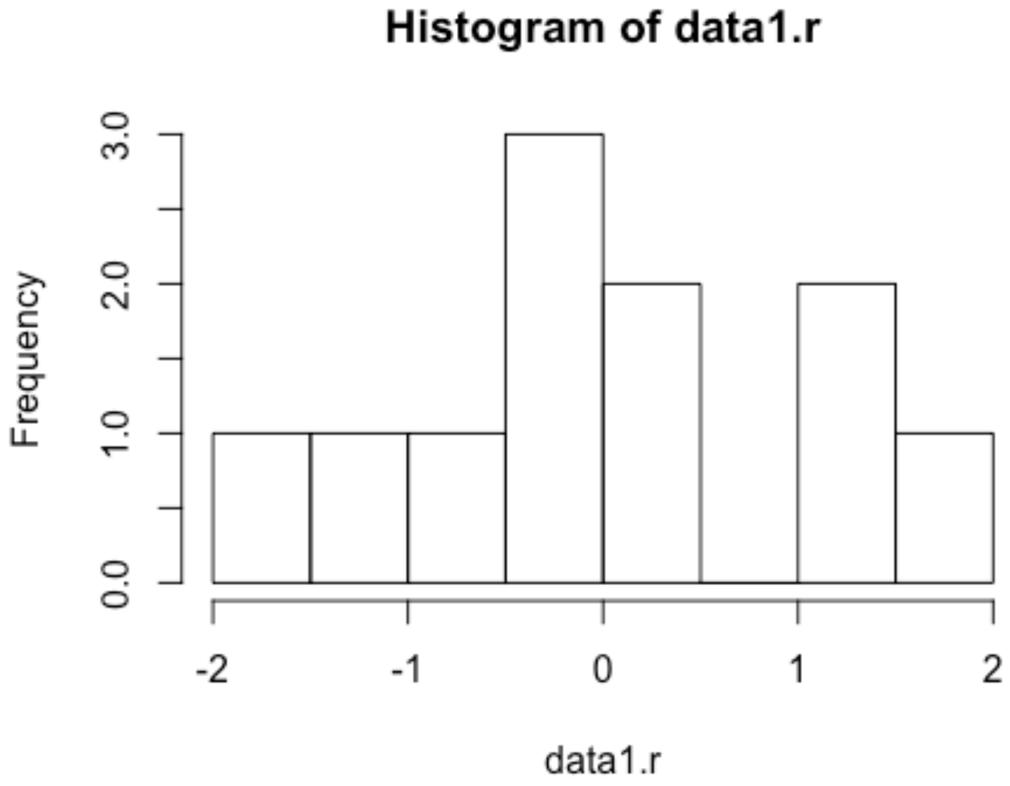
```
plot(data1$x ~ data1$y)
abline(equation.lm1)
```



```
data1.r <- rstandard(equation.lm1)
qqnorm(data1.r, ylab="y", xlab = "x")
qqline(data1.r)
```



