DATA 606 Spring 2018 - Final Exam

Brian Weinfeld

Part I

Please put the answers for Part I next to the question number (2pts each):

1. B
2. B
3. A
4. C
5. B
6. D

7a. Describe the two distributions (2pts).

Center: Both distributions have a similar mean.

Shape: Distribution A is skew-right while distribution B is roughly normal.

Spread: A has a much larger spread, with a range of ≈ 18, while B has a much smaller spread, with a range of ≈ 3.5

7b. Explain why the means of these two distributions are similar but the standard deviations are not (2 pts).

The means are the same because the sampling distribution is used to approximate the mean of distribution A. As more samples are added to the sampling distribution it will begin to converge on the true mean of 5.05. This also explains why the sampling distribution has a smaller standard deviation. That is, the distribution is being used to approximate the mean of distribution A, not any of the other points. As more samples are added to B, the standard deviation will continue to shrink.

7c. What is the statistical principal that describes this phenomenon (2 pts)?

The statistical principal in effect with the sampling distribution is the **Central Limit Theorem**.

Part II

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

```
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 1

```r
mean(data1$x) %>% round(2)
```

```
## [1] 9
```

```r
mean(data1$y) %>% round(2)
```

```
## [1] 7.5
```

Data 2

```r
mean(data2$x) %>% round(2)
```

```
## [1] 9
```

```r
mean(data2$y) %>% round(2)
```

```
## [1] 7.5
```

Data 3

```r
mean(data3$x) %>% round(2)
```

```
## [1] 9
```

```r
mean(data3$y) %>% round(2)
```

```
## [1] 7.5
```

Data 4

```r
mean(data4$x) %>% round(2)
```

```
## [1] 9
```

```r
mean(data4$y) %>% round(2)
```

```
## [1] 7.5
```

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

Data 1

```r
median(data1$x) %>% round(2)
```

```
## [1] 9
```

```r
median(data1$y) %>% round(2)
```

```
## [1] 7.58
```

Data 2

```r
median(data2$x) %>% round(2)
```

```
## [1] 9
```

```r
median(data2$y) %>% round(2)
```

```
## [1] 8.14
```

Data 3

```r
median(data3$x) %>% round(2)
```

```
## [1] 9
```

```r
median(data3$y) %>% round(2)
```

```
## [1] 7.11
```

Data 4

```r
median(data4$x) %>% round(2)
```

```
## [1] 8
```

```r
median(data4$y) %>% round(2)
```

```
## [1] 7.04
```

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

Data 1

```r
sd(data1$x) %>% round(2)
```

```
## [1] 3.32
```

```r
sd(data1$y) %>% round(2)
```

```
## [1] 2.03
```

Data 2

```r
sd(data2$x) %>% round(2)
```

```
## [1] 3.32
```

```r
sd(data2$y) %>% round(2)
```

```
## [1] 2.03
```

Data 3

```r
sd(data3$x) %>% round(2)
```

```
## [1] 3.32
```

```r
sd(data3$y) %>% round(2)
```

```
## [1] 2.03
```

Data 4

```r
sd(data4$x) %>% round(2)
```

```
## [1] 3.32
```

```r
sd(data4$y) %>% round(2)
```

```
## [1] 2.03
```

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

d. The correlation (1 pt).

Data 1

```r
cor(data1$x, data1$y) %>% round(2)
```

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

Data 2

```r
cor(data2$x, data2$y) %>% round(2)
```

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

Data 3

```r
cor(data3$x, data3$y) %>% round(2)
```

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

Data 4

```r
cor(data4$x, data4$y) %>% round(2)
```

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

e. Linear regression equation (2 pts).

Data 1

```r
d1lm <- lm(y~x, data1)
summary(d1lm)
```

```
## 
## Call:
## lm(formula = y ~ x, data = data1)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.92127 -0.45577 -0.04136  0.70941  1.83882 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)   3.0001     1.1247   2.667  0.02573 * 
## x             0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295 
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

$$\widehat{y} = 0.5001 \times x + 3.0001$$

Data 2

```r
d2lm <- lm(y~x, data2)
summary(d2lm)
```

```
## 
## Call:
## lm(formula = y ~ x, data = data2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9009 -0.7609  0.1291  0.9491  1.2691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125   2.667  0.02576 *
## x              0.500      0.118   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

$$\widehat{y} = 0.5001 \times x + 3.0001$$

Data 3

```r
d3lm <- lm(y~x, data3)
summary(d3lm)
```

```
##
## Call:
## lm(formula = y ~ x, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1586 -0.6146 -0.2303  0.1540  3.2411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0025     1.1245   2.670  0.02562 *
## x             0.4997     0.1179   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

$$\widehat{y} = 0.4997 \times x + 3.0025$$

Data 4

```r
d4lm <- lm(y~x, data4)
summary(d4lm)
```

```
##
## 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)   3.0017     1.1239   2.671  0.02559 *
## x             0.4999     0.1178   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

$$\widehat{y} = 0.4999 \times x + 3.0017$$

f. R-Squared (2 pts).

Data 1 (Calculated in part e)

