DATA 606 Fall 2017 - Final Exam

Joshua Sturm

Part I

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

1. B. Quantitative - counts, discrete - whole numbers.
2. A. Since the histogram is left-skewed, we can assume that the median is larger than the mean. Between a, c, and e, a seems the most realistic.
3. A. B is an observational study, and cannot be used to draw causality inferences.
4. D. A large chi-square test means the data is ill-fitted, and there is no relationship.
5. B. We have the formulas $Q_1 - IQR \cdot 1.5$ and $Q_3 + IQR \cdot 1.5$.

```
q1 <- 37
q3 <- 49.8
iqr <- q3 - q1
iqr.15 <- 1.5 * iqr
q1 - iqr.15
```

```
## [1] 17.8
```

```
q3 + iqr.15
```

```
## [1] 69
```

6. D.

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

A - Unimodal, right-skewed, nearly-normal.

B - unimodal, symmetrical, nearly-normal.

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

Since B has many, random, independent samples from A, it will be nearly normal, and have a similar mean. With fewer samples, there is more spread, and samples farther from the mean. The formula for the SD in B is $\frac{sd}{\sqrt{n}}$.

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

This is known as the Central Limit Theorem. If the sample size is greater than 30, each one being independent of the other, and no significant skew, it follows (converges toward) a normal distribution.

Part II

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

```
#options(digits=2)
options(digits=10)
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).

```
#
# These may not be the most efficient way. Also, I would put them into a dataframe, or nicer display ou
#
mean = lapply(
  c(data1, data2, data3, data4),
  function(x)
  {
    y <- mean(x)
    round(y, 2)
  }
)
mean
```

```
## $x
## [1] 9
##
## $y
## [1] 7.5
##
## $x
## [1] 9
##
## $y
## [1] 7.5
##
## $x
## [1] 9
##
## $y
## [1] 7.5
##
## $x
## [1] 9
##
## $y
## [1] 7.5
```

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

```
median = lapply(
  c(data1, data2, data3, data4),
  function(x)
  {
    y <- median(x)
    round(y, 2)
  }
```

```
)
median
```

```
## $x
## [1] 9
##
## $y
## [1] 7.58
##
## $x
## [1] 9
##
## $y
## [1] 8.14
##
## $x
## [1] 9
##
## $y
## [1] 7.11
##
## $x
## [1] 8
##
## $y
## [1] 7.04
```

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

```
sd = lapply(
  c(data1, data2, data3, data4),
  function(x)
  {
    y <- sd(x)
    round(y, 2)
  }
)
sd
```

```
## $x
## [1] 3.32
##
## $y
## [1] 2.03
##
## $x
## [1] 3.32
##
## $y
## [1] 2.03
##
## $x
## [1] 3.32
##
```

```
## $y
## [1] 2.03
##
## $x
## [1] 3.32
##
## $y
## [1] 2.03
```

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

d. The correlation (1 pt).

```
round(cor(data1), 2)
```

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

```
round(cor(data2), 2)
```

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

```
round(cor(data3), 2)
```

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

```
round(cor(data4), 2)
```

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

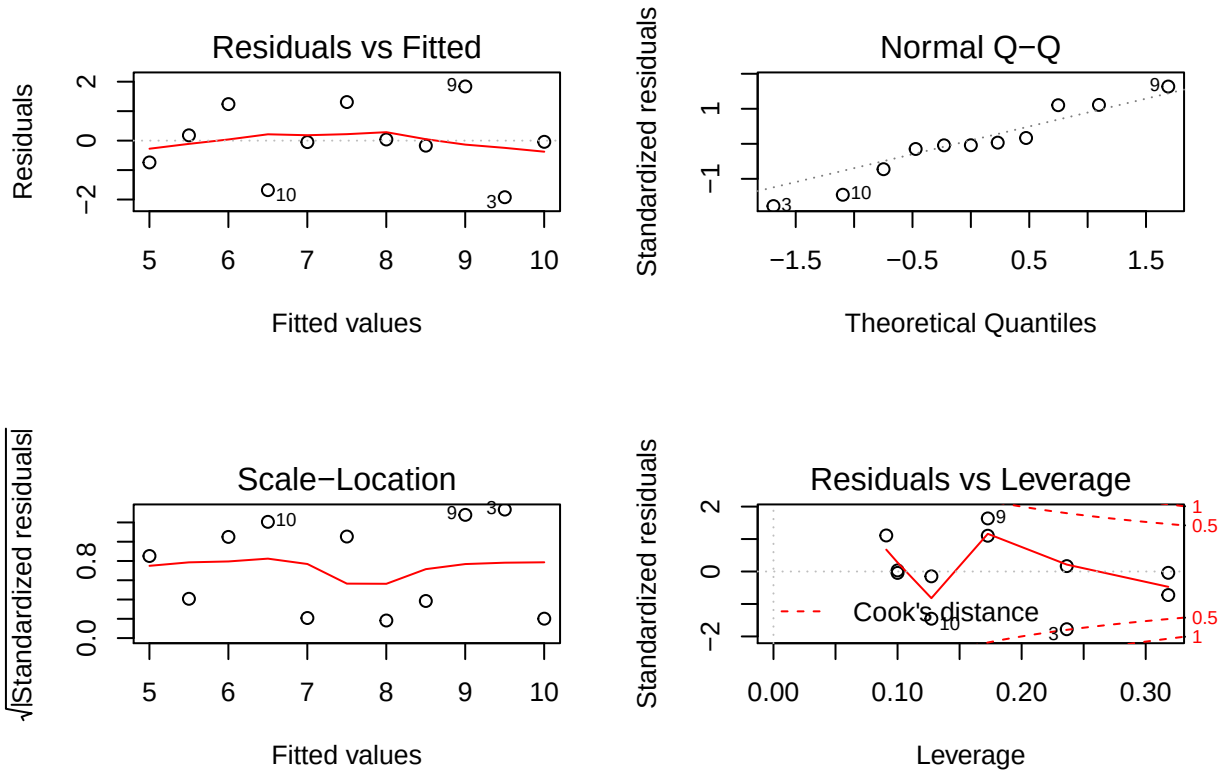
e. Linear regression equation (2 pts).