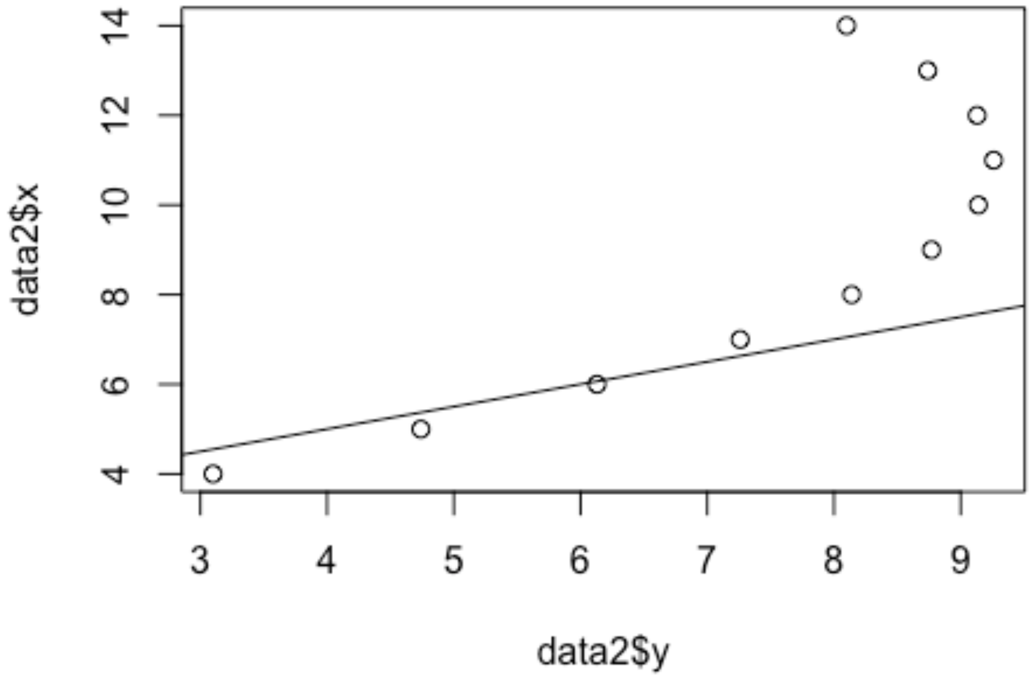
```
hist(data1.r)
```



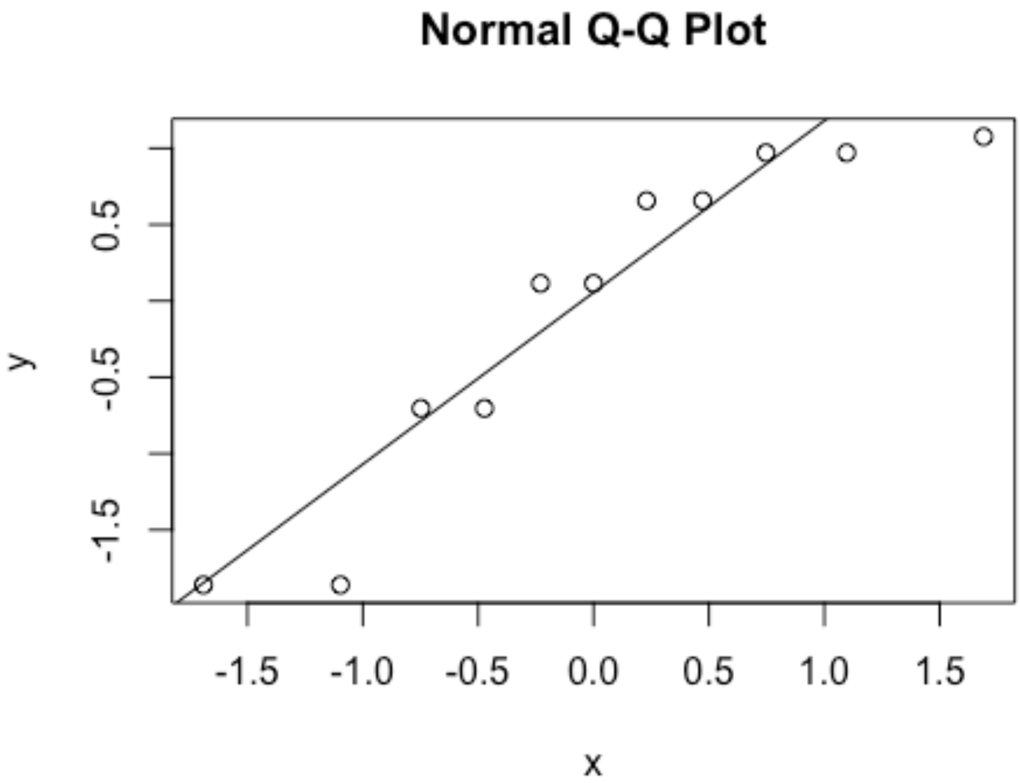
Data 1 appears that there is a positive linear relationship. Residual plots are very close to the line and the histogram for the residual appears to be an approximately normal shape. We also will assume independence of the observations. Data 1, it would be appropriate to estimate a regression line model.