```
print('Adjusted R-squared:  0.6295')
```

```
## [1] "Adjusted R-squared:  0.6295"
```

Data 2 (Calculated in part e)

```
print('Adjusted R-squared:  0.6292')
```

```
## [1] "Adjusted R-squared:  0.6292"
```

Data 3 (Calculated in part e)

```
print('Adjusted R-squared:  0.6292')
```

```
## [1] "Adjusted R-squared:  0.6292"
```

Data 4 (Calculated in part e)

```
print('Adjusted R-squared:  0.6297')
```

```
## [1] "Adjusted R-squared:  0.6297"
```

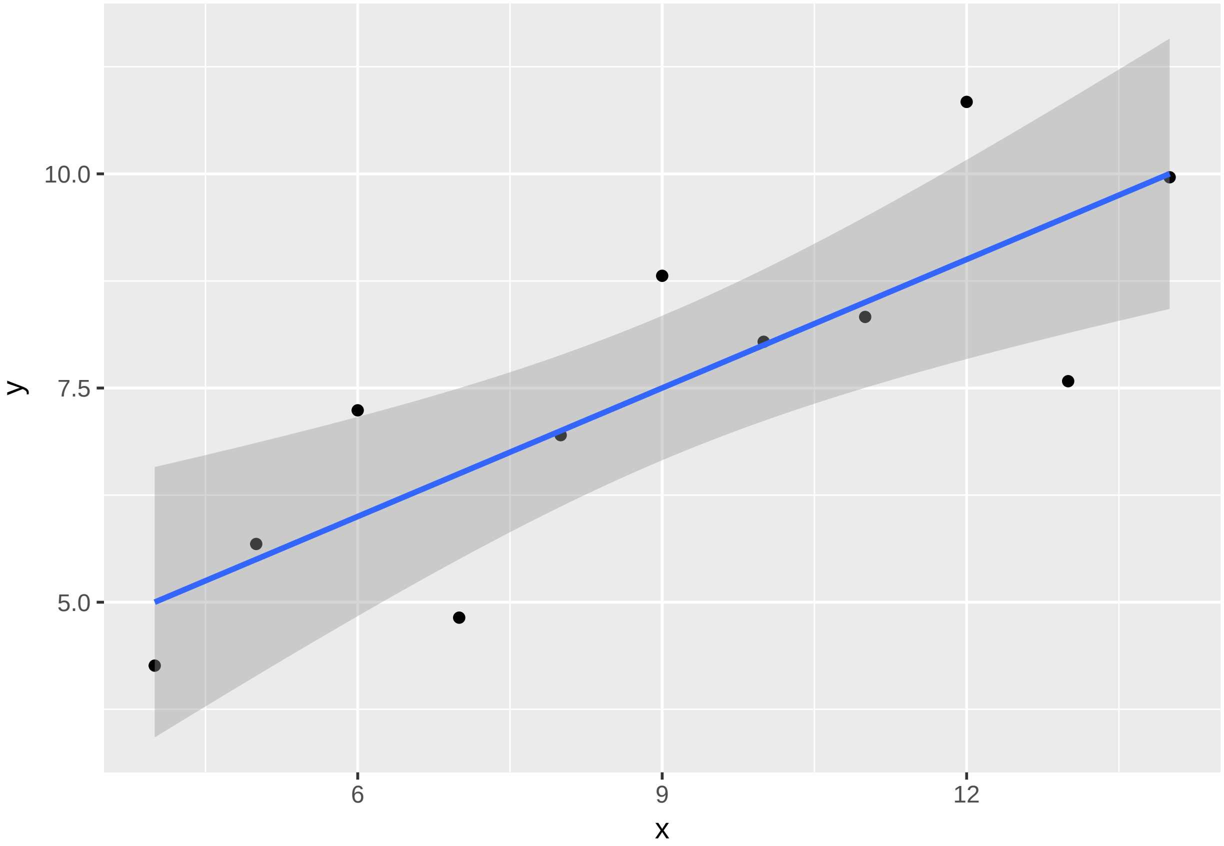
For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

Data 1

3 conditions for inference must be met:

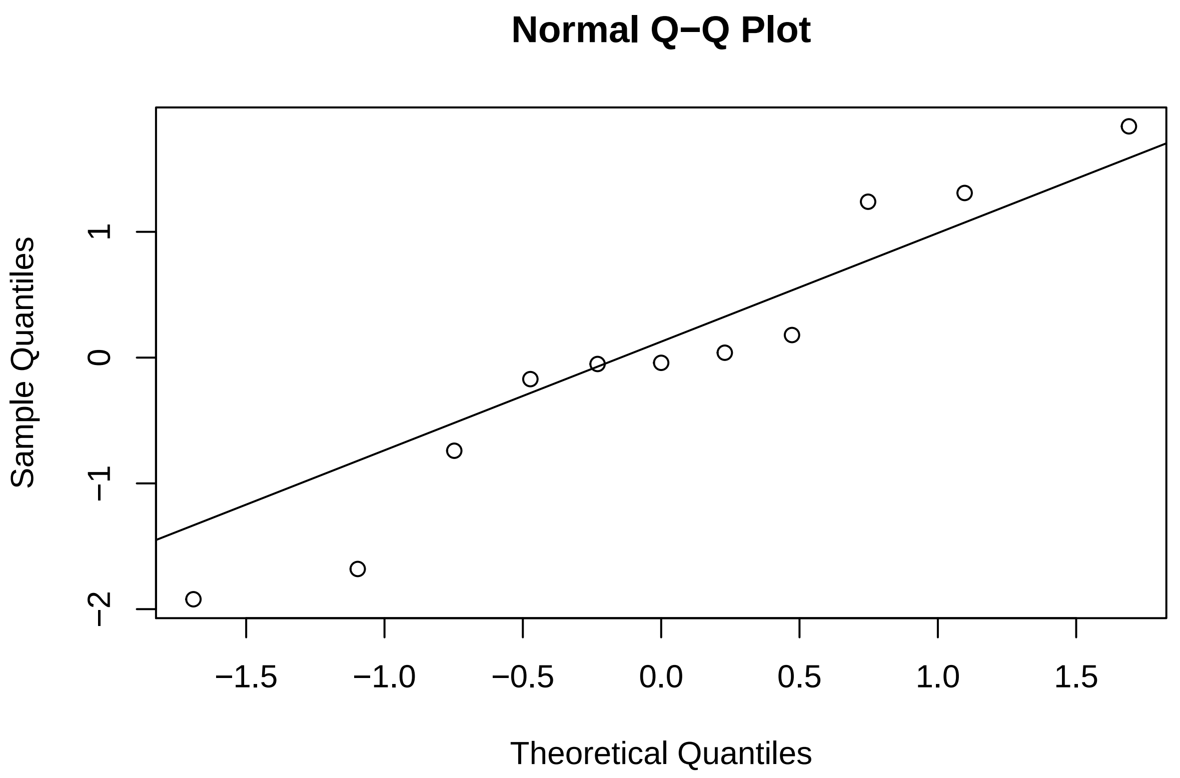
1: Linearity: Met

