Spring 2017 - Final Exam

CUNY MSDS DATA 606

Duubar Villalobos Jimenez mydvtech@gmail.com

May 25, 2017

Part I

1. Quantitative and Discrete variables

A student is gathering data on the driving experiences of other college students. A description of the data car color is presented below. Which of the variables are quantitative and discrete?

Car: 1 = compact, 2 = standard size, 3 = mini van, 4 = SUV, and 5 = truck **Color:** red, blue, green, black, white **daysDrive:** number of days per week the student drives **gasMonth:** the amount of money the student spends on gas per month

a. car
b. daysDrive
c. `daysDrive, car` **<- ANSWER**
d. daysDrive, gasMonth
e. car, daysDrive, gasMonth

Answer

- The Correct answer will be option **b** since `daysDrive` quantifies whole non negative numbers with jumps.
- Car is a numerical categorical variable since it's used for categories in different numerical levels.
- gasMonth is considered a numerical variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values.

2. GPA histogram

A histogram of the GPA of 132 students from this course in Fall 2012 class is presented below. Which estimates of the mean and median are most plausible?

a. `mean = 3.3, median = 3.5` **<- ANSWER**
b. mean = 3.5, median = 3.3
c. mean = 2.9, median = 3.8
d. mean = 3.8, median = 2.9
e. mean = 2.5, median = 3.8

Answer:

Since the distribution is left skewed, the mean is smaller than the median, leaving us with 3 choices a, c, e.

From the remaining options, if we look at the histogram, we can observe that the median could be about 3.5 since 3.8 is a little too high leaving us with option **a**. Hence the plausible option will be option **a**.

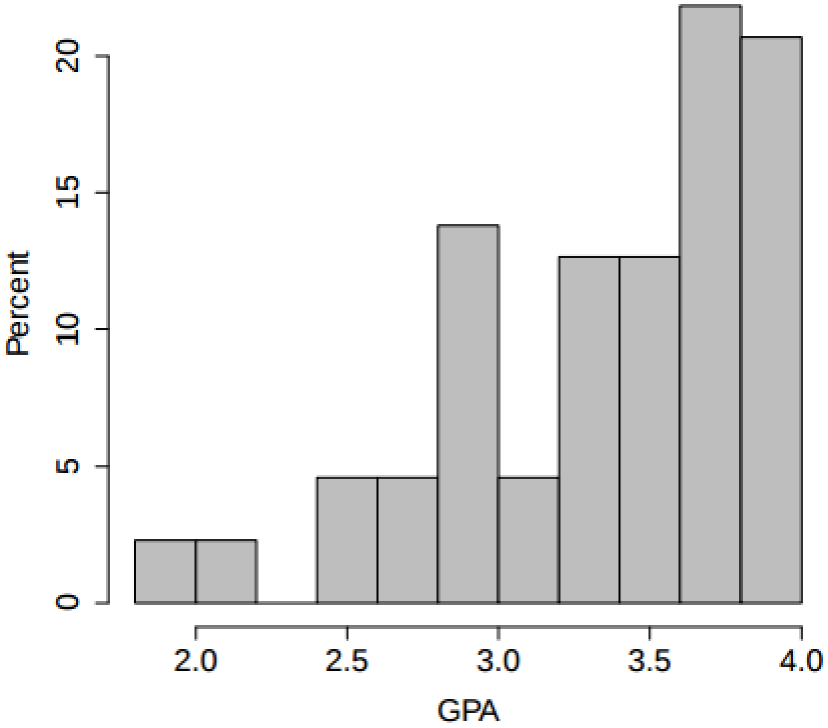


Figure 1:

3. Ebola fever

A researcher wants to determine if a new treatment is effective for reducing Ebola related fever. What type of study should be conducted in order to establish that the treatment does indeed cause improvement in Ebola patients?

a. Randomly assign Ebola patients to one of two groups, either the treatment or placebo group, and then compare the fever of the two groups.
b. Identify Ebola patients who received the new treatment and those who did not, and then compare the fever of those two groups.
c. Identify clusters of villages and then stratify them by gender and compare the fevers of male and female groups.
d. `Both studies (a) and (b)` can be conducted in order to establish that the treatment does indeed cause improvement with regards to fever in Ebola patients. **<- ANSWER**

Answer:

The answer will be option **d** since observational studies come in two forms: prospective and retrospective studies.

4. Natural Hair color relationship

A study is designed to test whether there is a relationship between natural hair color (brunette, blond, red) and eye color (blue, green, brown). If a large χ^2 test statistic is obtained, this suggests that:

a. `there is a difference` between average eye color and average hair color. **<- ANSWER**
b. a person's hair color is determined by his or her eye color.
c. there is an association between natural hair color and eye color.
d. eye color and natural hair color are independent.

Answer:

Because larger chi-square values correspond to stronger evidence against the null hypothesis and usually we take our null hypothesis as the average being the same; hence, we can conclude that the option **a** is the answer.

5. Standar Memory Task

A researcher studying how monkeys remember is interested in examining the distribution of the score on a standard memory task. The researcher wants to produce a boxplot to examine this distribution. Below are summary statistics from the memory task. What values should the researcher use to determine if a particular score is a potential outlier in the boxplot?

min	Q1	median	Q3	max	mean	sd	n
26	37	45	49.8	65	44.4	8.4	50

a. 37.0 and 49.8
b. `17.8 and 69.0` **<- ANSWER**
c. 36.0 and 52.8
d. 26.0 and 50.0

e. 19.2 and 69.9

Answer:

Since the $IQR = Q_3 - Q_1$

We have that the $IQR = 49.8 - 37$

Hence $IQR = 12.8$

And in order to identify the outliers, we can find them as any value below $Q_1 - 1.5 \cdot IQR$ for the lower value and $Q_3 + 1.5 \cdot IQR$ any value above for the upper value.

Hence:

The lower outlier limit will be $37 - 1.5 \cdot 12.8 = 17.8$

The upper outlier limit will be $49.8 + 1.5 \cdot 12.8 = 69$

Based on those results, any value lower than 17.8 or upper than 69 will be considered outliers. Giving us the option **b**.

6. Complete the Sentence

The ____ are resistant to outliers, whereas the ____ are not.

a. mean and median; standard deviation and interquartile range
b. mean and standard deviation; median and interquartile range
c. standard deviation and interquartile range; mean and median
d. `median and interquartile range; mean and standard deviation` **<- ANSWER**
e. median and standard deviation; mean and interquartile range

Answer:

Option **d**.