for Data 2

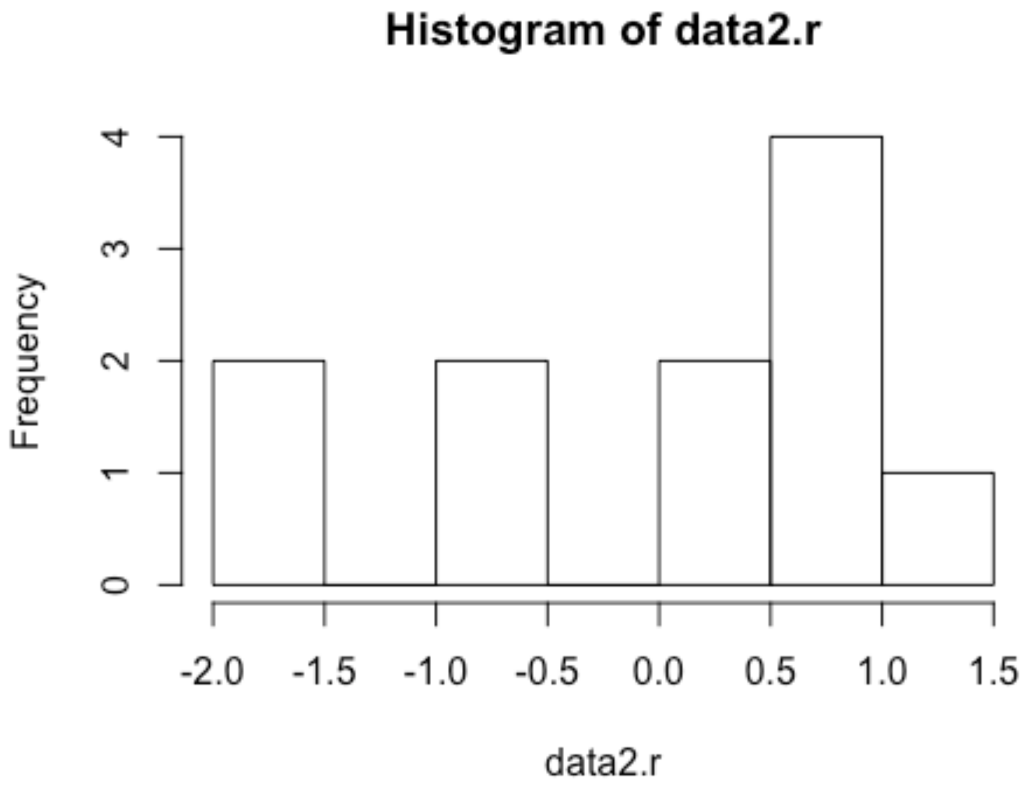
```
plot(data2$x ~ data2$y)
abline(equation.lm2)
```



```
data2.r <- rstandard(equation.lm2)
qqnorm(data2.r, ylab="y", xlab = "x")
qqline(data2.r)
```