```
par(mfrow = c(2, 2))
lr1 <- lm(y ~ x, data1)
print(summary(lr1), round(2))
```

```
##
## Call:
## lm(formula = y ~ x, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.921 -0.456 -0.041  0.709  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.00       1.12     2.7    0.026 *
## x               0.50       0.12     4.2    0.002 **
## ---
```

```
## 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:  0.67,   Adjusted R-squared:  0.63
## F-statistic:  18 on 1 and 9 DF,  p-value: 0.0022
```

```
plot(lr1)
```

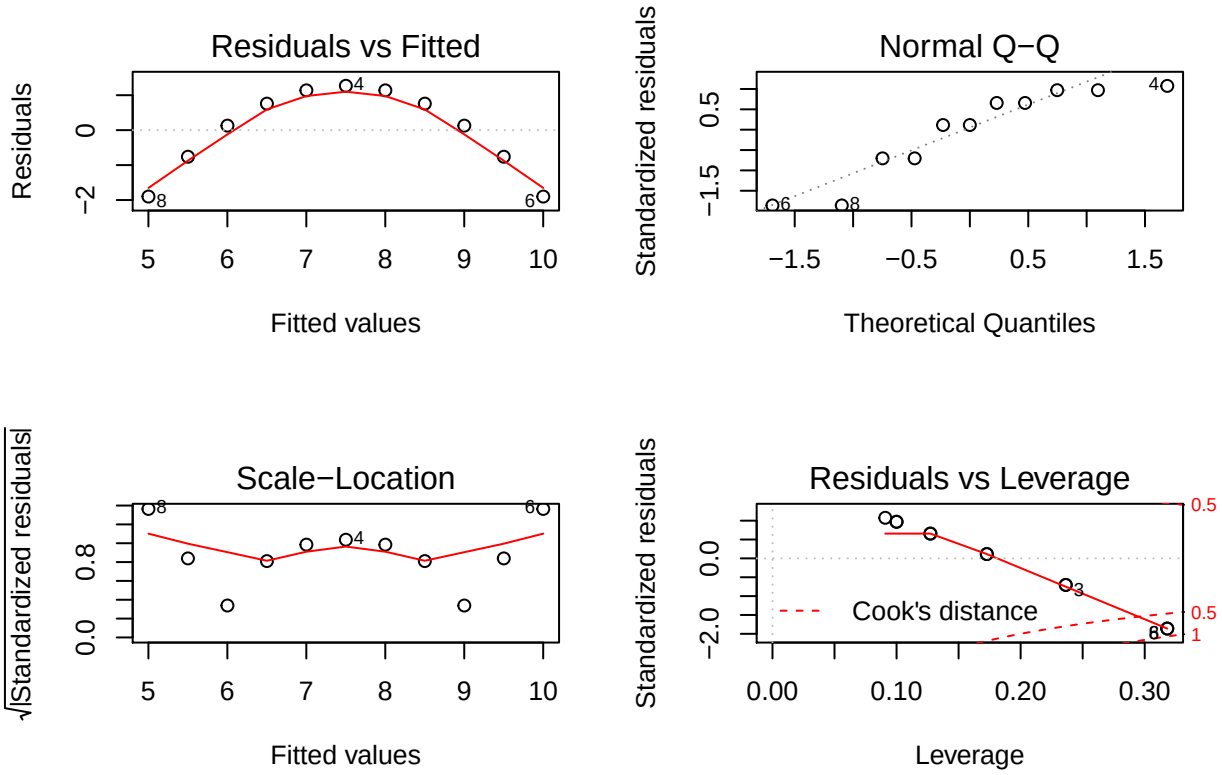



```
lr2 <- lm(y ~ x, data2)
print(summary(lr2), round(2))
```

```
##
## Call:
## lm(formula = y ~ x, data = data2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.90  -0.76   0.13   0.95   1.27
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.00       1.13     2.7    0.026 *
## x               0.50       0.12     4.2    0.002 **
## ---
## 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:  0.67,   Adjusted R-squared:  0.63
## F-statistic:  18 on 1 and 9 DF,  p-value: 0.0022
```

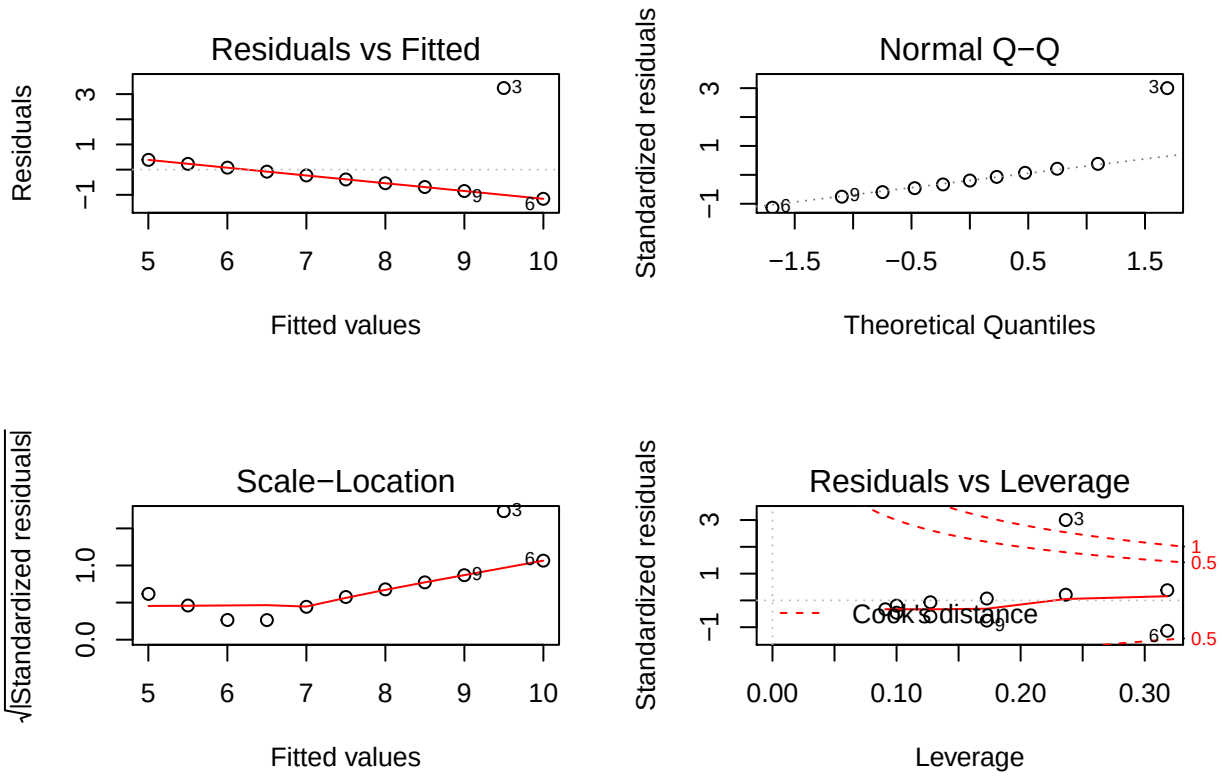
