DATA 606 Spring 2018 - Final Exam

Ohannes Ohannessian

Part I

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

1. b
2. a
3. d
4. a
5. b
6. d

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

A: right skewed, unimodal
B: normal distribution, unimodal

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

The standard deviation (sd) of the observation describes the spread of its values in the sample. The standard deviation of the sample mean is the standard error of the mean (SE). It describes the accuracy as an estimate of the population mean μ.

$SE = sd_B = \frac{\sigma}{\sqrt{n}} = \frac{3.22}{\sqrt{30}} = 0.5878889$

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

It's the 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.frame(data=c("data1","data2","data3","data4"), mean_x=c(round(mean(data1$x),2),roun
d(mean(data2$x),2),round(mean(data3$x),2),round(mean(data4$x),2)), mean_y=c(round(mean(d
ata1$y),2),round(mean(data2$y),2),round(mean(data3$y),2),round(mean(data4$y),2)))
```

data <fctr>	mean_x <dbl>	mean_y <dbl>
data1	9	7.5
data2	9	7.5
data3	9	7.5
data4	9	7.5

4 rows

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

```
data.frame(data=c("data1","data2","data3","data4"), median_x=c(round(median(data1$x),2),
round(median(data2$x),2),round(median(data3$x),2),round(median(data4$x),2)), median_y=c
(round(median(data1$y),2),round(median(data2$y),2),round(median(data3$y),2),round(median
(data4$y),2)))
```

data <fctr>	median_x <dbl>	median_y <dbl>
data1	9	7.58
data2	9	8.14
data3	9	7.11
data4	8	7.04

4 rows

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

```
data.frame(data=c("data1","data2","data3","data4"), sd_x=c(round(sd(data1$x),2),round(sd
(data2$x),2),round(sd(data3$x),2),round(sd(data4$x),2)), sd_y=c(round(sd(data1$y),2),rou
nd(sd(data2$y),2),round(sd(data3$y),2),round(sd(data4$y),2)))
```

data <fctr>	sd_x <dbl>	sd_y <dbl>
data1	3.32	2.03
data2	3.32	2.03
data3	3.32	2.03
data4	3.32	2.03

4 rows

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

d. The correlation (1 pt).

```
data.frame(data1 = cor(data1))
```

	data1.x <dbl>	data1.y <dbl>
x	1.0000000	0.8164205
y	0.8164205	1.0000000

2 rows

```
data.frame(data2 = cor(data2))
```

	data2.x <dbl>	data2.y <dbl>
x	1.0000000	0.8162365
y	0.8162365	1.0000000

2 rows

```
data.frame(data3 = cor(data3))
```

	data3.x <dbl>	data3.y <dbl>
x	1.0000000	0.8162867
y	0.8162867	1.0000000

2 rows

```
data.frame(data4 = cor(data4))
```

	data4.x <dbl>	data4.y <dbl>
x	1.0000000	0.8165214
y	0.8165214	1.0000000

2 rows

e. Linear regression equation (2 pts).

```
lm_data1 <- lm(data1$y ~ data1$x)
lm_data2 <- lm(data2$y ~ data2$x)
lm_data3 <- lm(data3$y ~ data3$x)
lm_data4 <- lm(data4$y ~ data4$x)
```

Linear regression equation ($y = \beta_0 + \beta_1 * x$) for:

data1: $y = 3 + 0.5 * x$

data2: $y = 3.001 + 0.5 * x$

data3 :$y = 3.002 + 0.5 * x$

data4: $y = 3.002 + 0.5 * x$

f. R-Squared (2 pts).