7. Sampling Distribution

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

Answer:

a. Describe the two distributions (2 pts).

- Both distributions are uni modal.
- The observations distribution is skewed to the right.
- The sampling distributions seems to be normal due to it's 500 random mean sampling of size 30.
- If it is true that their means are not equal; these are near and could be improved if the samples size increase from 30.

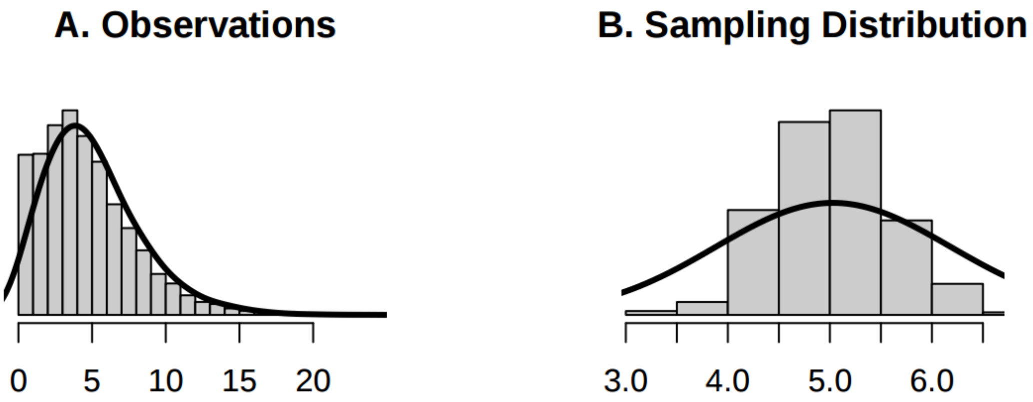


Figure 2:

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

The reason for the means to be similar is because when the mean from 500 random samples of size 30 from the Observations distribution was performed, it followed a normal distribution since it satisfies the randomness and minimum number of samples in order to follow a normal distribution. Also,the law of large numbers (LLN) describes the result of performing the same experiment a large number of times. According to the law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed. In regards of the standard deviation is because the Observations distribution has observations farther from the mean than in the Sampling Distribution as previously explained by the law of large numbers.

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

The statistical principal is the normality condition provided by the Central Limit Theorem. This ensures that the distribution of sample means will be nearly normal, regardless of sample size, when the data come from a nearly normal distribution. Also, explained in different words: the central limit theorem (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. In the above example, in order to obtain the mean, the independent selected random values were added in order to calculate the mean, satisfying the condition.

Part II

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

```r
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).

```
mxdata1 <- mean(data1$x)
mydata1 <- mean(data1$y)

mxdata2 <- mean(data2$x)
mydata2 <- mean(data2$y)

mxdata3 <- mean(data3$x)
mydata3 <- mean(data3$y)

mxdata4 <- mean(data4$x)
mydata4 <- mean(data4$y)

summary(data1)
```

```
##        x              y         
##  Min.   : 4.0   Min.   : 4.3  
##  1st Qu.: 6.5   1st Qu.: 6.3  
##  Median : 9.0   Median : 7.6  
##  Mean   : 9.0   Mean   : 7.5  
##  3rd Qu.:11.5   3rd Qu.: 8.6  
##  Max.   :14.0   Max.   :10.8  
```

```
summary(data2)
```

```
##        x              y        
##  Min.   : 4.0   Min.   :3.1  
##  1st Qu.: 6.5   1st Qu.:6.7  
##  Median : 9.0   Median :8.1  
##  Mean   : 9.0   Mean   :7.5  
##  3rd Qu.:11.5   3rd Qu.:8.9  
##  Max.   :14.0   Max.   :9.3  
```

```
summary(data3)
```

```
##        x              y         
##  Min.   : 4.0   Min.   : 5.4  
##  1st Qu.: 6.5   1st Qu.: 6.2  
##  Median : 9.0   Median : 7.1  
##  Mean   : 9.0   Mean   : 7.5  
##  3rd Qu.:11.5   3rd Qu.: 8.0  
##  Max.   :14.0   Max.   :12.7  
```

```
summary(data4)
```

```
##        x            y         
##  Min.   : 8   Min.   : 5.2  
##  1st Qu.: 8   1st Qu.: 6.2  
##  Median : 8   Median : 7.0  
##  Mean   : 9   Mean   : 7.5  
##  3rd Qu.: 8   3rd Qu.: 8.2  
##  Max.   :19   Max.   :12.5  
```

Data	Mean x	Mean y
data1	9	7.5
data2	9	7.5
data3	9	7.5
data4	9	7.5

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

```
mdxdata1 <- median(data1$x)
mdydata1 <- median(data1$y)

mdxdata2 <- median(data2$x)
mdydata2 <- median(data2$y)

mdxdata3 <- median(data3$x)
mdydata3 <- median(data3$y)

mdxdata4 <- median(data4$x)
mdydata4 <- median(data4$y)

summary(data1)
```

```
##        x              y
##  Min.   : 4.0   Min.   : 4.3
##  1st Qu.: 6.5   1st Qu.: 6.3
##  Median : 9.0   Median : 7.6
##  Mean   : 9.0   Mean   : 7.5
##  3rd Qu.:11.5   3rd Qu.: 8.6
##  Max.   :14.0   Max.   :10.8
```

```
summary(data2)
```

```
##        x              y
##  Min.   : 4.0   Min.   :3.1
##  1st Qu.: 6.5   1st Qu.:6.7
##  Median : 9.0   Median :8.1
##  Mean   : 9.0   Mean   :7.5
##  3rd Qu.:11.5   3rd Qu.:8.9
##  Max.   :14.0   Max.   :9.3
```

```
summary(data3)
```

```
##        x              y
##  Min.   : 4.0   Min.   : 5.4
##  1st Qu.: 6.5   1st Qu.: 6.2
##  Median : 9.0   Median : 7.1
##  Mean   : 9.0   Mean   : 7.5
##  3rd Qu.:11.5   3rd Qu.: 8.0
##  Max.   :14.0   Max.   :12.7
```

```
summary(data4)
```

```
##        x            y
##  Min.   : 8   Min.   : 5.2
##  1st Qu.: 8   1st Qu.: 6.2
##  Median : 8   Median : 7.0
##  Mean   : 9   Mean   : 7.5
##  3rd Qu.: 8   3rd Qu.: 8.2
##  Max.   :19   Max.   :12.5
```

Data	Median x	Median y
data1	9	7.58
data2	9	8.14
data3	9	7.11
data4	8	7.04

c. The standard deviation (for x and y separately; 1 pt).

