DATA 606 Fall 2021 - Final Exam

Leticia Salazar

Part I

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

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

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

Both distributions are unimodal and appear to be normally distributed with no equal means. Figure A, Observations, is right skewed while Figure B, Sampling Distribution, appears to be symmetric.

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

The mean of the two distributions are similar because Figure B is a sample distribution of the total population in Figure A. The standard deviations for both distributions are different because Figure B has a wider spread since it is a sample distribution from the total population, it is only taking into account the sample size and not the population as a whole.

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

The Central Limit Theorem which according to the Open Intro Statistics book: When observations are independent and the sample size is sufficiently large, the sample proportion p-hat will tend to follow a normal distribution

Part II

Consider the three 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.

For each column, calculate (to four decimal places):

```
options(digits = 4)
data1.x.mean <- mean(data1$x)
data1.x.mean
```

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

```
## [1] 54.26
```

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

```
## [1] 47.83
```

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

```
## [1] 54.27
```

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

```
## [1] 47.84
```

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

```
## [1] 54.27
```

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

```
## [1] 47.83
```

```
options(digits = 4)
data1.x.median <- median(data1$x)
data1.y.median <- median(data1$y)
data2.x.median <- median(data2$x)
data2.y.median <- median(data2$y)
data3.x.median <- median(data3$x)
data3.y.median <- median(data3$y)

data1.x.median
```

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

```
## [1] 53.33
```

```
data1.y.median
```

```
## [1] 46.03
```

```
data2.x.median
```

```
## [1] 53.14
```

```
data2.y.median
```

```
## [1] 46.4
```

```
data3.x.median
```

```
## [1] 53.34
```

```
data3.y.median
```

```
## [1] 47.54
```

```
options(digits = 4)
data1.x.sd <- sd(data1$x)
data1.y.sd <- sd(data1$y)
data2.x.sd <- sd(data2$x)
data2.y.sd <- sd(data2$y)
data3.x.sd <- sd(data3$x)
data3.y.sd <- sd(data3$y)

data1.x.sd
```

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

```
## [1] 16.77
```

```
data1.y.sd
```

```
## [1] 26.94
```

```
data2.x.sd
```

```
## [1] 16.77
```

```
data2.y.sd
```

```
## [1] 26.94
```

```
data3.x.sd
```

```
## [1] 16.77
```

```
data3.y.sd
```

```
## [1] 26.94
```

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

```
options(digits = 2)
data1.correlation <- cor(data1)[1,2]
data1.correlation
```

d. The correlation (1 pt).

```
## [1] -0.064
```

```
data2.correlation <- cor(data2)[1,2]
data2.correlation
```

```
## [1] -0.069
```

```
data3.correlation <- cor(data3)[1,2]
data3.correlation
```

```
## [1] -0.064
```

```
options(digits = 2)

# Data 1
lm_data1 <- lm(y ~ x, data = data1)
lm_data1
```

e. Linear regression equation (2 pts).

```
##
## Call:
## lm(formula = y ~ x, data = data1)
##
## Coefficients:
## (Intercept)            x
##      53.453       -0.104
```

```r
data1.slope <- coef(lm_data1)["x"]
data1.slope
```

```
##    x
## -0.1
```

```r
data1.intercept <- coef(lm_data1)["(Intercept)"]
data1.intercept
```

```
## (Intercept)
##          53
```

```r
# Data 2
lm_data2 <- lm(y ~ x, data = data2)
lm_data2
```

```
##
## Call:
## lm(formula = y ~ x, data = data2)
##
## Coefficients:
## (Intercept)            x
##      53.850       -0.111
```

```r
data2.slope <- coef(lm_data2)["x"]
data2.slope
```

```
##     x
## -0.11
```

```r
data2.intercept <- coef(lm_data2)["(Intercept)"]
data2.intercept
```

```
## (Intercept)
##          54
```

```r
# Data 3
lm_data3 <- lm(y ~ x, data = data3)
lm_data3
```

```
##
## Call:
## lm(formula = y ~ x, data = data3)
##
## Coefficients:
## (Intercept)            x
##      53.425       -0.103
```

```
data3.slope <- coef(lm_data3)["x"]
data3.slope
```

```
##    x
## -0.1
```

```
data3.intercept <- coef(lm_data3)["(Intercept)"]
data3.intercept
```

```
## (Intercept)
##          53
```

```
options(digits = 2)
data1.rsquared <- summary(lm_data1)$r.squared
data1.rsquared
```