```
ggplot(data1, aes(x, y)) +
  geom_point() +
  geom_smooth(method=lm)
```



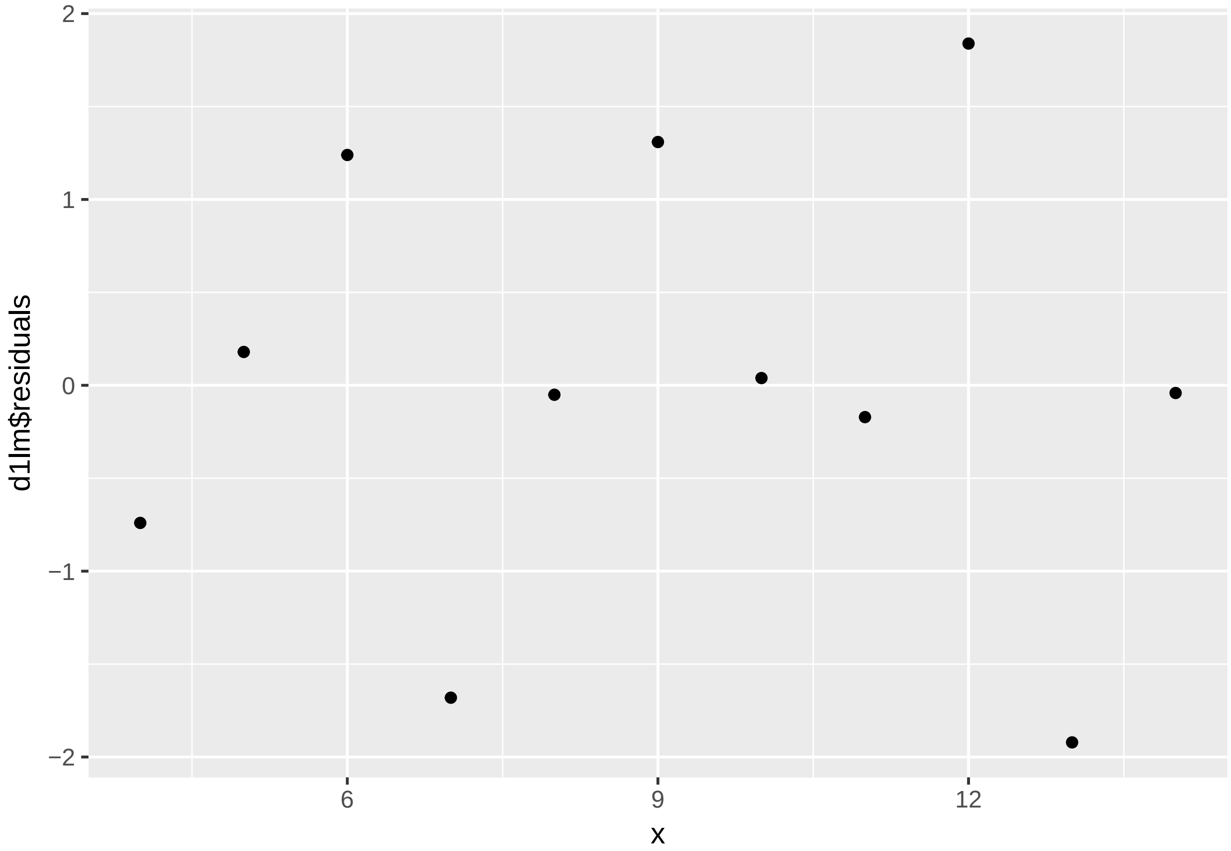
2: Nearly Normal Residuals: Met