```
data.frame(data=c("data1","data2","data3","data4"), r_squared=c(summary(lm_data1)$r.squared,summary(lm_data2)$r.squared,summary(lm_data3)$r.squared,summary(lm_data4)$r.squared))
```

data <fctr>	r_squared <dbl>
data1	0.6665425
data2	0.6662420
data3	0.6663240
data4	0.6667073

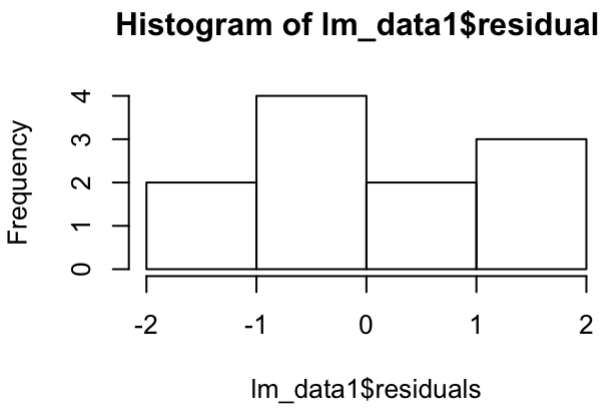
4 rows

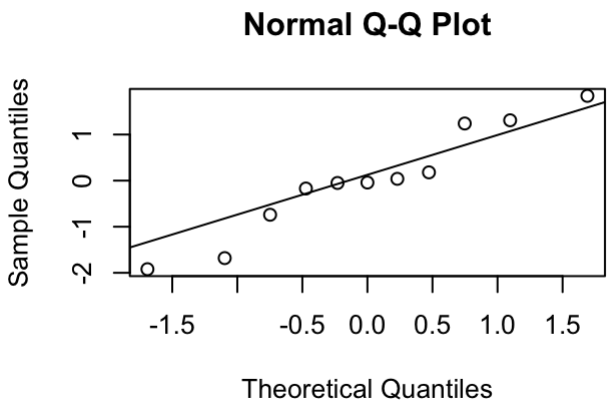
For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

- Data1: No, the plot seem to look an S-shape

```
par(mfrow=c(2,2))
hist(lm_data1$residuals)

qqnorm(lm_data1$residuals)
qqline(lm_data1$residuals)
```

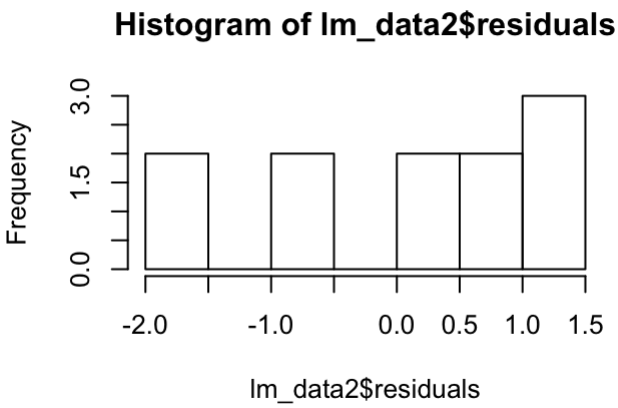


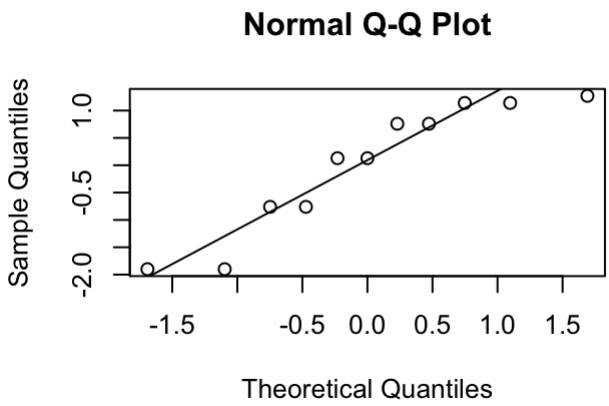


- Data2: No, the plot seem to look an S-shape