```
plot(lr2)
```




```
lr3 <- lm(y ~ x, data3)
print(summary(lr3), round(2))
```

```
##
## Call:
## lm(formula = y ~ x, data = data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.16  -0.61  -0.23   0.15   3.24
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.00       1.12     2.7    0.026 *
## x               0.50       0.12     4.2    0.002 **
## ---
## 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:  0.67,   Adjusted R-squared:  0.63
## F-statistic:  18 on 1 and 9 DF,  p-value: 0.0022
```

```r
plot(lr3)
```




```r
lr4 <- lm(y ~ x, data4)
print(summary(lr4), round(2))
```

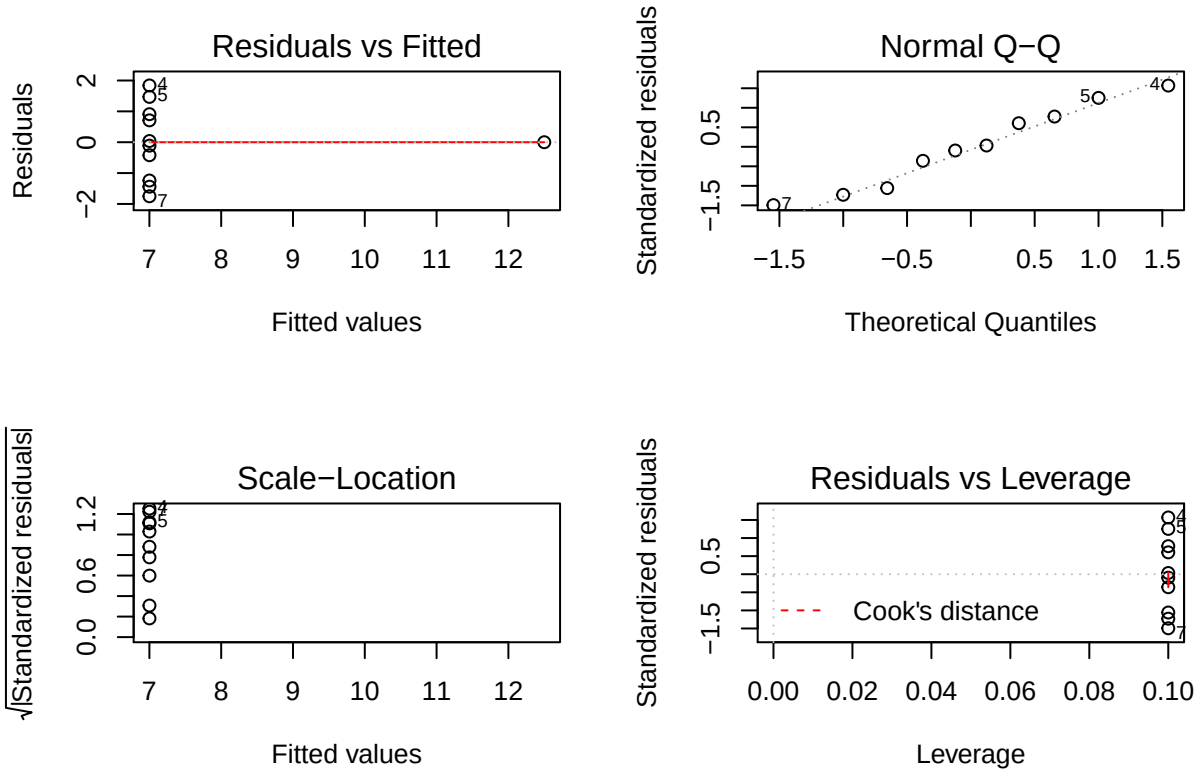
```
##
## Call:
## lm(formula = y ~ x, data = data4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.75  -0.83   0.00   0.81   1.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.00       1.12     2.7    0.026 *
## x               0.50       0.12     4.2    0.002 **
## ---
## 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:  0.67,   Adjusted R-squared:  0.63
## F-statistic:  18 on 1 and 9 DF,  p-value: 0.0022
```