f. R-Squared (2 pts).

```
## [1] 0.0042
```

```
data2.rsquared <- summary(lm_data2)$r.squared
data2.rsquared
```

```
## [1] 0.0048
```

```
data3.rsquared <- summary(lm_data3)$r.squared
data3.rsquared
```

```
## [1] 0.0041
```

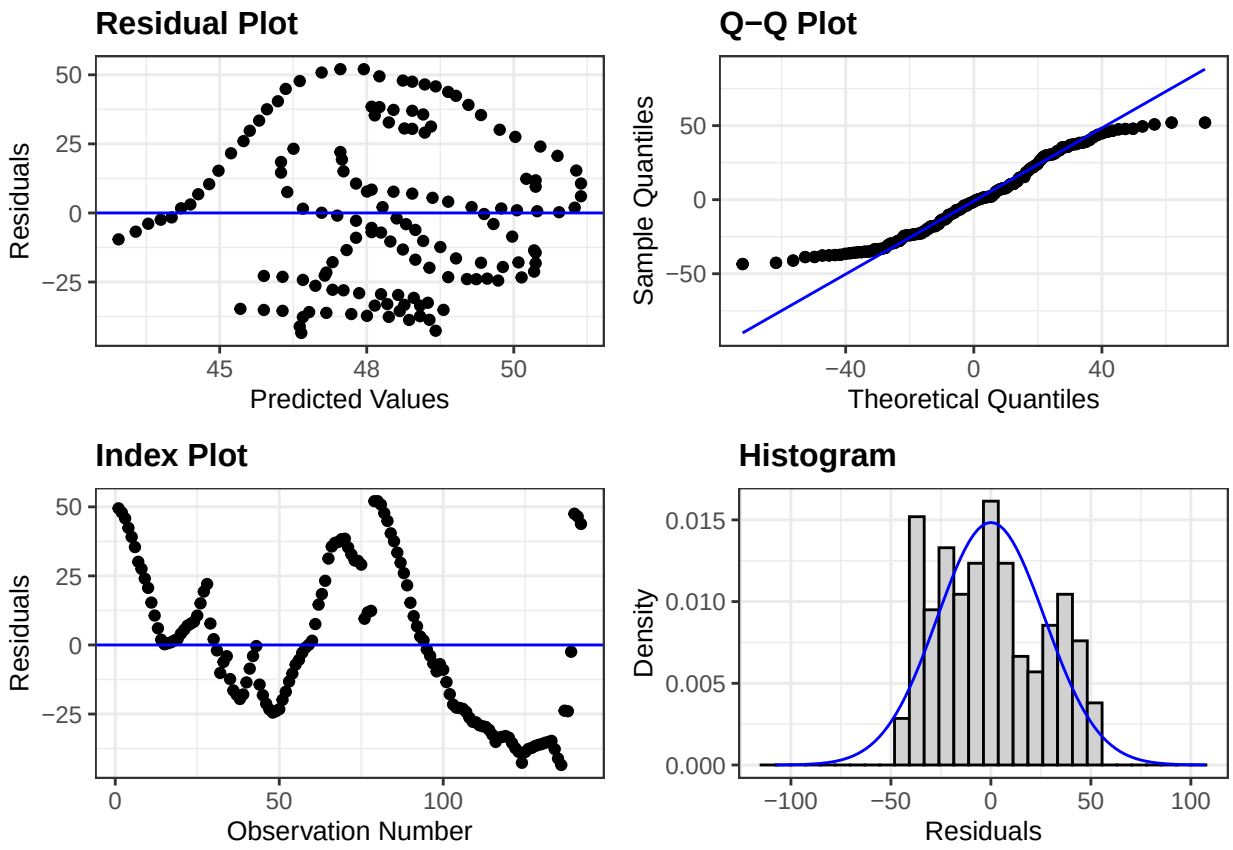
Summary Table

	Data 1		Data 2		Data 3	
	x	y	x	y	x	y
Mean	54.2633	47.8323	54.2678	47.8359	54.2661	47.8347
Median	53.3333	46.0256	53.1352	46.4013	53.3403	47.5353
SD	16.7651	26.9354	16.7668	26.9361	16.7698	26.9397
r	-0.0645		-0.0690		-0.0641	
Intercept	53.4530		53.8497		53.4251	
Slope	-0.1036		-0.1108		-0.1030	
R-Squared	0.0042		0.0048		0.0041	

