DATA 606 Spring 2020 - Final Exam

Chunjie Nan

Part I

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

1. b. daysDrive
2. a. mean = 3.3, median = 3.5
3. 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.
4. a. there is a difference between average eye color and average hair color.
5. b. 17.8 and 69.0
6. d. median and interquartile range; mean and standard deviation

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

Answer: Both distributions are unimodal, and A is skewed right compare to the B that seems symmetrically and normally distributed. Also, The spread of B is less than the spread A.

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

Answer: Since the distribution B is from the distribution A as sampled, therefore, the mean sould be similar. The standard deviation is different because the standard error whichi is the standard deviation of sampling is from the equation SE=SD/Mean, so the SE usually smaller value than the population standard deviation.

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

Answer: I believe the phenomenon is from the principal of Central Limit Theorem. The sample are independent and in random without strong skewness and normally distributed.

Part II

Consider the four datasets, each with two columns (x and y), provided below. Be sure to replace the `NA` with your answer for each part (e.g. assign the mean of `x` for `data1` to the `data1.x.mean` variable). When you Knit your answer document, a table will be generated with all the answers.

```
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):

```
data1.x.mean <- mean(data1$x)
data1.x.mean
```

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

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

```
data1.y.mean <- mean(data1$y)
data1.y.mean
```

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

```
data2.x.mean <- mean(data2$x)
data2.x.mean
```

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

```
data2.y.mean <- mean(data2$y)
data2.y.mean
```

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

```
data3.x.mean <- mean(data3$x)
data3.x.mean
```

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

```
data3.y.mean <- mean(data3$y)
data3.y.mean
```

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

```
data4.x.mean <- mean(data4$x)
data4.x.mean
```

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

```
data4.y.mean <- mean(data4$y)
data4.y.mean
```

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

```
data1.x.median <- median(data1$x)
data1.x.median
```

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

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

```
data1.y.median <- median(data1$y)
data1.y.median
```

```
## [1] 7.6
```

```
data2.x.median <- median(data2$x)
data2.x.median
```

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

```
data2.y.median <- median(data2$y)
data2.y.median
```

```
## [1] 8.1
```

```
data3.x.median <- median(data3$x)
data3.x.median
```

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

```
data3.y.median <- median(data3$y)
data3.y.median
```

```
## [1] 7.1
```

```
data4.x.median <- median(data4$x)
data4.x.median
```

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

```
data4.y.median <- median(data4$y)
data4.y.median
```

```
## [1] 7
```

```
data1.x.sd <- sd(data1$x)
data1.x.sd
```

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

```
## [1] 3.3
```

```
data1.y.sd <- sd(data1$y)
data1.y.sd
```

```
## [1] 2
```

```
data2.x.sd <- sd(data2$x)
data2.x.sd
```

```
## [1] 3.3
```

```
data2.y.sd <- sd(data2$y)
data2.y.sd
```

```
## [1] 2
```

```
data3.x.sd <- sd(data3$x)
data3.x.sd
```

```
## [1] 3.3
```

```
data3.y.sd <- sd(data3$y)
data3.y.sd
```

```
## [1] 2
```

```
data4.x.sd <- sd(data4$x)
data4.x.sd
```

```
## [1] 3.3
```

```
data4.y.sd <- sd(data4$y)
data4.y.sd
```

```
## [1] 2
```

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

```
data1.correlation <- cor(data1)
data1.correlation
```

d. The correlation (1 pt).

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

```
data2.correlation <- cor(data2)
data2.correlation
```

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

```
data3.correlation <- cor(data3)
data3.correlation
```

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

```
data4.correlation <- cor(data4)
data4.correlation
```

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

```
model1<-lm(y ~ x, data=data1)
model2<-lm(y ~ x, data=data2)
model3<-lm(y ~ x, data=data3)
model4<-lm(y ~ x, data=data4)
```

```
data1.slope <- coef(model1)[2]
data1.slope
```

e. Linear regression equation (2 pts).

```
##   x
## 0.5
```

```
data2.slope <- coef(model2)[2]
data2.slope
```

```
##   x
## 0.5
```

```
data3.slope <- coef(model3)[2]
data3.slope
```

```
##   x
## 0.5
```

```
data4.slope <- coef(model4)[2]
data4.slope
```

```
##   x
## 0.5
```

```
data1.intercept <- coef(model1)[1]
data1.intercept
```

```
## (Intercept)
##           3
```

```
data2.intercept <- coef(model2)[1]
data2.intercept
```

```
## (Intercept)
##           3
```

```
data3.intercept <- coef(model3)[1]
data3.intercept
```

```
## (Intercept)
##           3
```

```
data4.intercept <- coef(model4)[1]
data4.intercept
```

```
## (Intercept)
##           3
```

Answer: All the data has the same equation of yhat= 3 + 0.5 * x