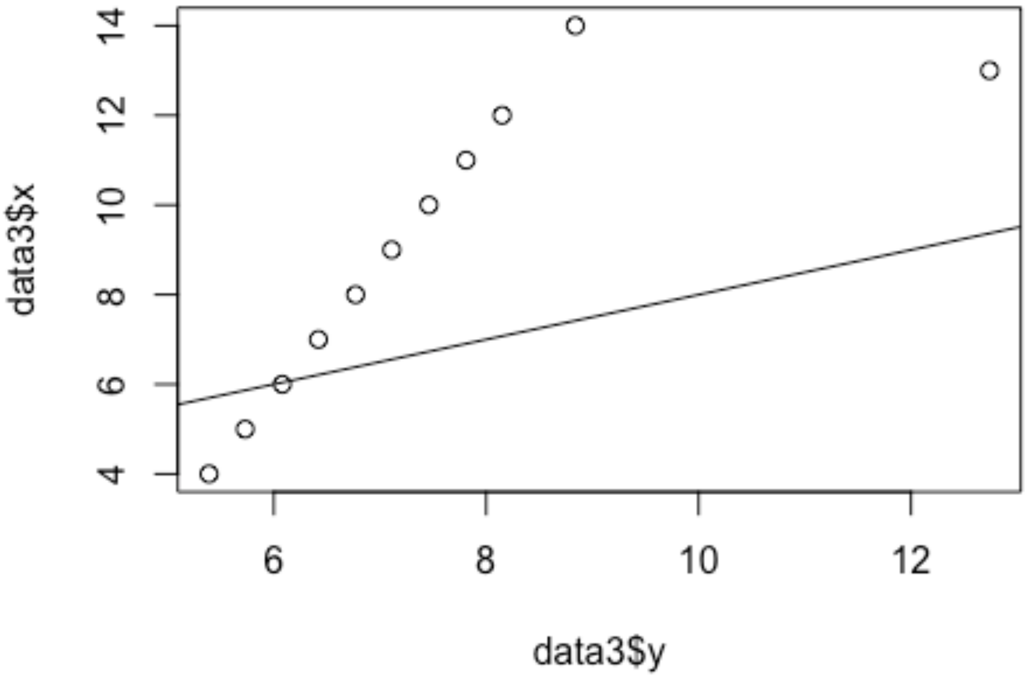
```
hist(data2.r)
```



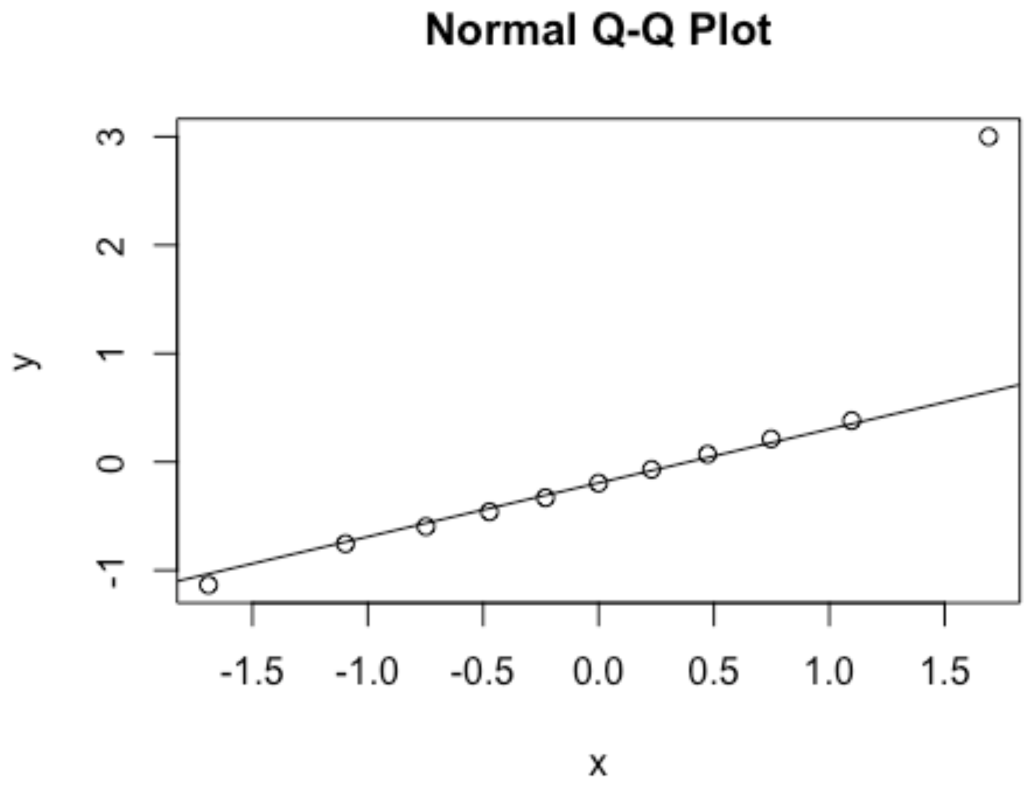
Data 2 appears to be a positive linear relationship. Residual plots are close to the line but the histogram for the residual appears to be skewed left shape which is not normal. We'll assume the independence of the observations. Data 2 may not be appropriate to estimate a regression line.