```r
sdxdata1 <- sd(data1$x)
sdydata1 <- sd(data1$y)

sdxdata2 <- sd(data2$x)
sdydata2 <- sd(data2$y)

sdxdata3 <- sd(data3$x)
sdydata3 <- sd(data3$y)

sdxdata4 <- sd(data4$x)
sdydata4 <- sd(data4$y)

summary(data1)
```

```
##        x              y        
##  Min.   : 4.0   Min.   : 4.3  
##  1st Qu.: 6.5   1st Qu.: 6.3  
##  Median : 9.0   Median : 7.6  
##  Mean   : 9.0   Mean   : 7.5  
##  3rd Qu.:11.5   3rd Qu.: 8.6  
##  Max.   :14.0   Max.   :10.8
```

```r
summary(data2)
```

```
##        x              y      
##  Min.   : 4.0   Min.   :3.1  
##  1st Qu.: 6.5   1st Qu.:6.7  
##  Median : 9.0   Median :8.1  
##  Mean   : 9.0   Mean   :7.5  
##  3rd Qu.:11.5   3rd Qu.:8.9  
##  Max.   :14.0   Max.   :9.3
```

```r
summary(data3)
```

```
##        x              y        
##  Min.   : 4.0   Min.   : 5.4  
##  1st Qu.: 6.5   1st Qu.: 6.2  
##  Median : 9.0   Median : 7.1  
##  Mean   : 9.0   Mean   : 7.5  
##  3rd Qu.:11.5   3rd Qu.: 8.0  
##  Max.   :14.0   Max.   :12.7
```

```
summary(data4)
```

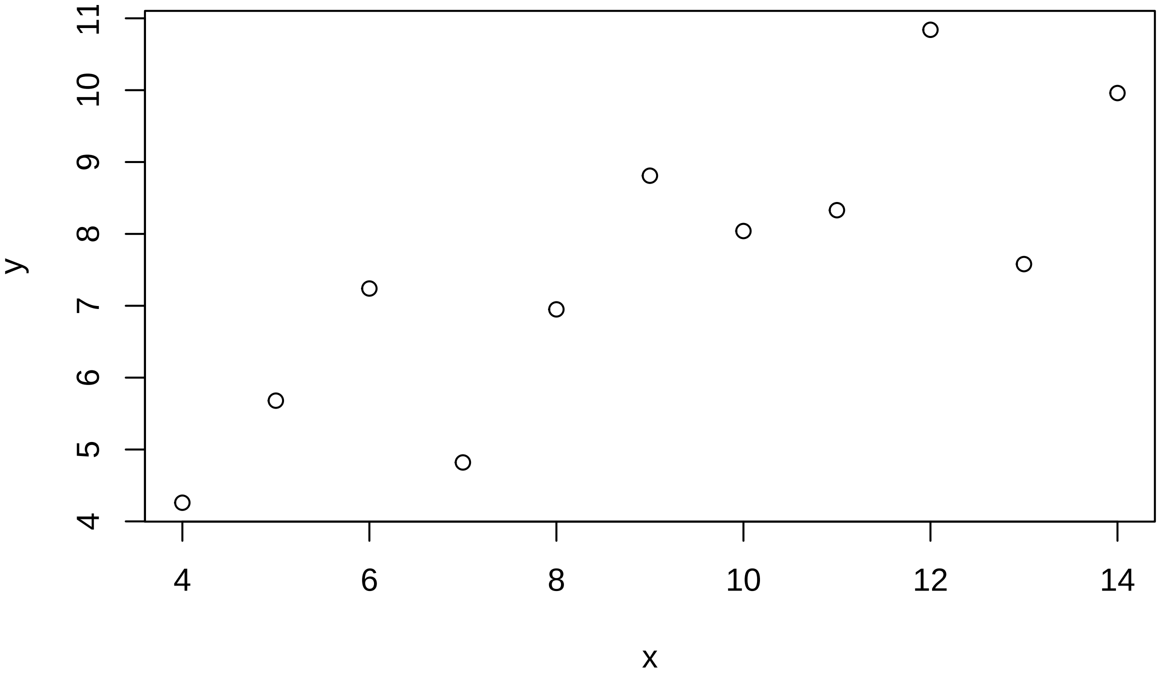
```
##        x              y        
##  Min.   : 8   Min.   : 5.2  
##  1st Qu.: 8   1st Qu.: 6.2  
##  Median : 8   Median : 7.0  
##  Mean   : 9   Mean   : 7.5  
##  3rd Qu.: 8   3rd Qu.: 8.2  
##  Max.   :19   Max.   :12.5  
```

Data	SD x	SD y
data1	3.32	2.03
data2	3.32	2.03
data3	3.32	2.03
data4	3.32	2.03

For each x and y pair, calculate (also to two decimal places; 1 pt):

d. The correlation (1 pt).

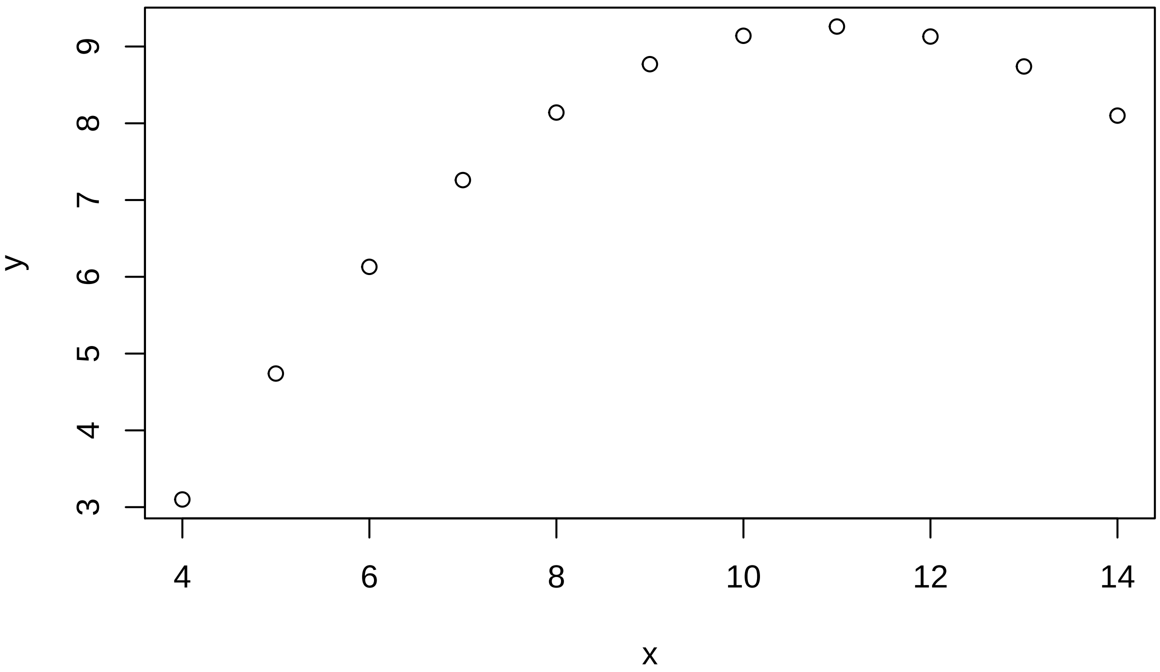
```
plot(data1)
```