```
data1.rsquared <- summary(model1)$r.squared
data1.rsquared
```

f. R-Squared (2 pts).

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

```
data2.rsquared <- summary(model2)$r.squared
data2.rsquared
```

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

```
data3.rsquared <- summary(model3)$r.squared
data3.rsquared
```

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

```
data4.rsquared <- summary(model4)$r.squared
data4.rsquared
```

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

```
## Error in data.frame(data1.x = c(data1.x.mean, data1.x.median, data1.x.sd, : arguments imply differing
```

```
## Error in row.names(results) <- c("Mean", "Median", "SD", "r", "Intercept", : object 'results' not fou
```

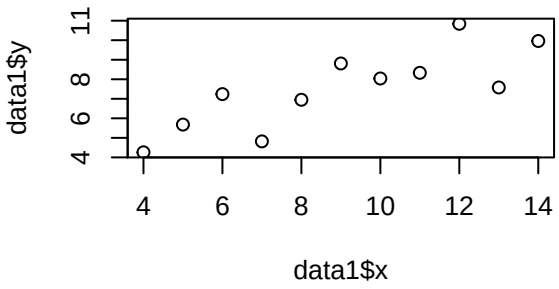
```
## Error in names(results) <- c("x", "y", "x", "y", "x", "y", "x", "y"): object 'results' not found
```

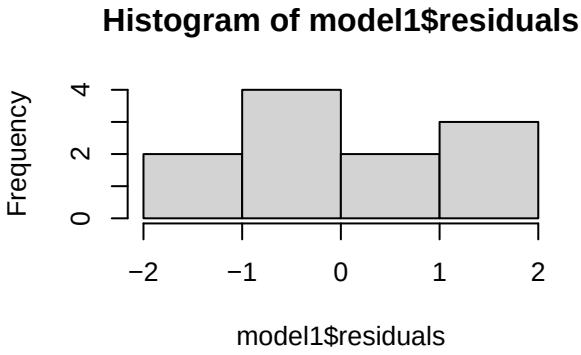
```
## Error in kable(results, digits = 2, align = "r", row.names = TRUE, format.args = list(nsmall = 2)): o
```

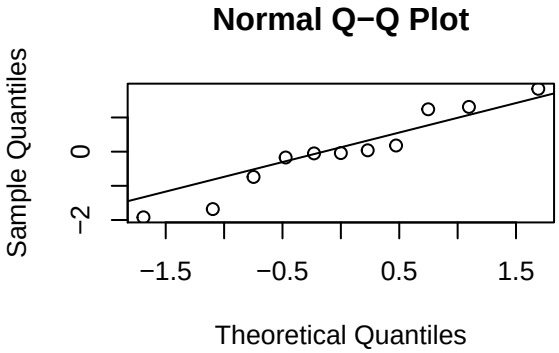
g. For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts) Answer: We need to see assumptions as below to check a linear regression model. 1. Linearity 2. Risiduals are normal 3. constant variables 4. independence

data1

```
par(mfrow=c(2,2))
plot(data1$x,data1$y)
hist(model1$residuals)
qqnorm(model1$residuals)
qqline(model1$residuals)
```



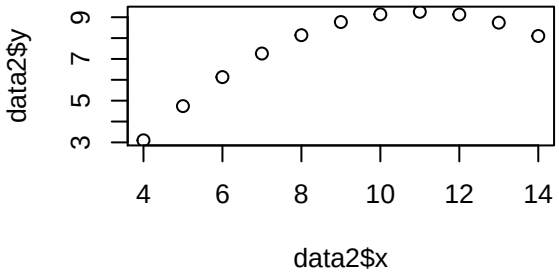


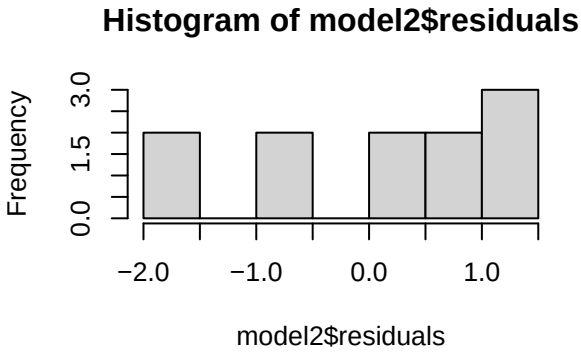


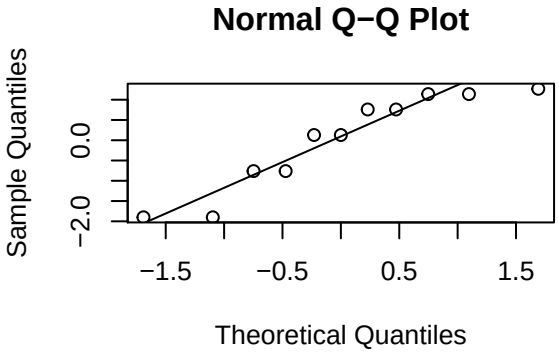
The graph for data 1 shows slight linearity and fair Q-Q plot, but the histogram for residuals is not clear to show a normal, so data 1 isn't appropriate to create a linear model.

data2