for Data 3

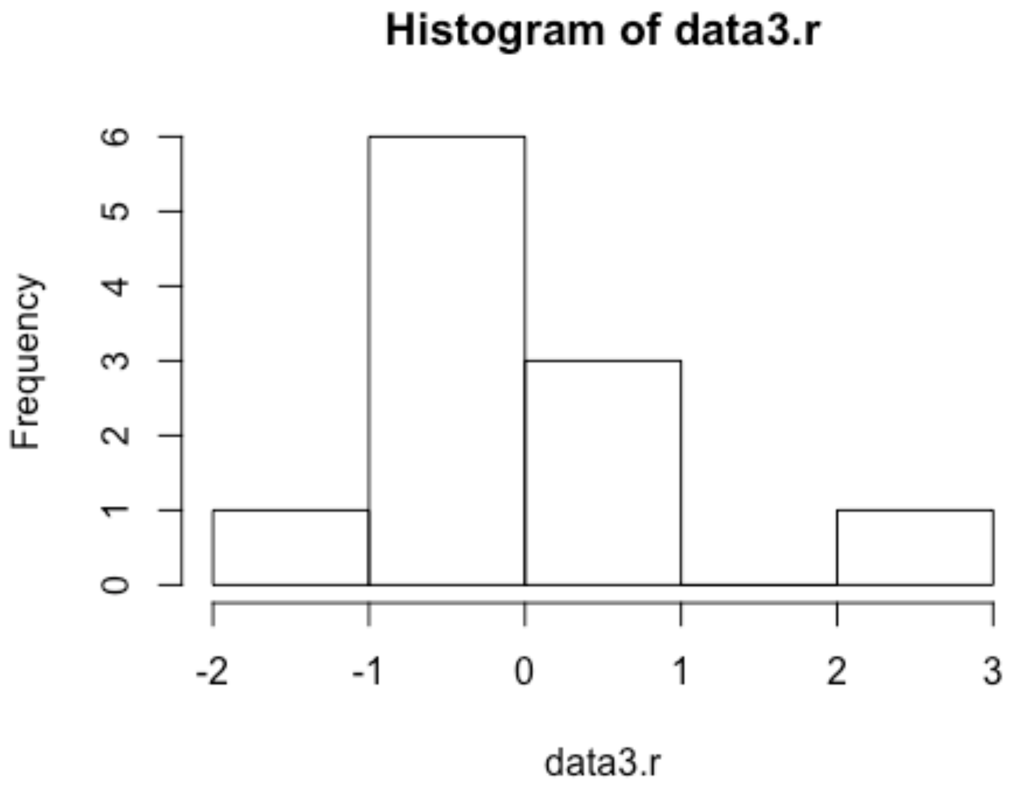
```
plot(data3$x ~ data3$y)
abline(equation.lm3)
```



```
data3.r <- rstandard(equation.lm3)
qqnorm(data3.r, ylab="y", xlab = "x")
qqline(data3.r)
```