```
qqnorm(d1lm$residuals)
qqline(d1lm$residuals)
```



3: Constant Variability: Met

```
ggplot(data1, aes(x, d1lm$residuals)) +
  geom_point()
```



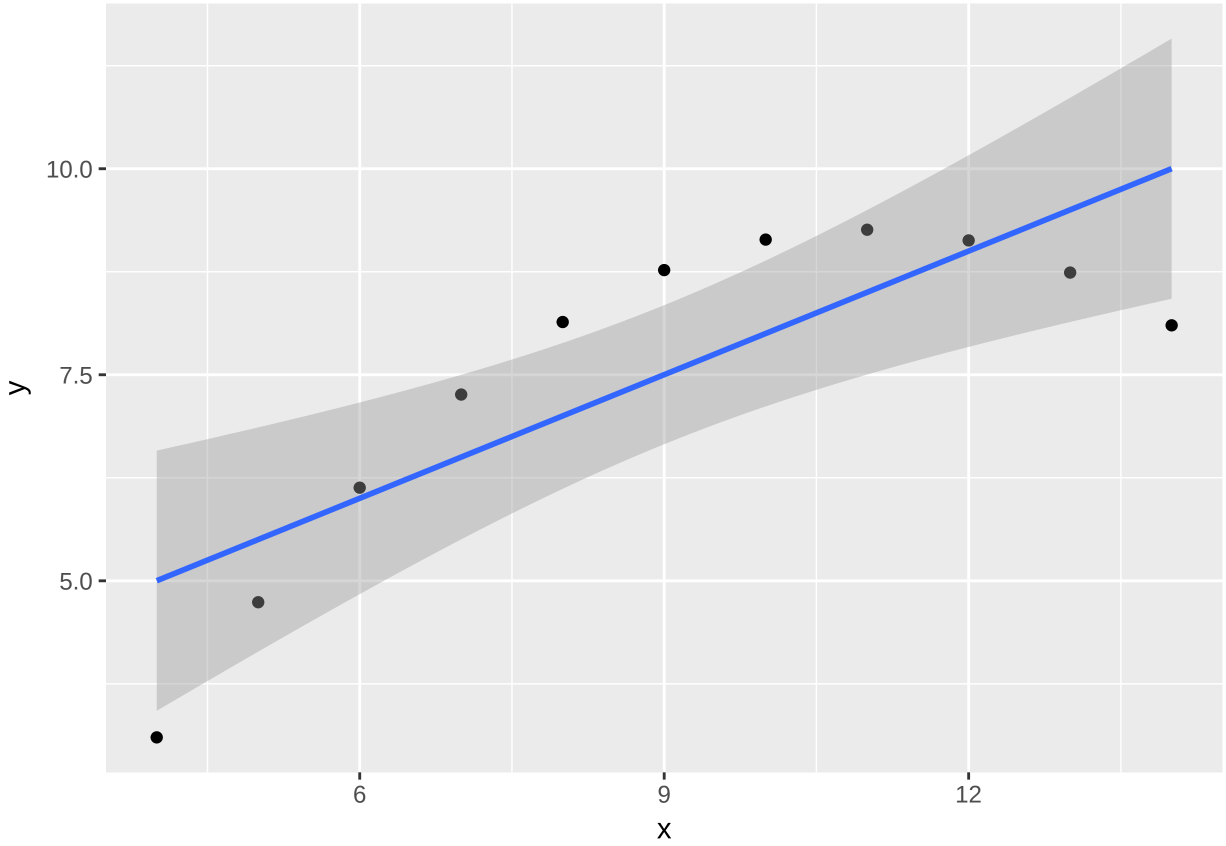
As all the conditions for inference have been met, a linear regression model **is** appropriate for Data 1

Data 2

3 conditions for inference must be met:

1: Linearity: Not Met