```
par(mfrow=c(2,2))
plot(data2$x,data2$y)
hist(model2$residuals)
qqnorm(model2$residuals)
qqline(model2$residuals)
```



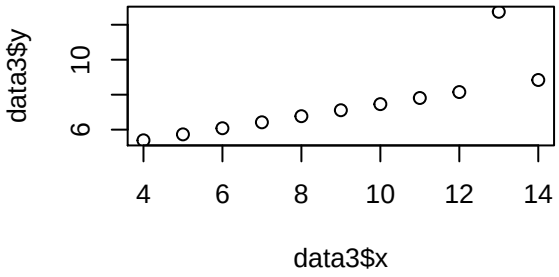


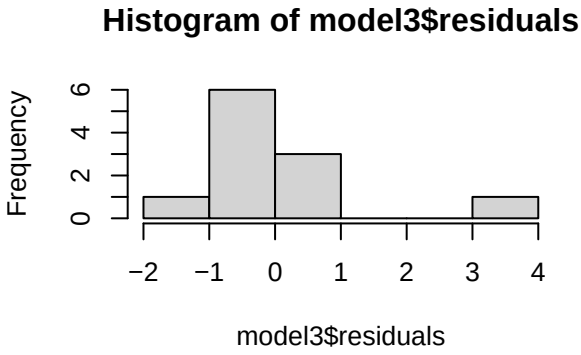


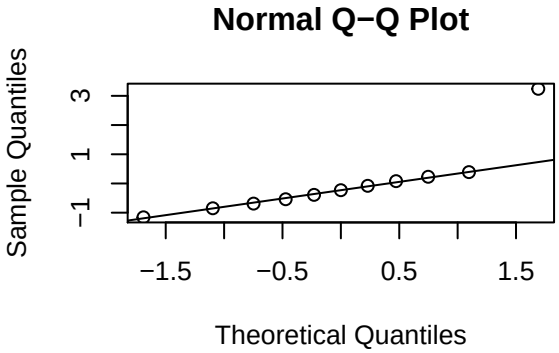
The graph for data 2 doesn't show linearity which is curved. Therefore, data 2 is not appropriate to create a linear model.

data3