```
cor(data1)
```

```
##      x    y
## x 1.00 0.82
## y 0.82 1.00
```

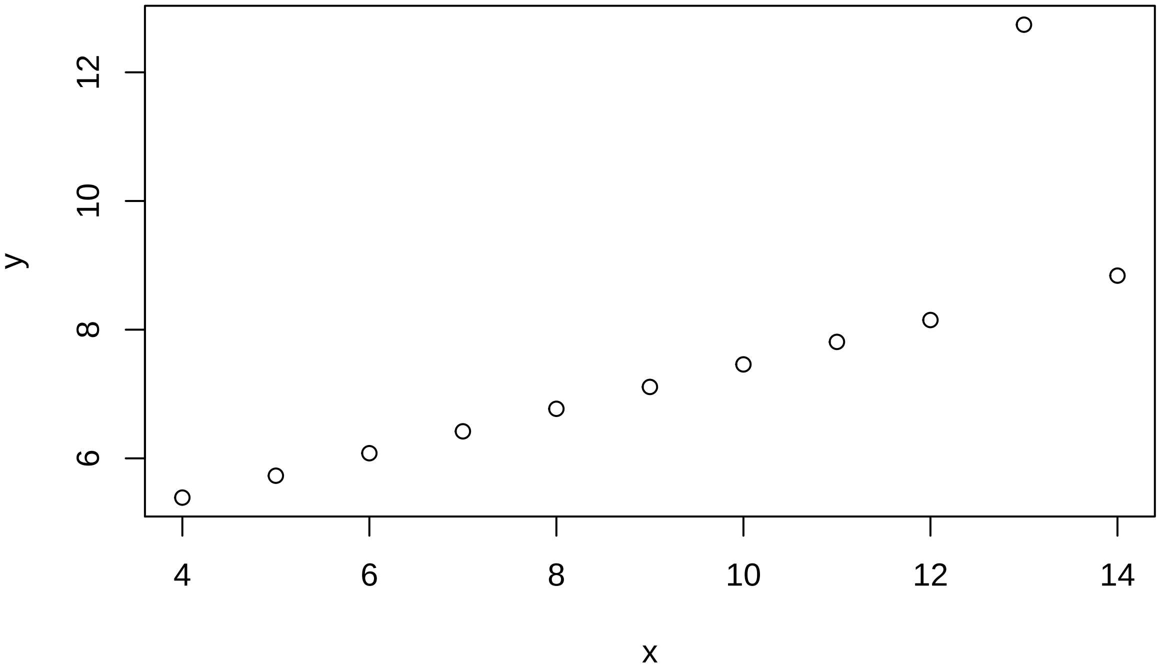
```r
plot(data2)
```



```r
cor(data2)
```

```
##      x    y
## x 1.00 0.82
## y 0.82 1.00
```

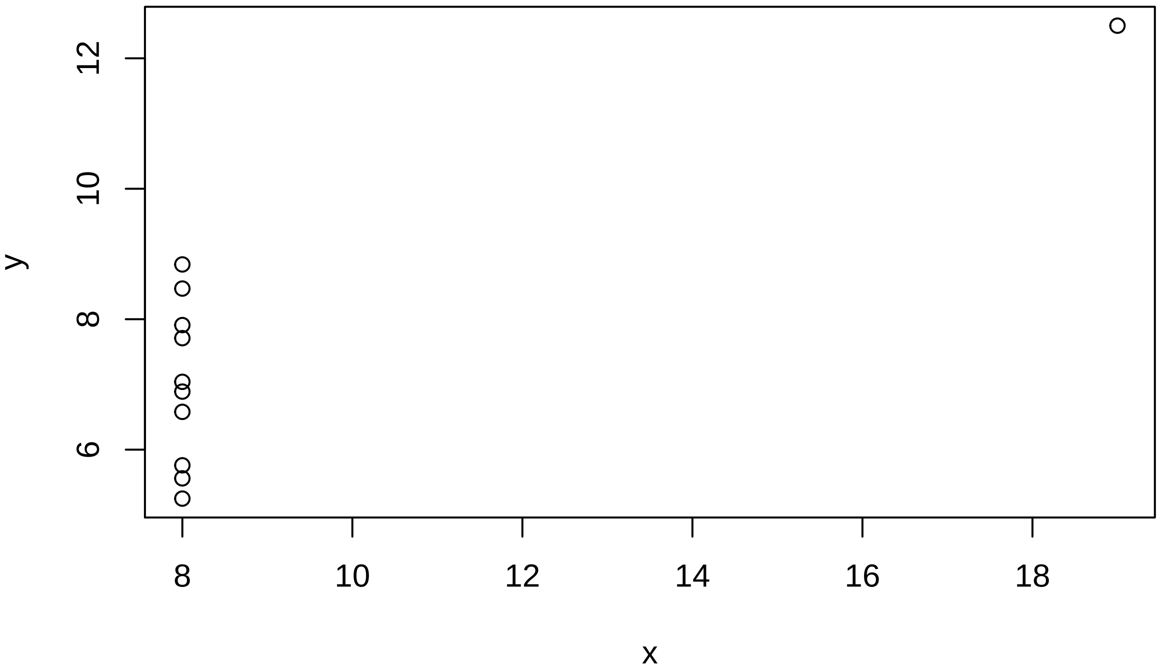
```r
plot(data3)
```