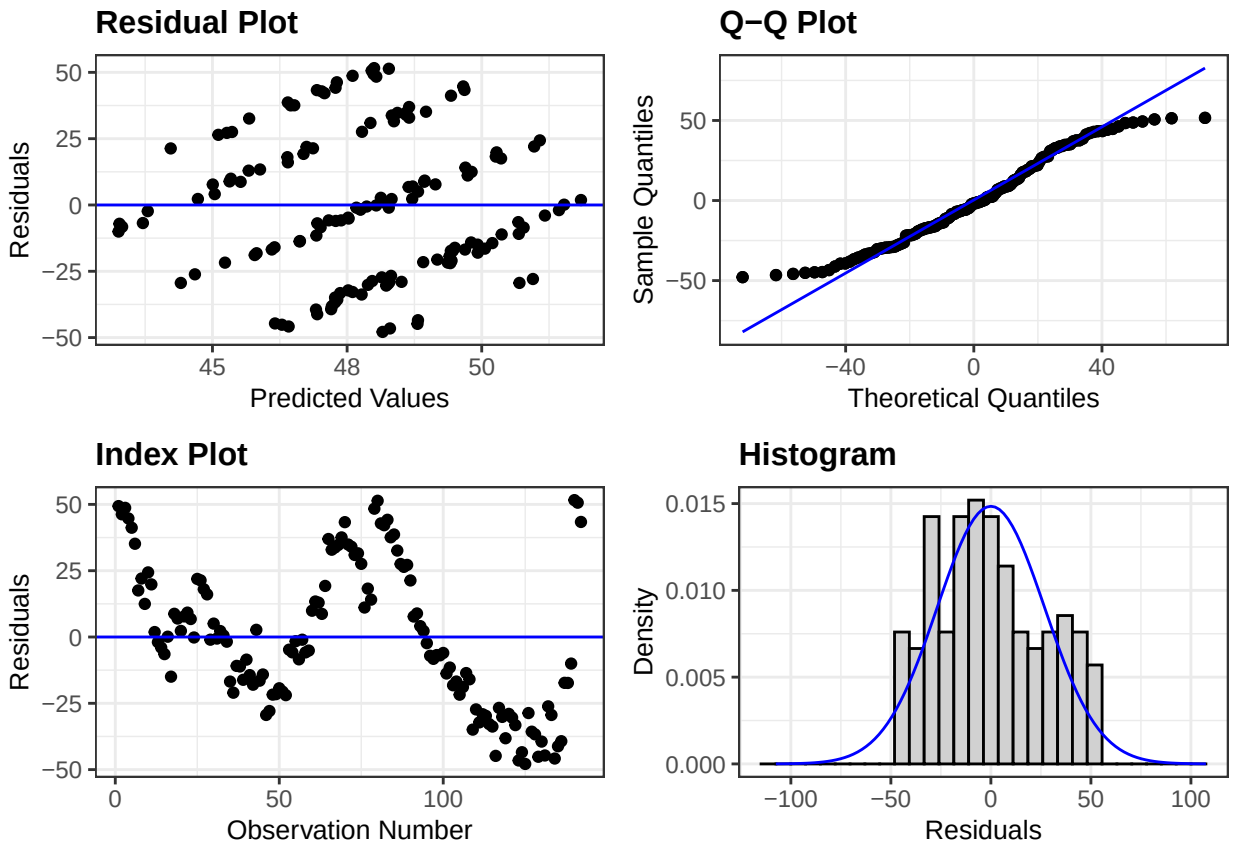
g. For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

It is appropriate to estimate a linear regression model for all data sets because they follow the linearity test. All 3 data sets appear to be normally distributed and the residuals are centered across 0.