```r
plot(lr4)
```

```
## Warning: not plotting observations with leverage one:
```

```
##   8

## Warning: not plotting observations with leverage one:
##   8
```



f. R-Squared (2 pts).

```
print(summary(lr1)$r.squared, round(2))
```

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

```
print(summary(lr2)$r.squared, round(2))
```

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

```
print(summary(lr3)$r.squared, round(2))
```

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

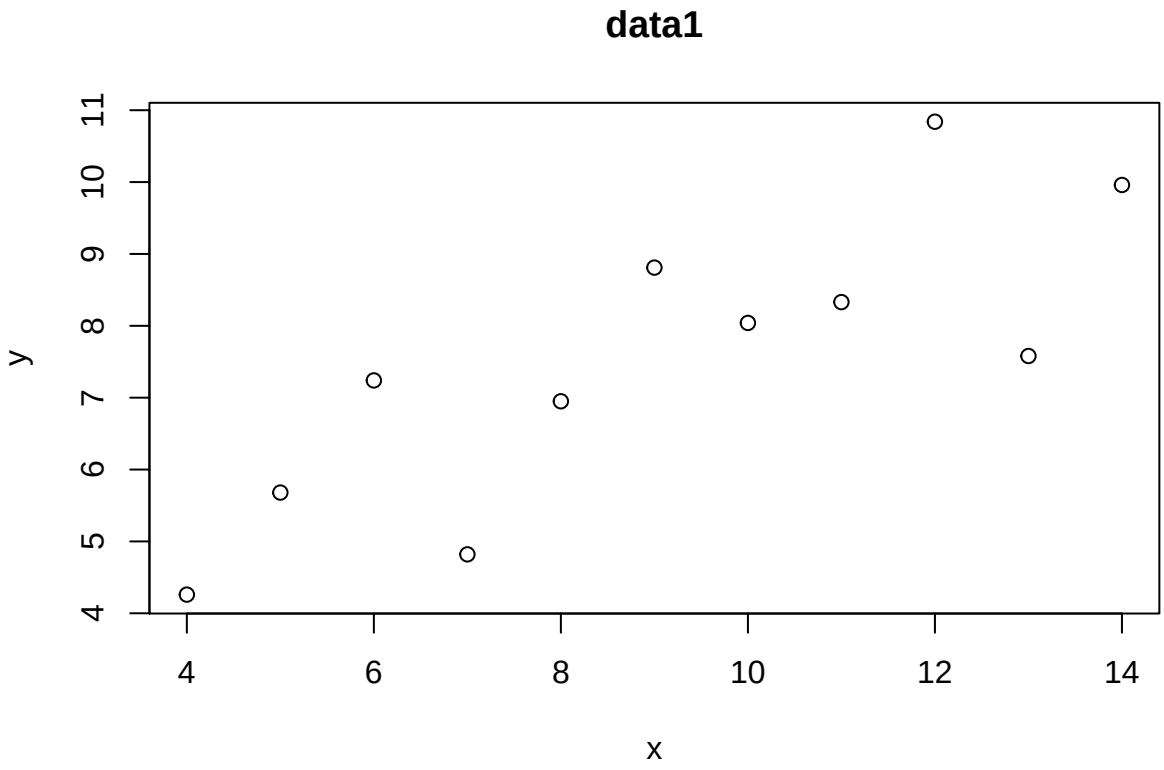
```
print(summary(lr4)$r.squared, round(2))
```

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

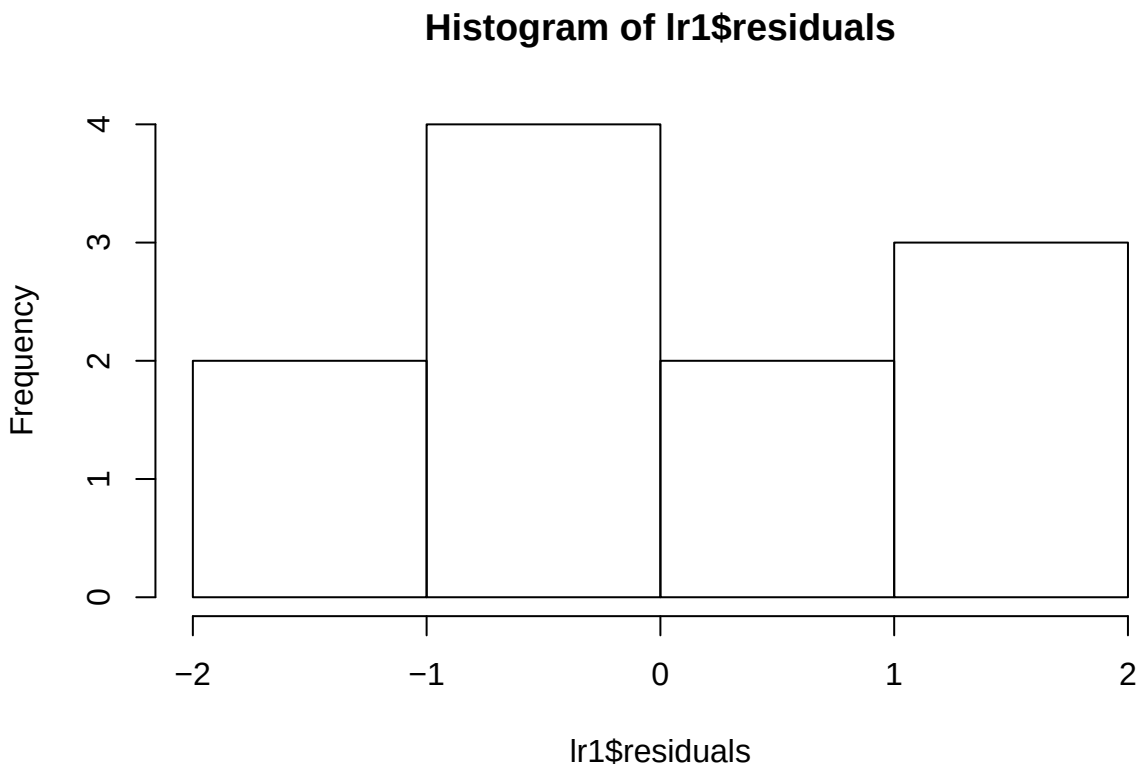
For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

There are four requirements for linear regression: - Linearity - Nearly normal residuals - Constant Variability - Independent observations

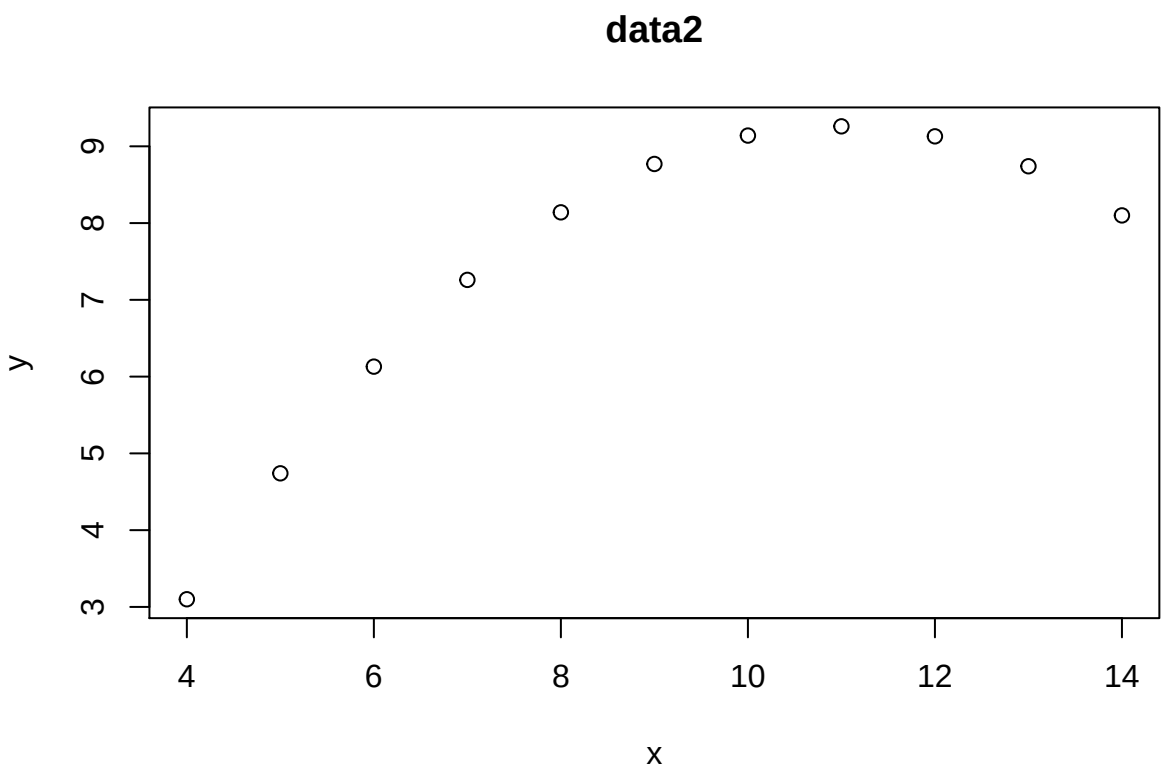
```r
plot(data1, main ='data1')
```



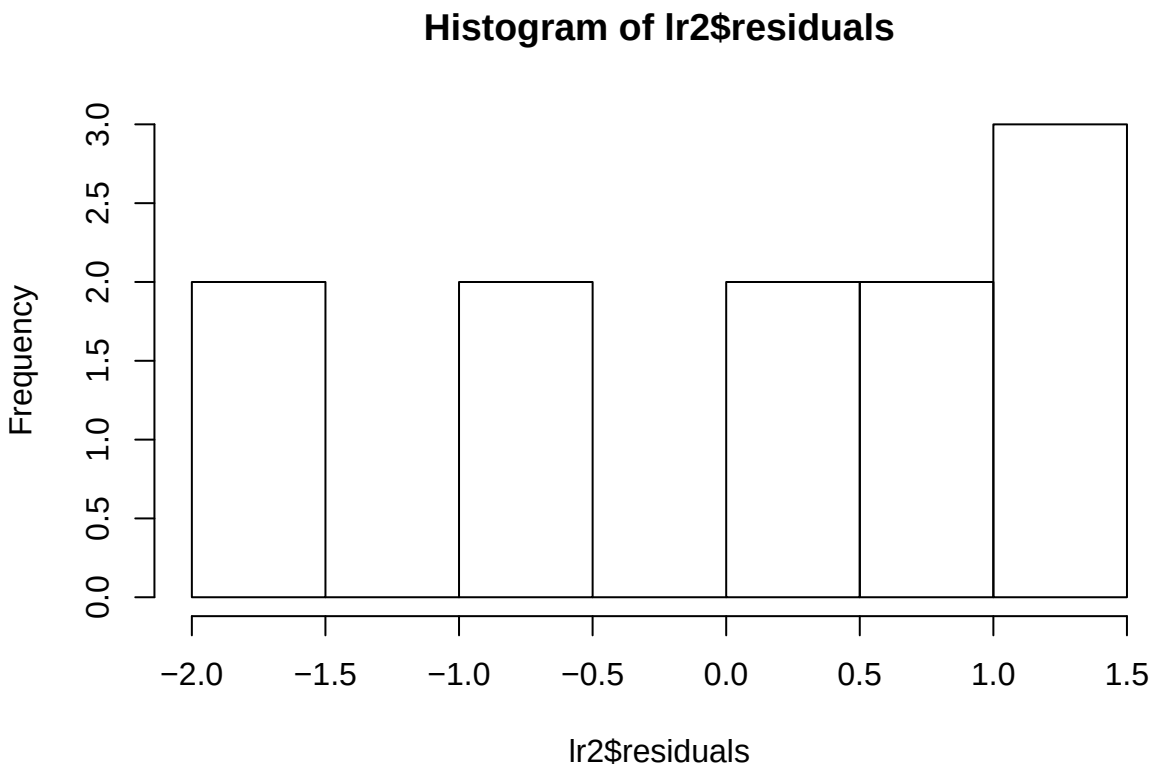