```r
cor(data3)
```

```
##      x    y
## x 1.00 0.82
## y 0.82 1.00
```

```
plot(data4)
```



```
cor(data4)
```

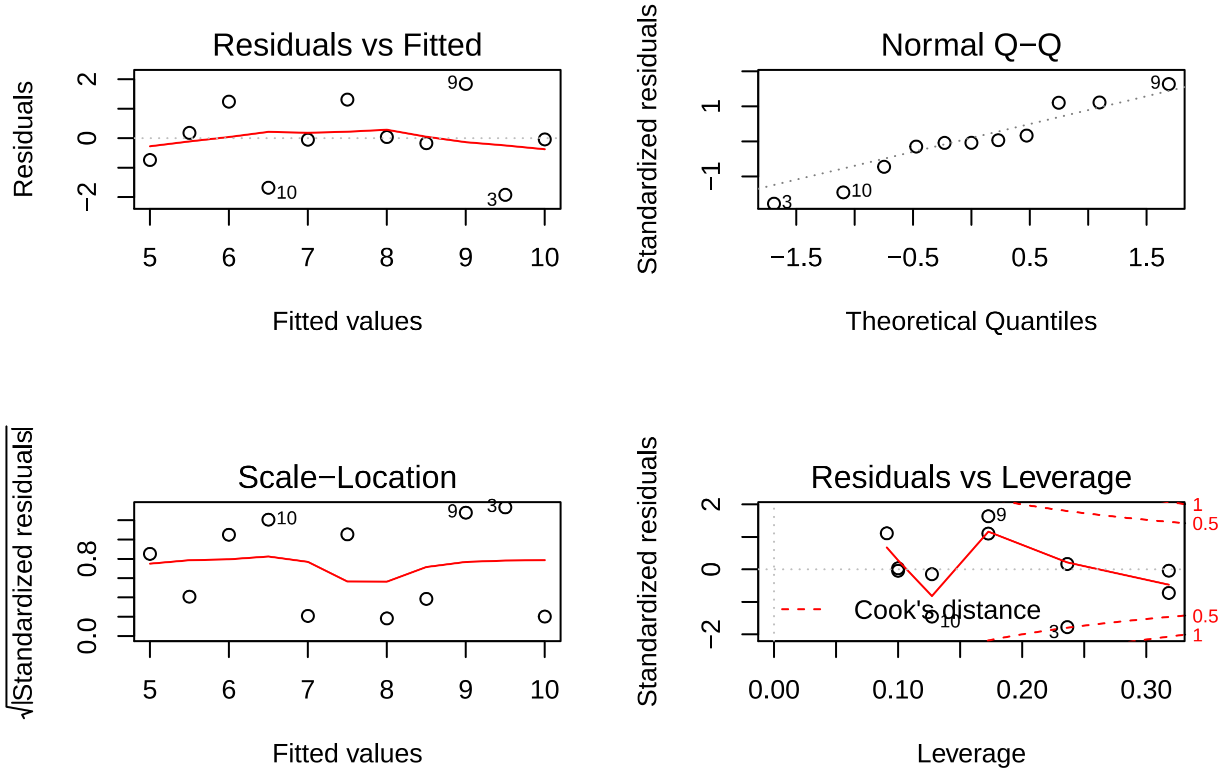
```
##      x    y
## x 1.00 0.82
## y 0.82 1.00
```

e. Linear regression equation (2 pts).

```
m1<-lm(y~x,data=data1)
summary(m1)
```

```
##
## 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
```

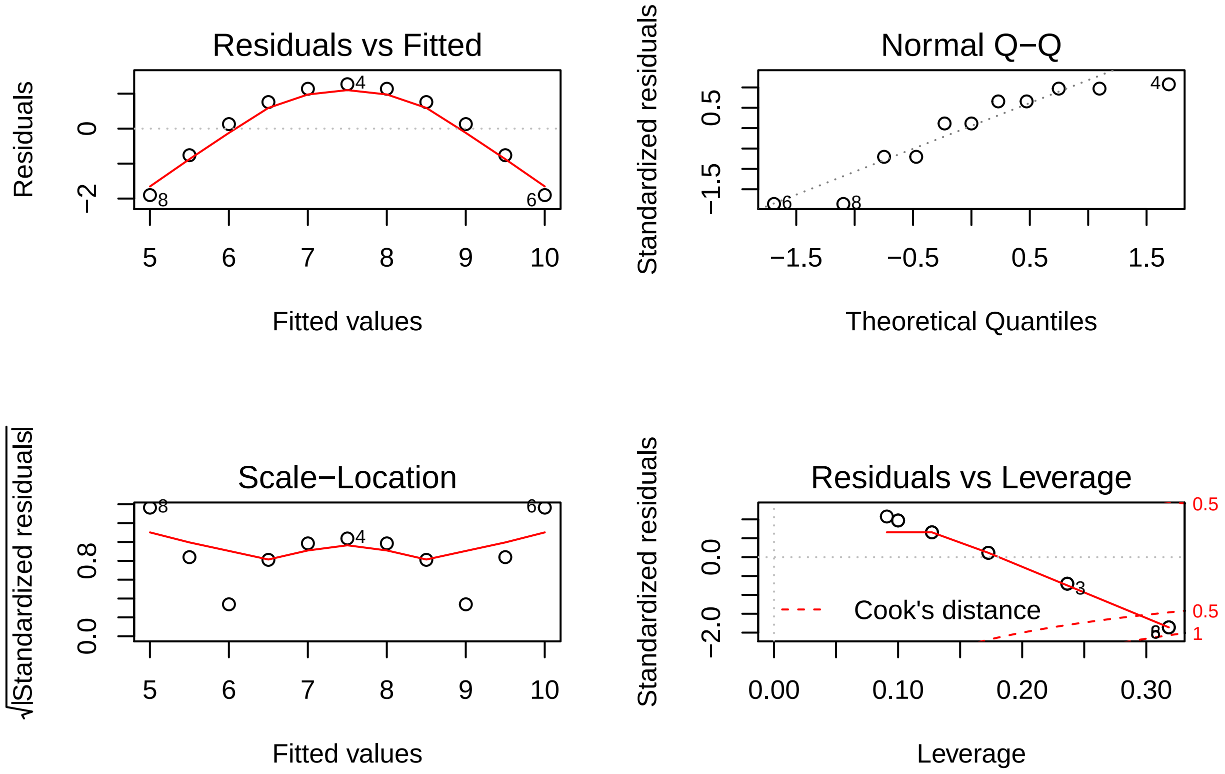
```
par(mfrow=c(2,2))
plot(m1)
```