```
par(mfrow=c(2,2))
plot(data3$x,data3$y)
hist(model3$residuals)
qqnorm(model3$residuals)
qqline(model3$residuals)
```



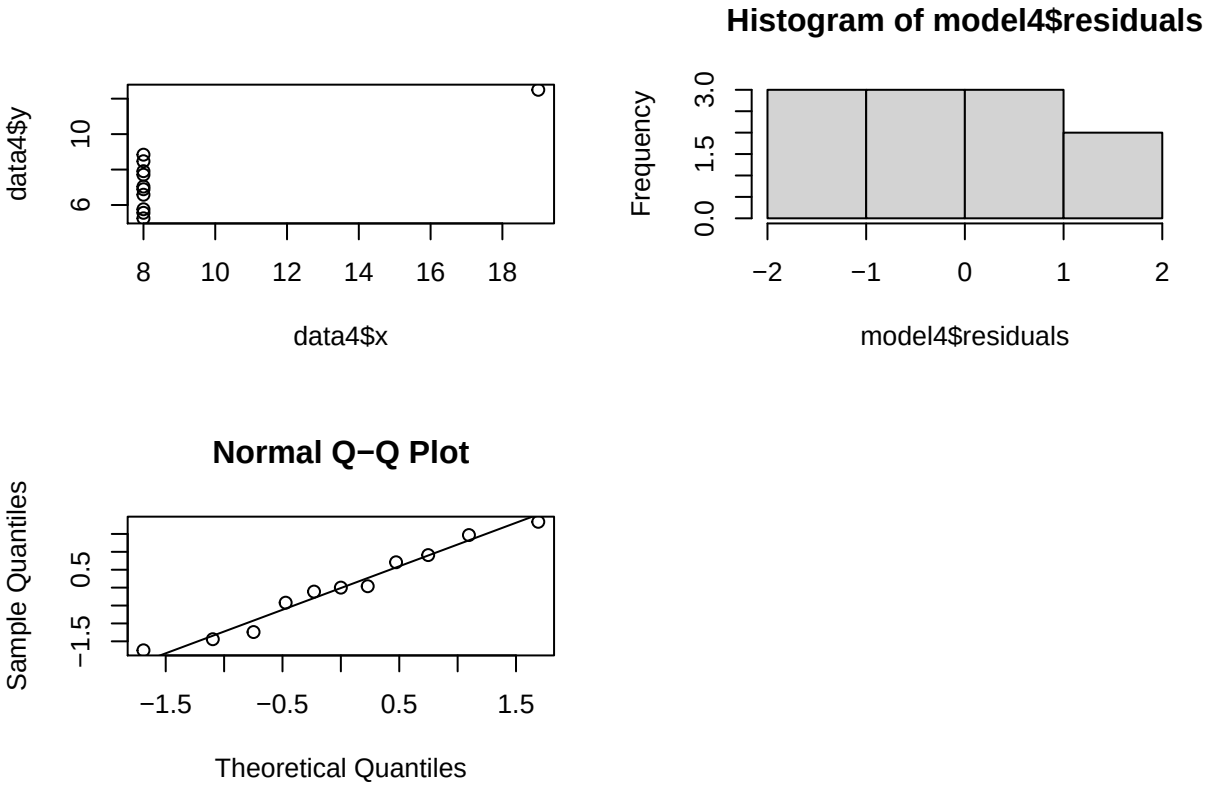




The graph for data 3 shows linearity, normal distribution of residuals, but there is a strong outlier that may affect the linear model. Therefore, not recommend use the data3 to create a linear model.

data4

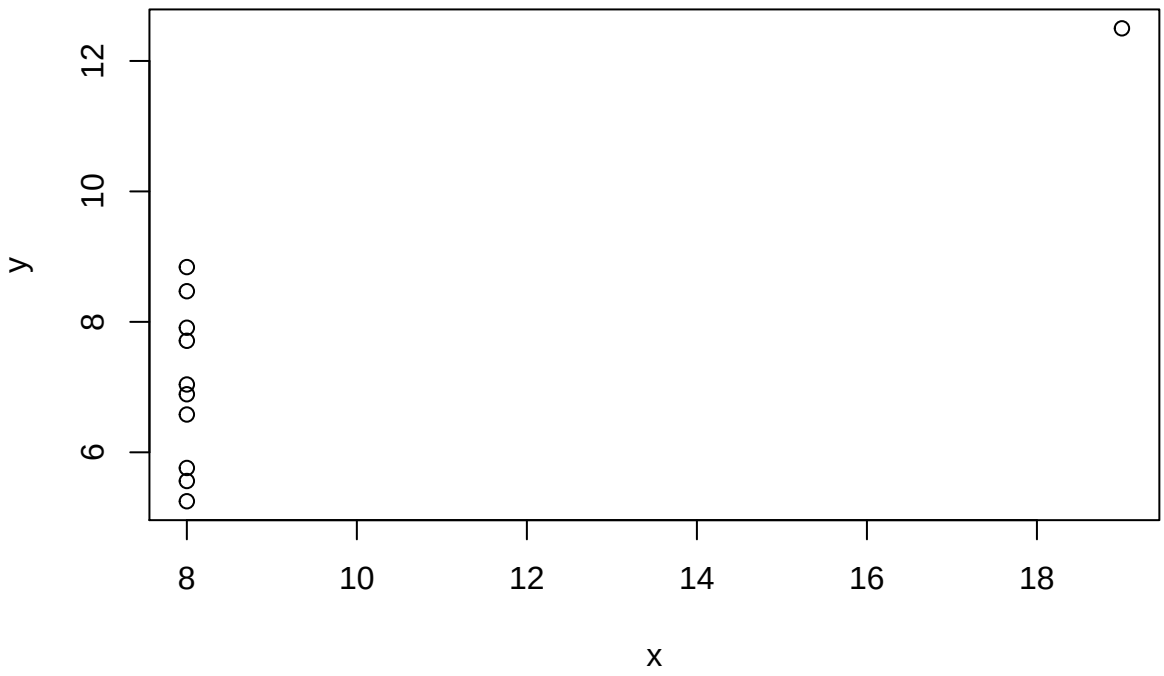
```
par(mfrow=c(2,2))
plot(data4$x,data4$y)
hist(model4$residuals)
qqnorm(model4$residuals)
qqline(model4$residuals)
```



The graph for data4, the first graph we see many ys at the same x, and the histogram is not normal, which tells us there are no relationships between x and y. so, we cannot create the linear model to see any relationship between x and y.

h. Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (2 pts) Answer: As the g above, when we determine if a data is appropriate to create a linear model, the easiest and convinient way is using plot and graph to see if the dataset meets the conditions of a linear model. For example, for the data4, we can very quickly figure out that the data doesn't show any relationships between x and y. So that, we can save out time to analyze the data compare with looking at the numbers in the data.

```
plot(data4)
```


y
12
10
8
6
8
10
12
14
16
18
x