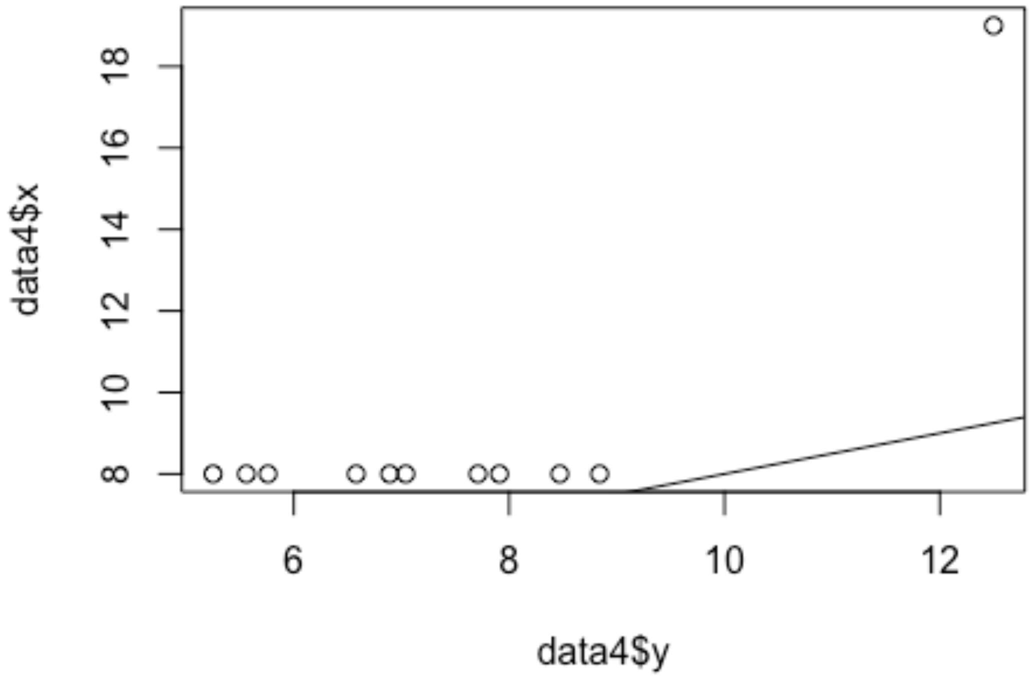
```
hist(data3.r)
```



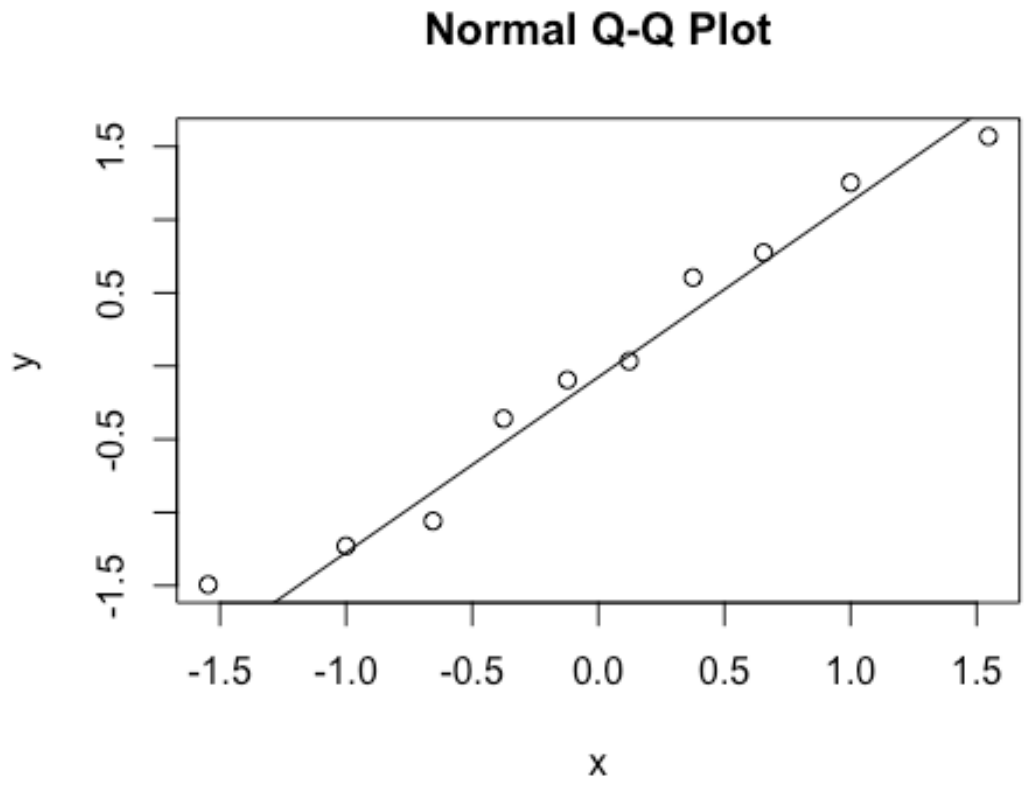
Data 3 appears that there is a positive linear relationship. Residual plots are very close to the line and the histogram for the residual appears to be an approximately normal shape. We also will assume independence of the observations. Data 3, it would be appropriate to estimate a regression line model.