```
m2<-lm(y~x,data=data2)
summary(m2)
```

```
##
## 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
```

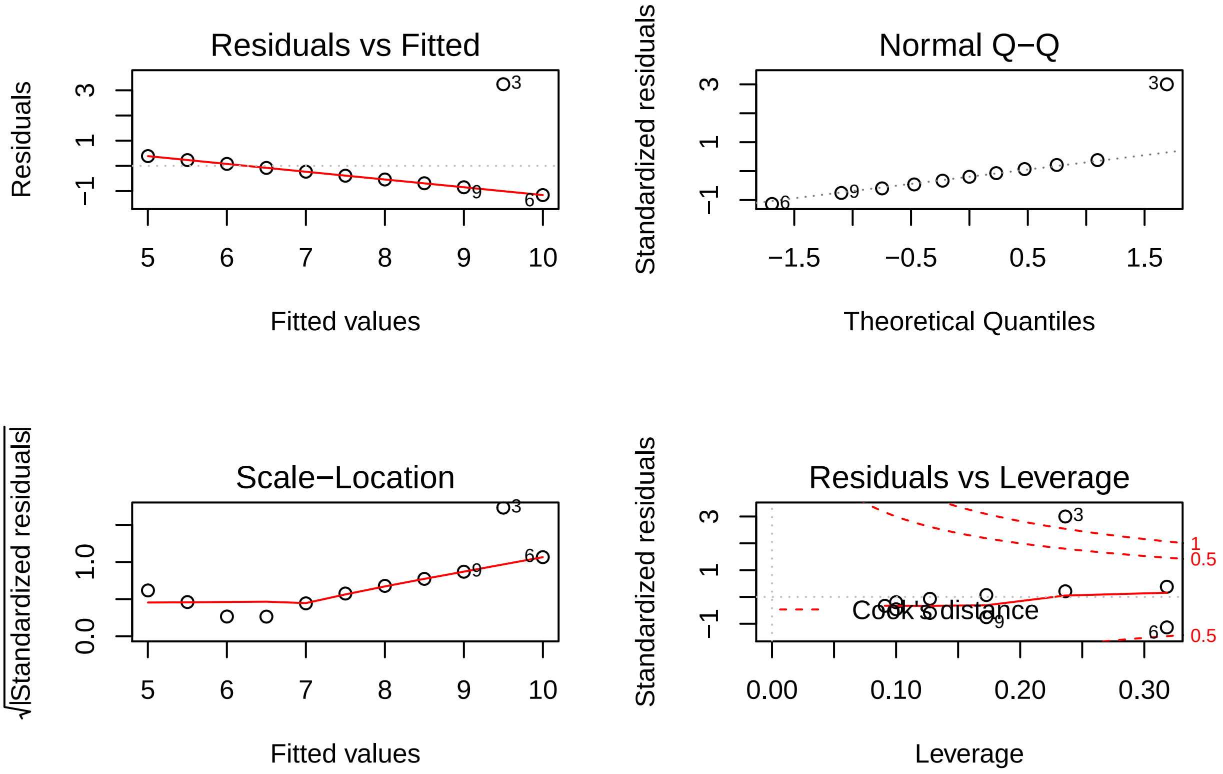
```
par(mfrow=c(2,2))
plot(m2)
```



```
m3<-lm(y~x,data=data3)
summary(m3)
```

```
##
## 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
```

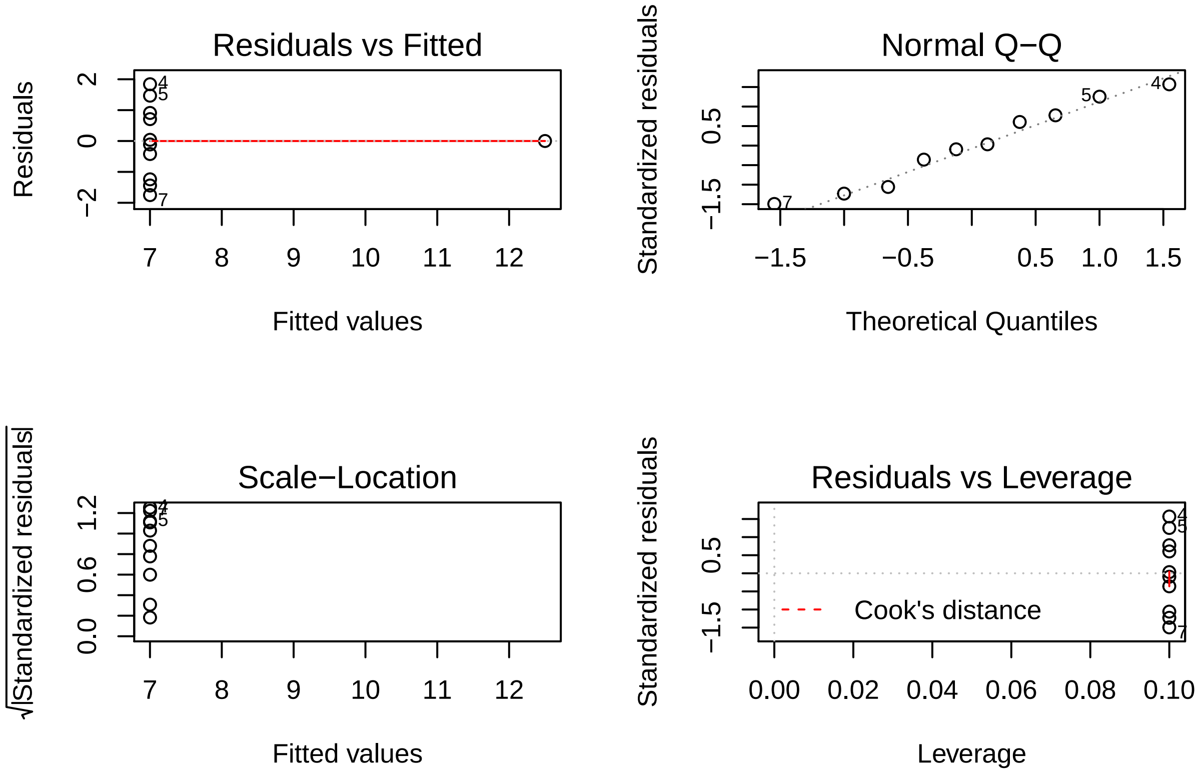
```
par(mfrow=c(2,2))
plot(m3)
```



```
m4<-lm(y~x,data=data4)
summary(m4)
```

```
##
## 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
```

```
par(mfrow=c(2,2))
plot(m4)
```



Answer:

$\hat{y}_1 = 3 + 0.5 \cdot x$

$\hat{y}_2 = 3 + 0.5 \cdot x$

$\hat{y}_3 = 3 + 0.5 \cdot x$

$\hat{y}_4 = 3 + 0.5 \cdot x$

f. R-Squared (2 pts).