```
par(mfrow=c(2,2))
hist(lm_data2$residuals)

qqnorm(lm_data2$residuals)
qqline(lm_data2$residuals)
```

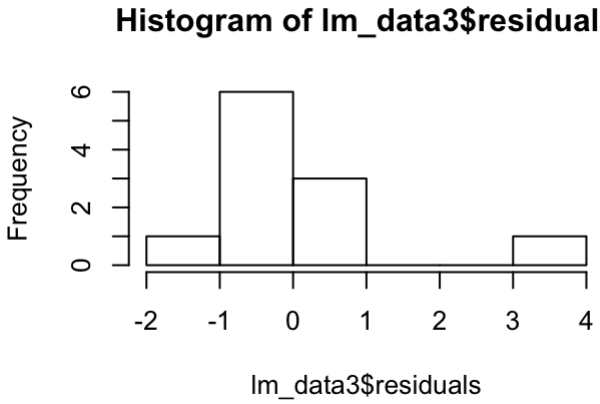


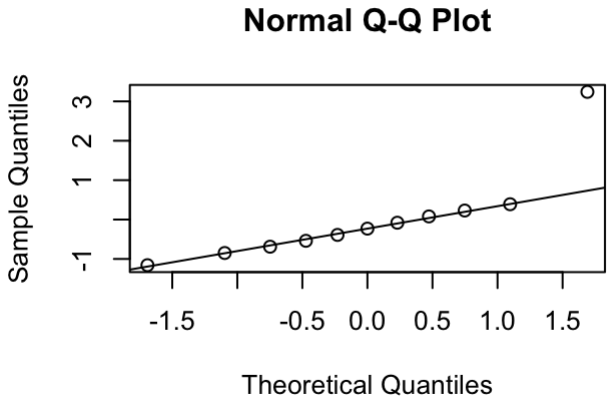


- Data3: Yes, the plot seem to look linear

```
par(mfrow=c(2,2))
hist(lm_data3$residuals)

qqnorm(lm_data3$residuals)
qqline(lm_data3$residuals)
```

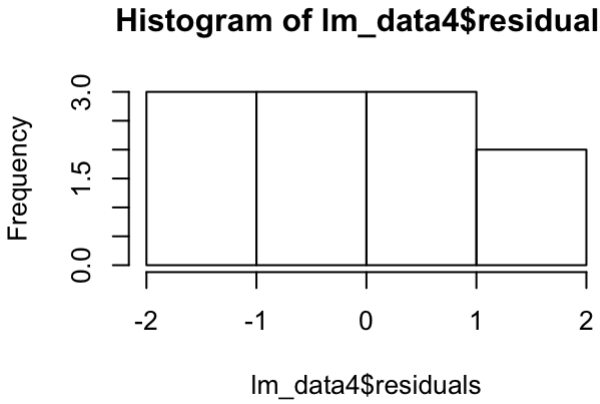


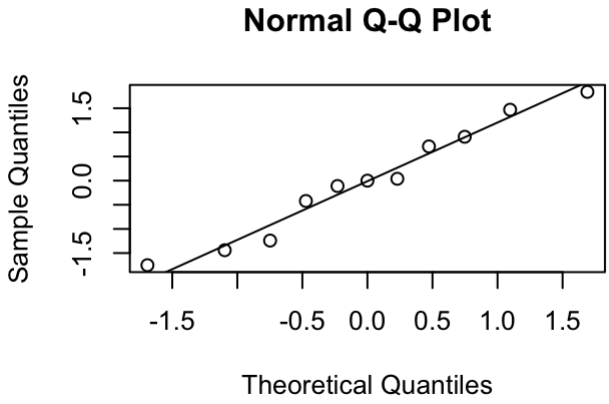


- Data4: No, the plot seem to look an S-shape

```
par(mfrow=c(2,2))
hist(lm_data4$residuals)

qqnorm(lm_data4$residuals)
qqline(lm_data4$residuals)
```





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

- In visualization, analytics presented visually can help see difficult concepts or identify new patterns
- Patterns or trends that might go unnoticed in text-based data can be exposed and recognized easier with data visualization