```
ggplot(data2, aes(x, y)) +
  geom_point() +
  geom_smooth(method=lm)
```



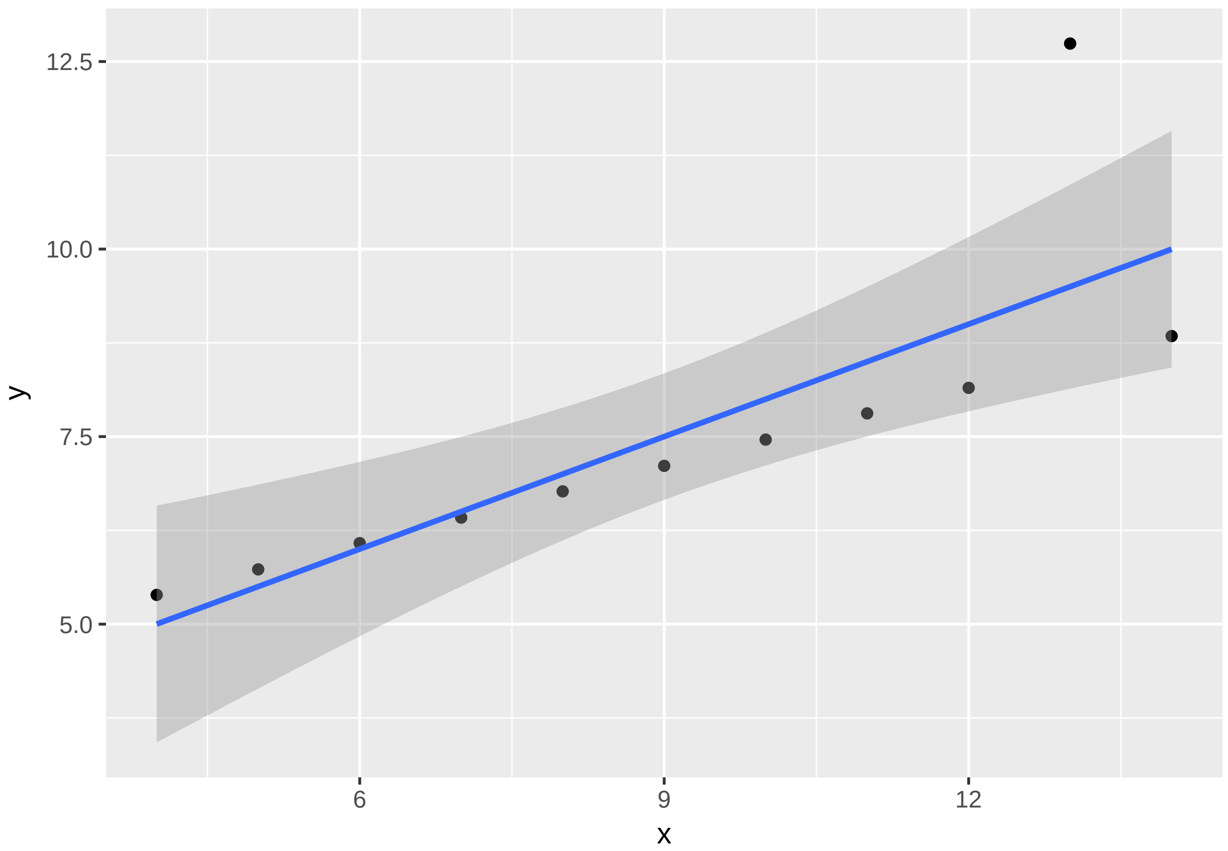
A linear regression model is **NOT** appropriate for Data 2

Data 3

3 conditions for inference must be met:

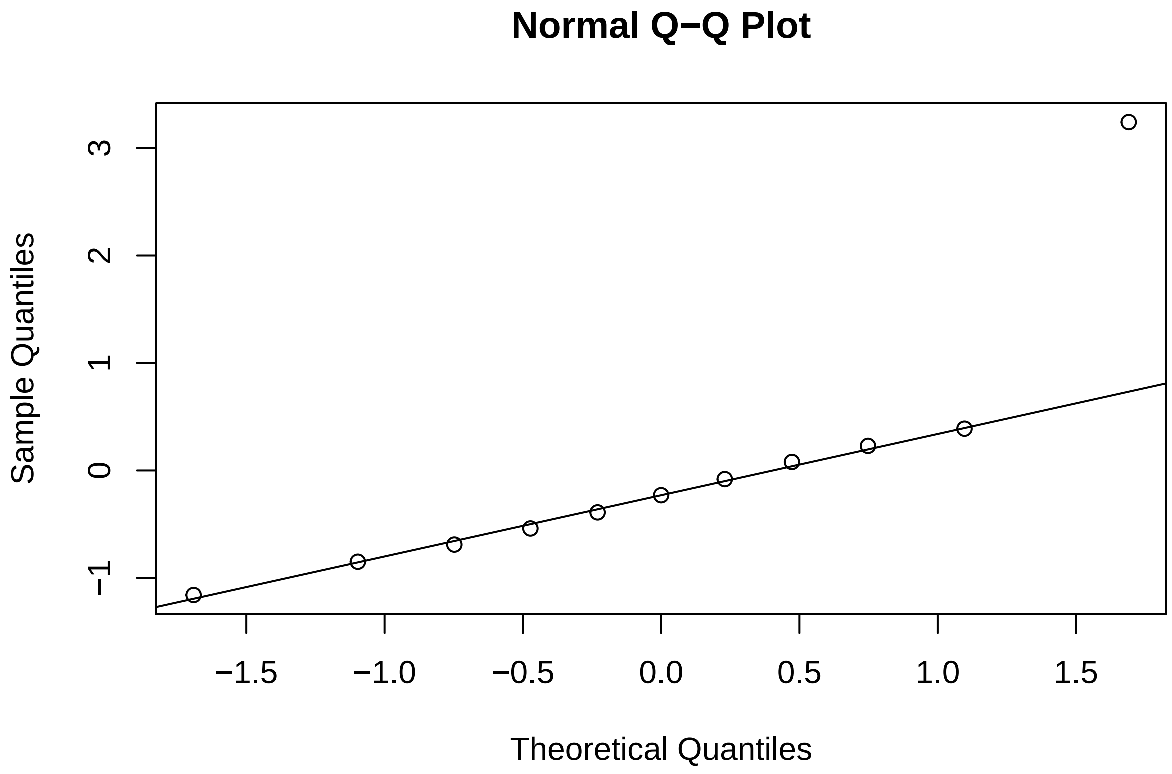
1: Linearity: Not Met