```
summary(m1)$r.squared
```

```
## [1] 0.67
```

```
summary(m2)$r.squared
```

```
## [1] 0.67
```

```
summary(m3)$r.squared
```

```
## [1] 0.67
```

```
summary(m4)$r.squared
```

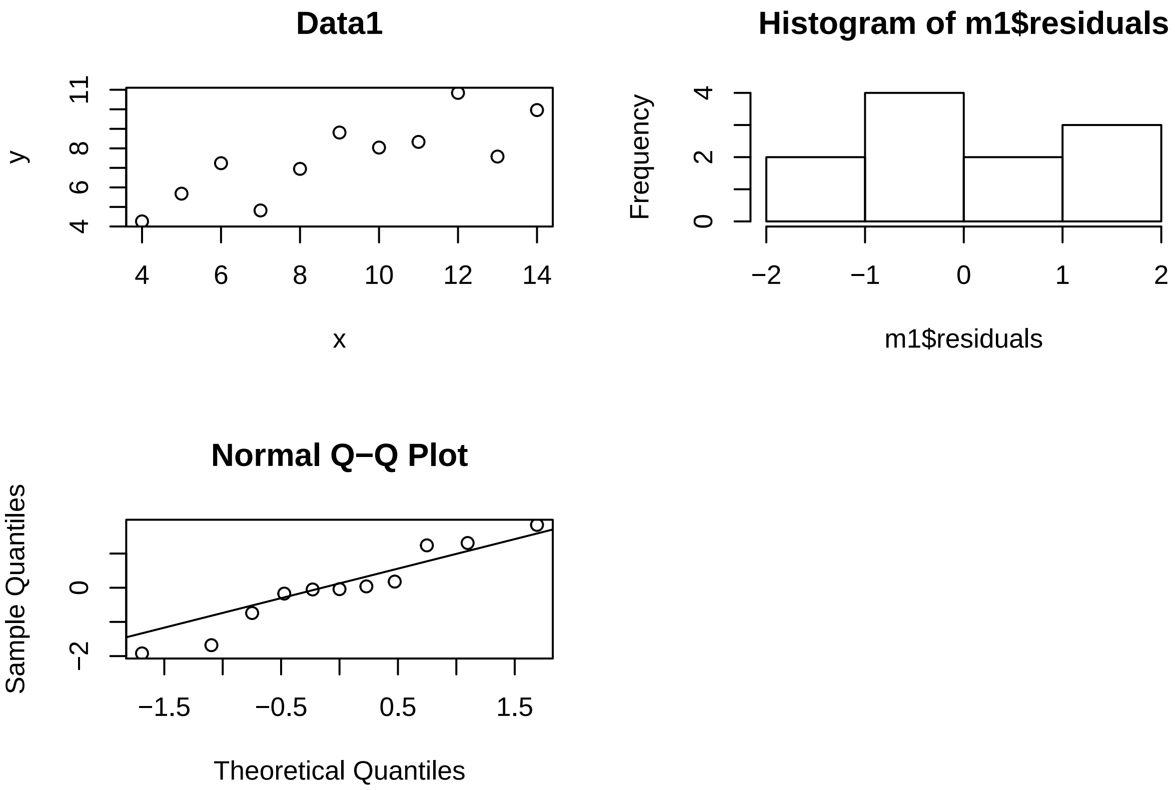
```
## [1] 0.67
```

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)

Data1