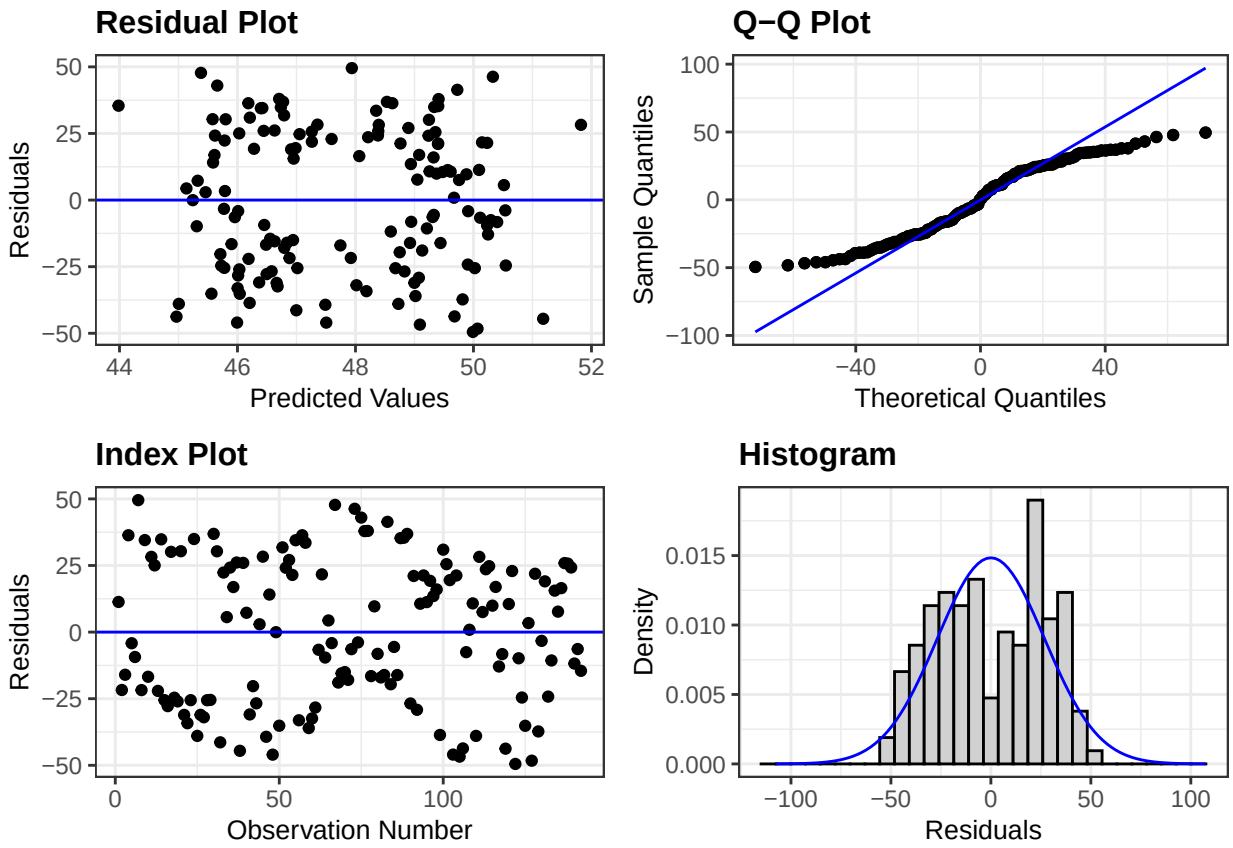
```
# Found this on https://goodekat.github.io/ggResidpanel/

#install.packages("ggResidpanel")
library(ggResidpanel)

# Data 1
resid_panel(lm_data1)
```




```
# Data 2
resid_panel(lm_data2)
```




```
# Data 3
resid_panel(lm_data3)
```



h. Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (2 pts) Without visualizations, it would be very hard to see any relationships and / or correlations within the data. They help identify any trends or outliers that may or may not be present and can help us support our hypothesis or conclusions we are making for any given data.