```
ggplot(data3, aes(x, y)) +
  geom_point() +
  geom_smooth(method=lm)
```



2: Nearly Normal Residuals: Not Met

```r
qqnorm(d3lm$residuals)
qqline(d3lm$residuals)
```



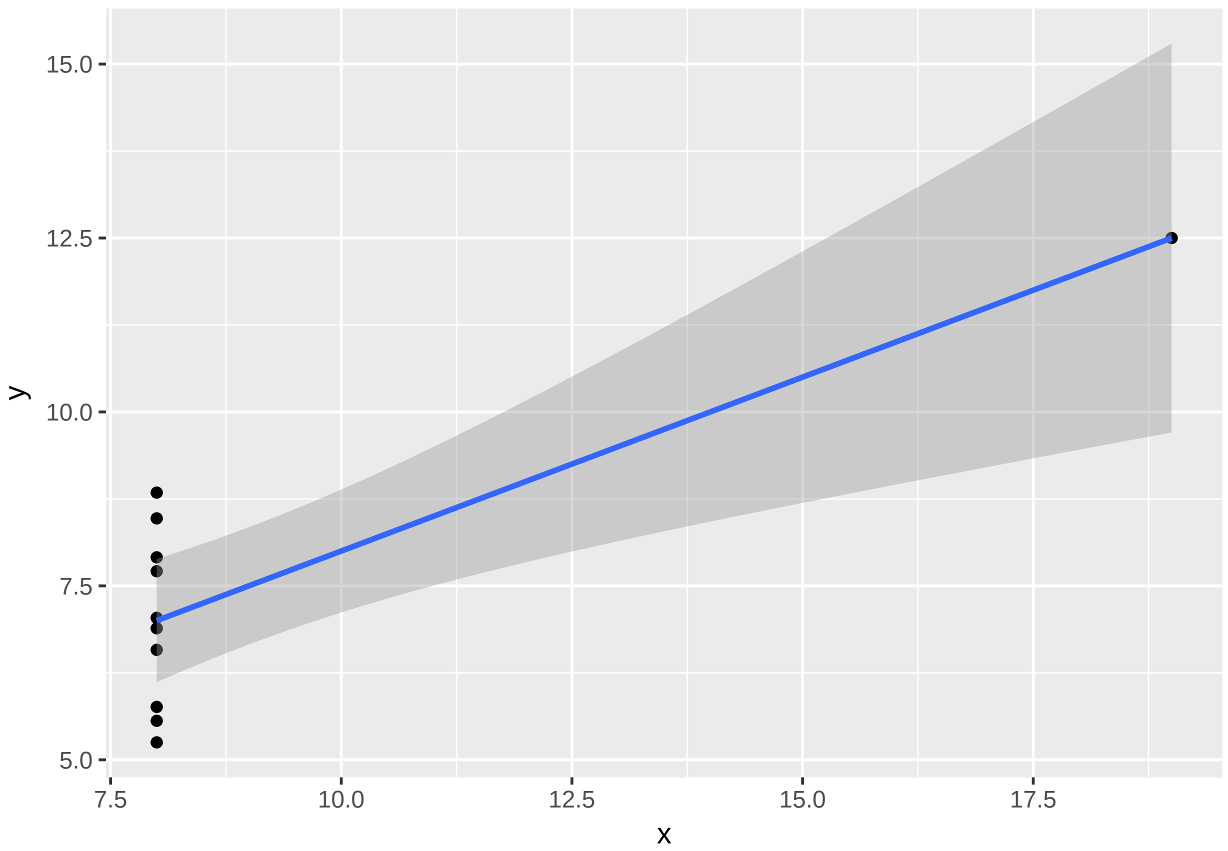
A linear regression model is **NOT** appropriate for Data 3

Data 4

3 conditions for inference must be met:

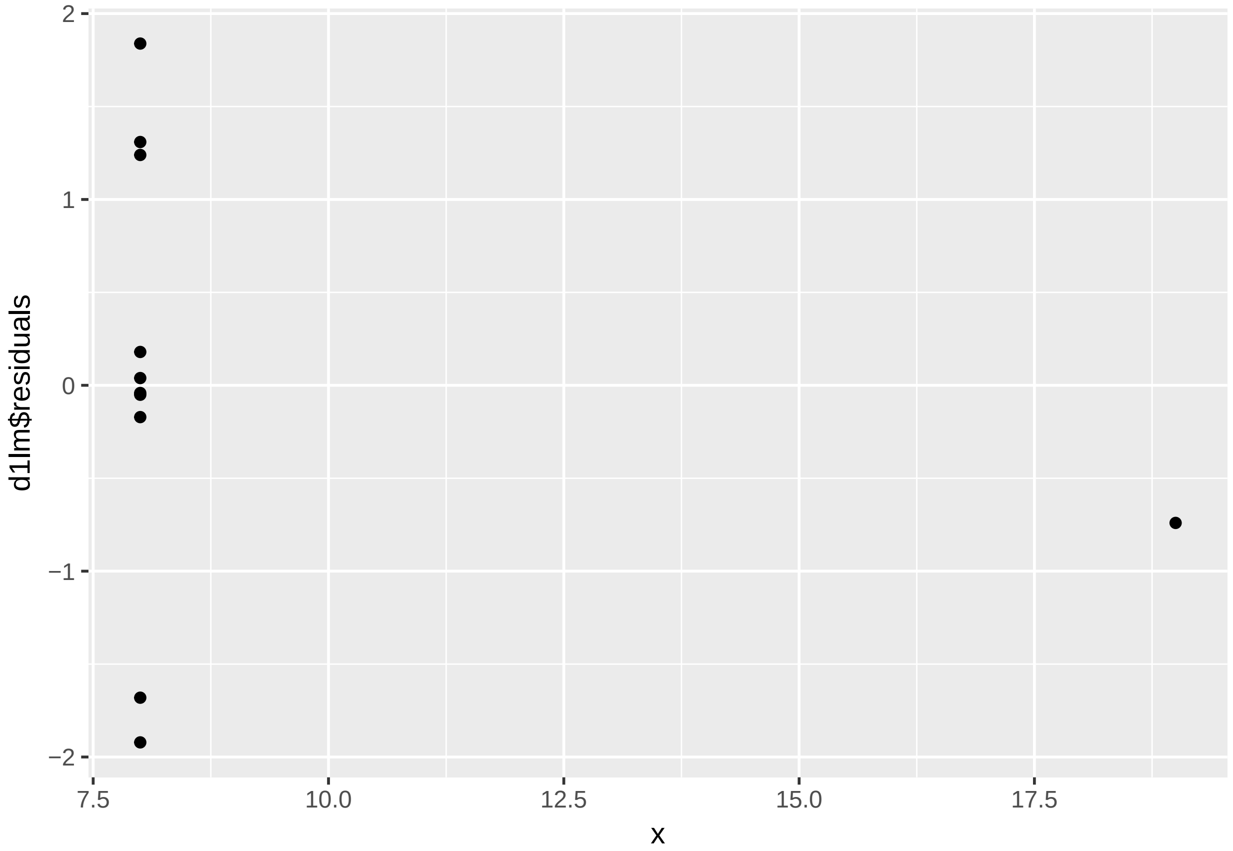
1: Linearity: Not Met