```r
hist(lr1$residuals)
```



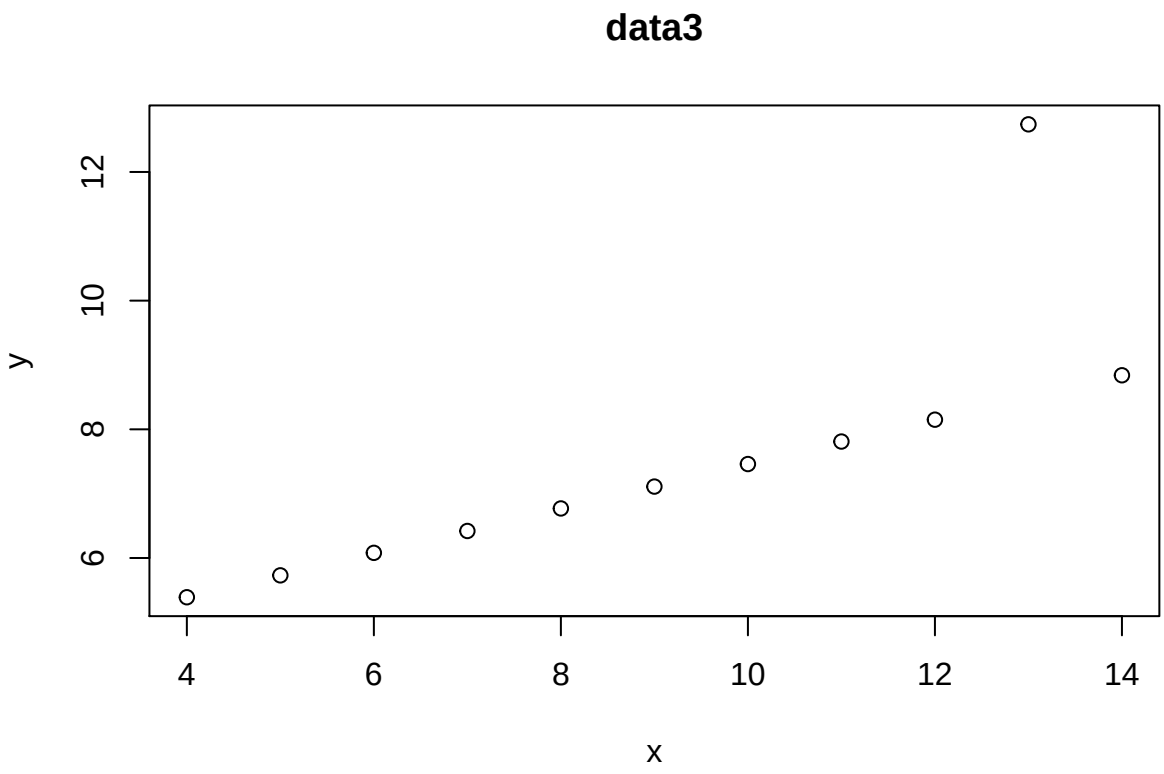

```
plot(data2, main = 'data2')
```



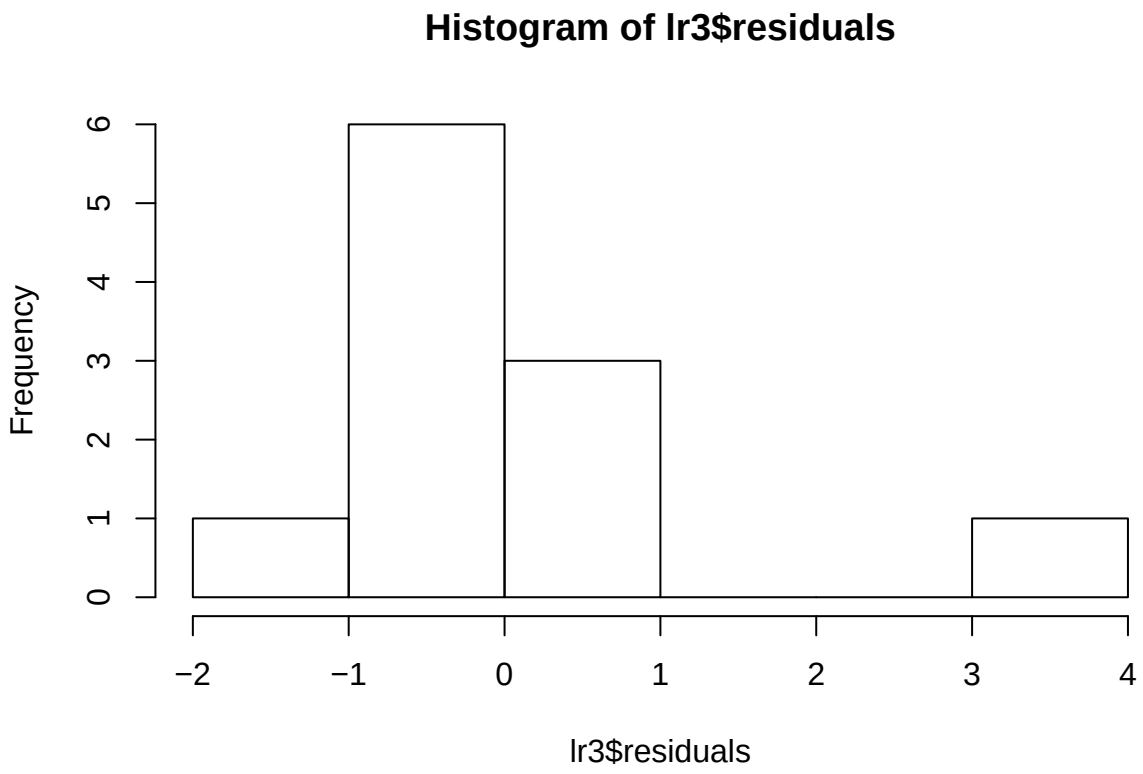

```
hist(lr2$residuals)
```



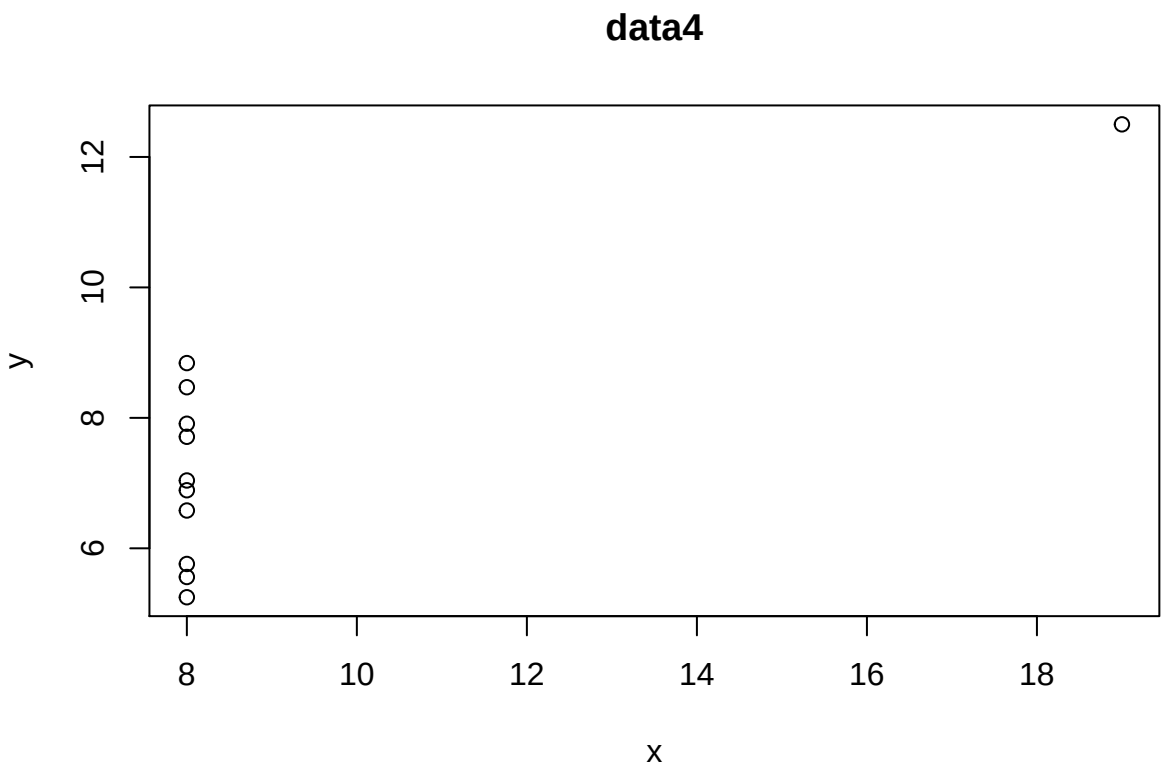

```
plot(data3, main = 'data3')
```



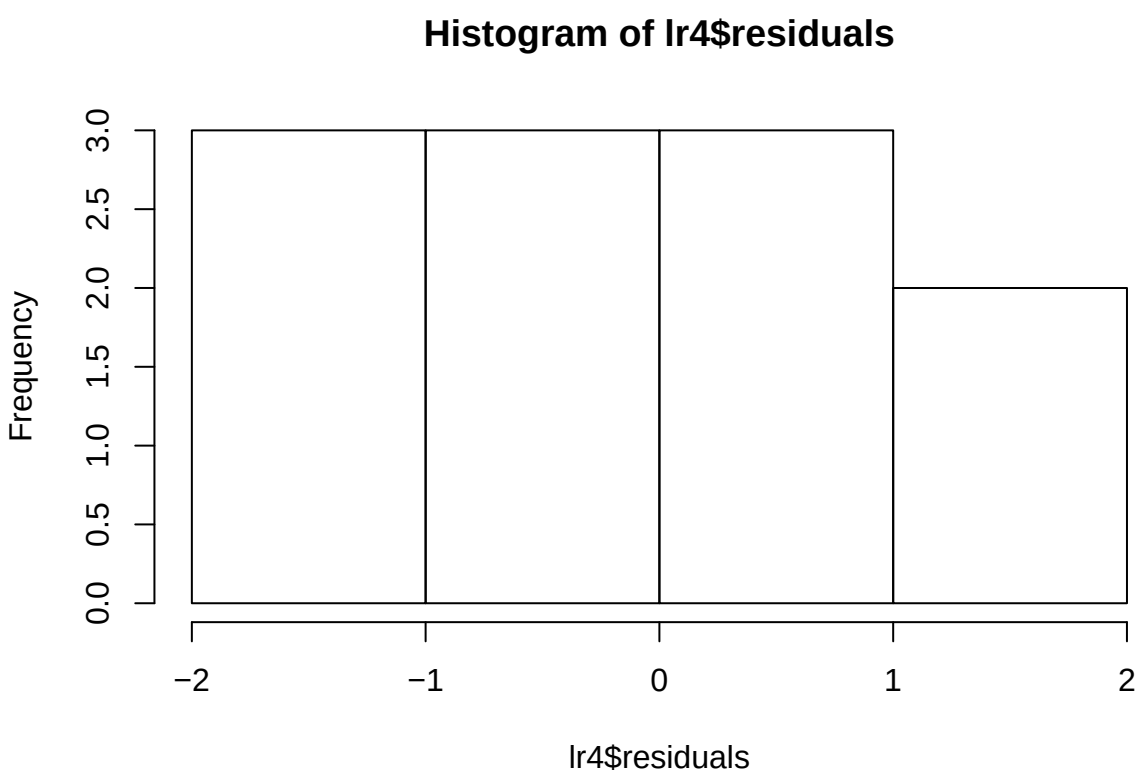

```
hist(lr3$residuals)
```




```
plot(data4, main = 'data4')
```




```
hist(lr4$residuals)
```



- *Dataset 1*: Plot seems normal, but the residuals are not.
- *Dataset 2*: Plot is not linear, and residuals are not normal.
- *Dataset 3*: Seems normal, except for the outlier, which skews the histogram.
- *Dataset 4*: No variability in the plot, and also has an outlier.

If any of the four were apropriate for linear regression, I'd go with three.

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

If you want a quick idea of the subject without delving into the numbers, it's easy to look at a graph. Also, you can spot things you might miss in the numbers, e.g. the outliers. Visualizations are also a good way to summarize a lot of information in a clean, concise, understandable format. Lastly, many people are visual learners. They need graphics to aid their understanding.