```
par(mfrow=c(2,2))
plot(data1, main = "Data1")
hist(m1$residuals)
```

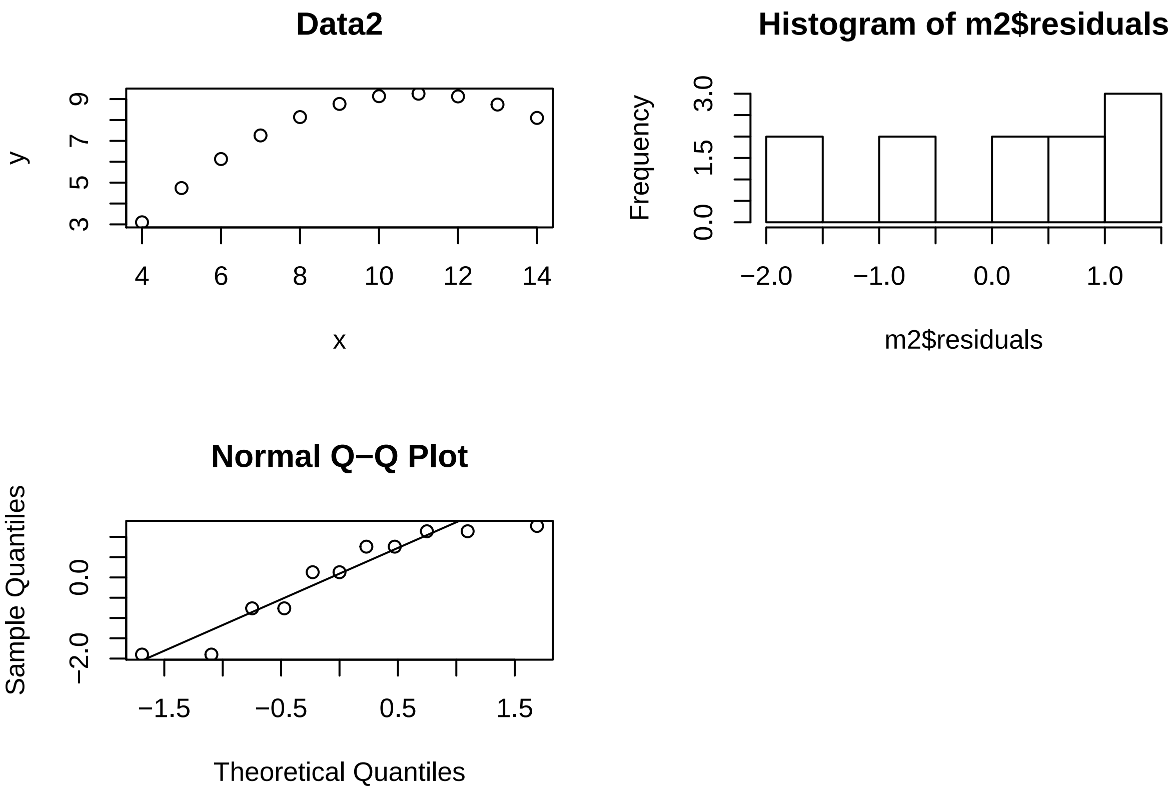
```
qqnorm(m1$residuals)
qqline(m1$residuals)
```



From the above information, we can identify that in the case of data1, the residuals do not seems to follow a nearly normal distribution even though the main data plot might suggest a linearity.

Data2

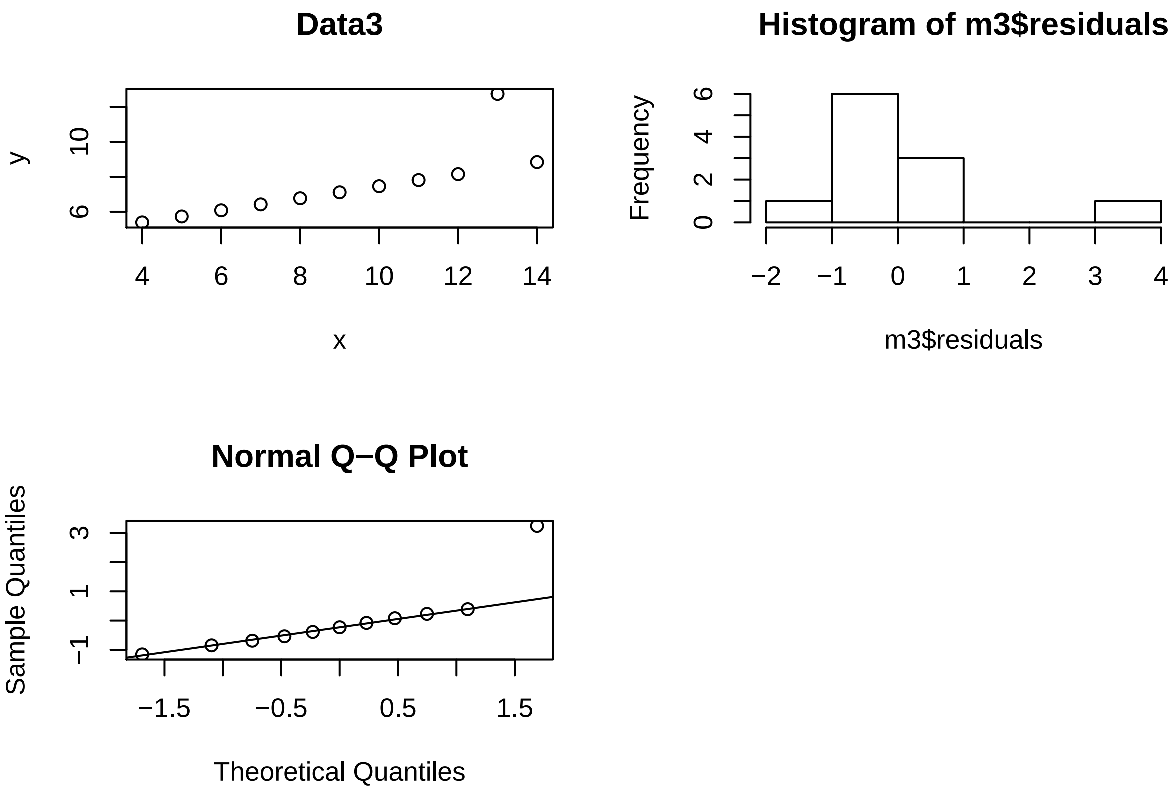
```
par(mfrow=c(2,2))
plot(data2, main = "Data2")
hist(m2$residuals)
qqnorm(m2$residuals)
qqline(m2$residuals)
```



From the above graphs, we can identify that the given data follows a curve and not necessarily a linear model; also the residuals do not seems to follow a normal distribution as well.

Data3

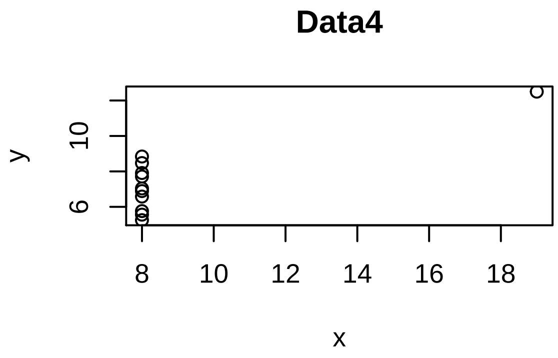
```
par(mfrow=c(2,2))
plot(data3, main = "Data3")
hist(m3$residuals)
qqnorm(m3$residuals)
qqline(m3$residuals)
```

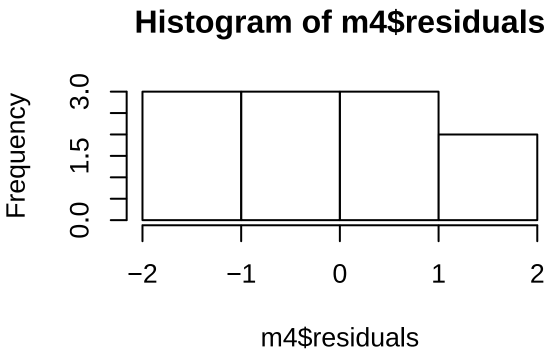


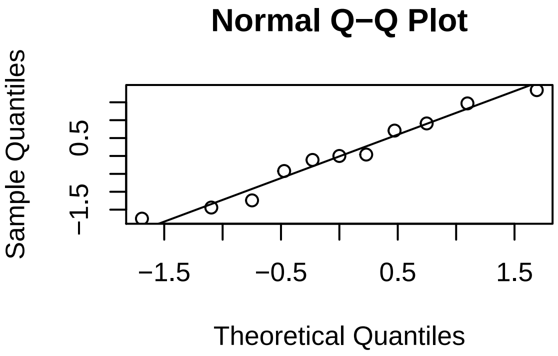
In the case of data3, it seems that the given data follows some sort of linearity with some outliers producing leverage and the distribution seems to be normal but due to leverage it affects the outcome.

Data4

```r
par(mfrow=c(2,2))
plot(data4, main = "Data4")
hist(m4$residuals)
qqnorm(m4$residuals)
qqline(m4$residuals)
```







In this case there's an outlier point producing leverage; and also the residuals distribution does not seems to be normal.

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

Appropriate visualizations are important since these help identify outliers and trends; it also helps to visualize the linear model of the plot and whether the residuals have constant variability from the regression line.

Links

In order to open, right click and select **"Open Link in New Tab"**.

dvillalobos.github.io

GitHub | Linkedin