for Data 4

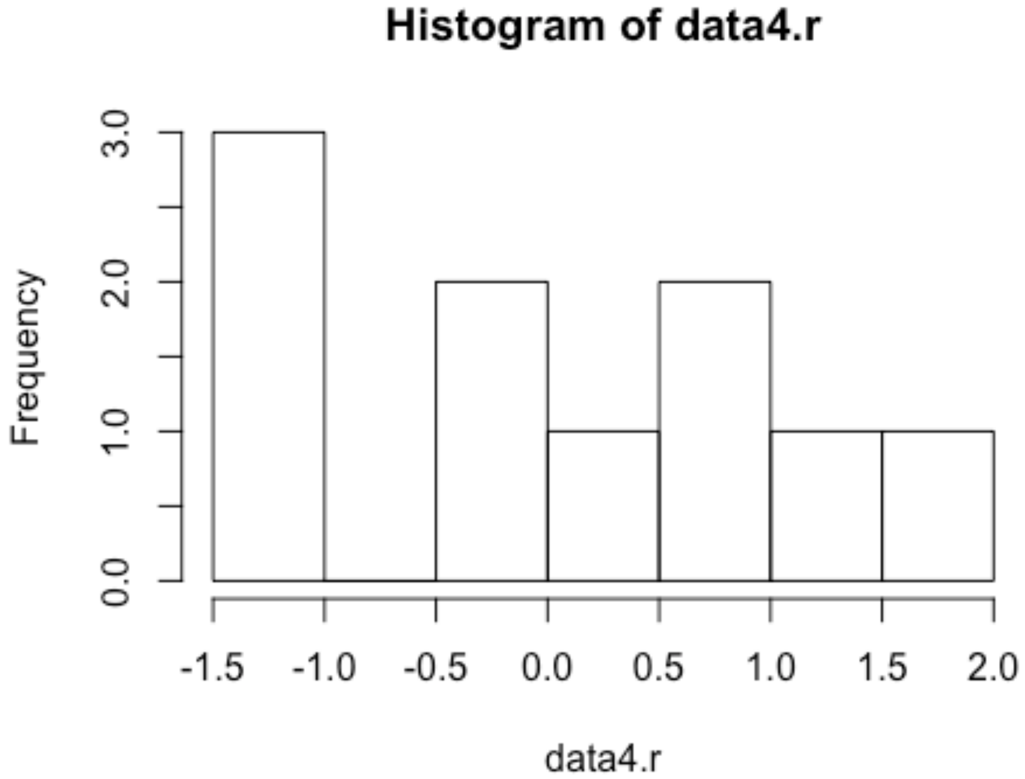
```
plot(data4$x ~ data4$y)
abline(equation.lm4)
```



```r
data4.r <- rstandard(equation.lm4)
qqnorm(data4.r, ylab="y", xlab = "x")
qqline(data4.r)
```



```r
hist(data4.r)
```



Data 4 is not appropriate to estimate a regression line model because its scattater diagram shows a horizontal line which means there is no correlation between x and y.

It is very important to include appropriate visualizations because by calculating the mean, standard deviation, r, and linear regression line, it is very difficult to tell if the data can be used to estimate a regression line. In our case, all the results appear to be the same but by using graph, it is clearly to see which data can be used.

Thank you.