```
ggplot(data4, aes(x, y)) +
  geom_point() +
  geom_smooth(method=lm)
```



3: Constant Variability: Not Met

```
ggplot(data4, aes(x, d1lm$residuals)) +
  geom_point()
```



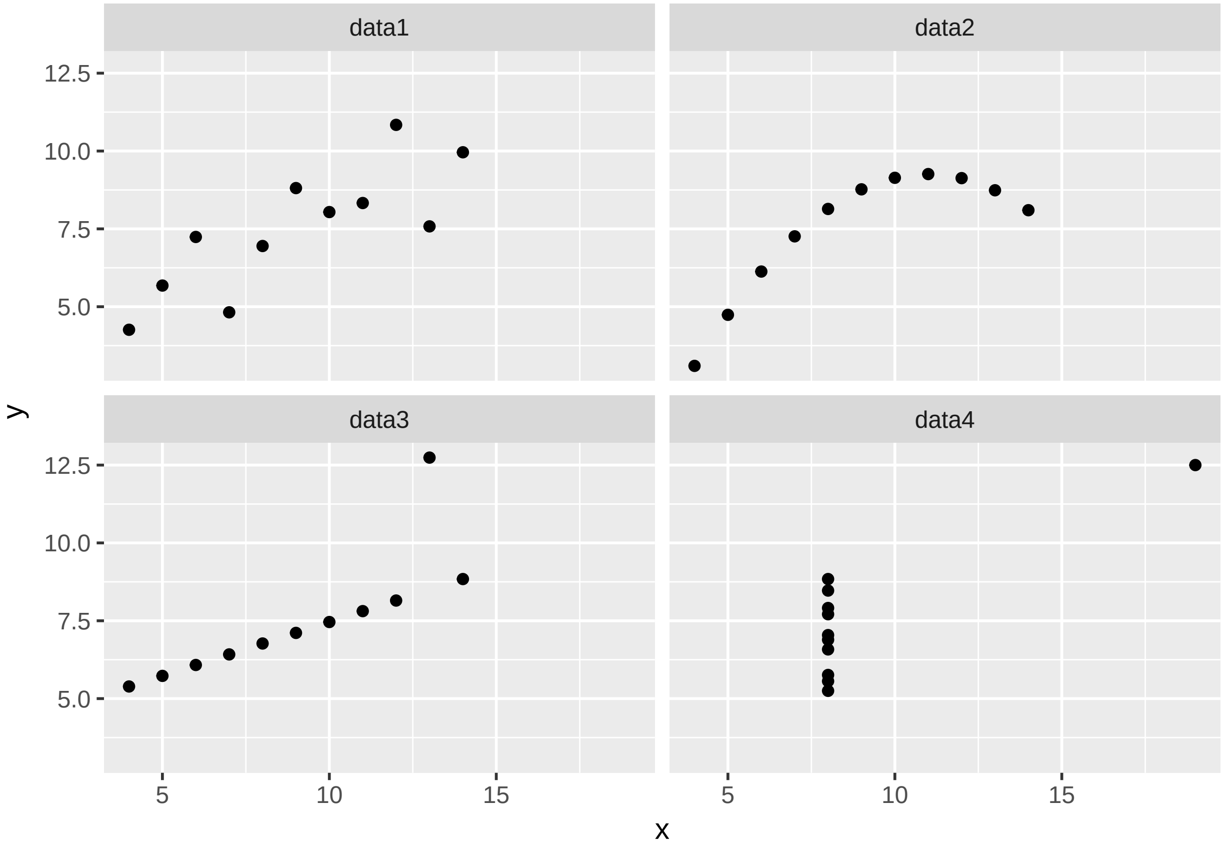
A linear regression model is **NOT** appropriate for Data 4

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

Visiualization are a critical tool for data analysis. For example, without visualizations, it would have been difficult to determine the very important differences in the four data sets. All of the data sets calculated statistical values were nearly identical and their linear regression models were as well. It was not until the data (and their respective conditions for inference) were plotted that the differences between the data sets became apparent.

All the calculations conducted above failed to accurately describe the differences between the data sets as well as the below single graph:

```
data1 %<>%
  mutate(type = 'data1')
data2 %<>%
  mutate(type = 'data2')
data3 %<>%
  mutate(type = 'data3')
data4 %<>%
  mutate(type = 'data4')
all.data <- rbind(data1, data2, data3, data4)
ggplot(all.data, aes(x, y)) +
  geom_point() +
  facet_wrap('type')
```


data1
data2
data3
data4
12.5
10.0
7.5
5.0
12.5
10.0
7.5
5.0
y
5
10
15
5
10
15
x