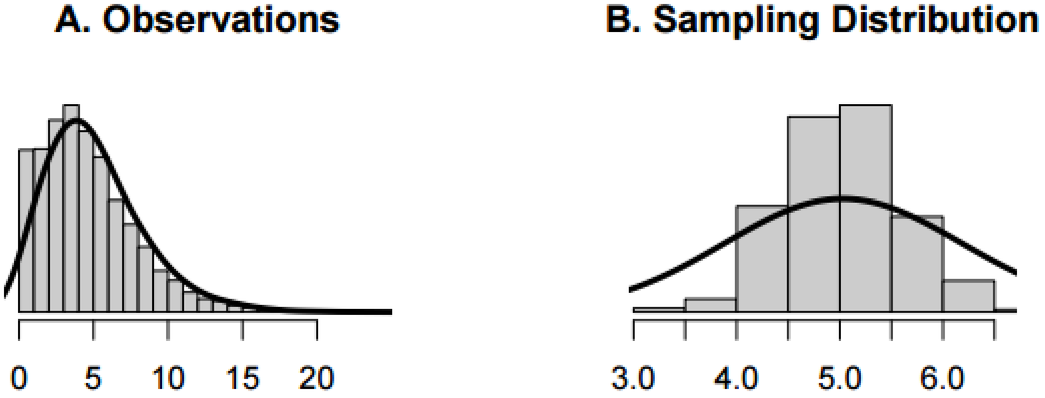
Final Exam

Max Wagner

December 14th, 2015

Part I

The graphs for the section are below:



a.

Observations: This graph is unimodal, asymmetrical, right skewed, and has no outliers represented. The spread is much larger than the sample distribution. The center is at the median, likely slightly below 5.

Sample: This is graph is unimodal, nearly symmetrical, has no skew, and no outliers. The spread is much smaller than the observation graph. The center is near 5.

The graphs above follow the general trend for population and sample mean distributions.

b.

The observations graph includes all observations in the population, which includes some values which will be far from the mean. The sampling distribution represents 500 means of size 30. This means that even with multiple outliers, the mean of the sample will still be near 5. The important difference is that only the means of the 500 samples are plotted on the sampling distribution, not the values within each sample.

c.

The central limit theorem describes the phenomenon above. The gist of the theory is that when random sampling from a population, and the size of samples is large enough, the sample distribution will be roughly normal, even if the population is not. Generally a sample size of 25-40 will be enough to normalize a sample distribution.

Part II

Putting all the x and y values into one frame.

```r
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))
data <- cbind(data1, data2, data3, data4)
colnames(data) <- paste(c("x","y"), c(1,1,2,2,3,3,4,4), sep = "")
kable(data)
```

x1	y1	x2	y2	x3	y3	x4	y4
10	8.0	10	9.1	10	7.5	8	6.6
8	7.0	8	8.1	8	6.8	8	5.8
13	7.6	13	8.7	13	12.7	8	7.7
9	8.8	9	8.8	9	7.1	8	8.8
11	8.3	11	9.3	11	7.8	8	8.5
14	10.0	14	8.1	14	8.8	8	7.0
6	7.2	6	6.1	6	6.1	8	5.2
4	4.3	4	3.1	4	5.4	19	12.5
12	10.8	12	9.1	12	8.2	8	5.6
7	4.8	7	7.3	7	6.4	8	7.9
5	5.7	5	4.7	5	5.7	8	6.9

A. Means

```r
means <- as.data.frame(apply(data, 2, mean))
kable(means)
```

	apply(data, 2, mean)
x1	9.0
y1	7.5
x2	9.0
y2	7.5
x3	9.0
y3	7.5
x4	9.0
y4	7.5

B. Medians

```r
medians <- as.data.frame(apply(data, 2, median))
kable(medians)
```

	apply(data, 2, median)
x1	9.0
y1	7.6
x2	9.0
y2	8.1
x3	9.0
y3	7.1
x4	8.0
y4	7.0

C. Standard Deviation

```r
sd <- as.data.frame(apply(data, 2, sd))
kable(sd)
```

	apply(data, 2, sd)
x1	3.3
y1	2.0
x2	3.3
y2	2.0
x3	3.3
y3	2.0
x4	3.3
y4	2.0

D. Correlation

With two decimal places the correlations appear to be the same, however they are all slightly different.

```r
cor1 <- cor(data$x1, data$y1)
cor2 <- cor(data$x2, data$y2)
cor3 <- cor(data$x3, data$y3)
cor4 <- cor(data$x4, data$y4)
kable(as.data.frame(rbind(cor1, cor2, cor3, cor4)))
```

	V1
cor1	0.82
cor2	0.82
cor3	0.82
cor4	0.82

E. Linear Regression Equations

```r
lr1 <- lm(data$y1 ~ data$x1)
lr2 <- lm(data$y2 ~ data$x2)
lr3 <- lm(data$y3 ~ data$x3)
lr4 <- lm(data$y4 ~ data$x4)
```

1:

$$\hat{y} = 3 + 0.5x$$

2:

$$\hat{y} = 3 + 0.5x$$

3:

$$\hat{y} = 3 + 0.5x$$

4:

$$\hat{y} = 3 + 0.5x$$

F. R-Squared

```r
lrs1 <- summary(lr1)
lrs2 <- summary(lr2)
lrs3 <- summary(lr3)
lrs4 <- summary(lr4)
r_squared <- c(lrs1$r.squared, lrs2$r.squared, lrs3$r.squared, lrs4$r.squared)
r_squared <- cbind(c(1:4), r_squared)
kable(as.data.frame(r_squared))
```

V1	r_squared
1	0.67
2	0.67
3	0.67
4	0.67

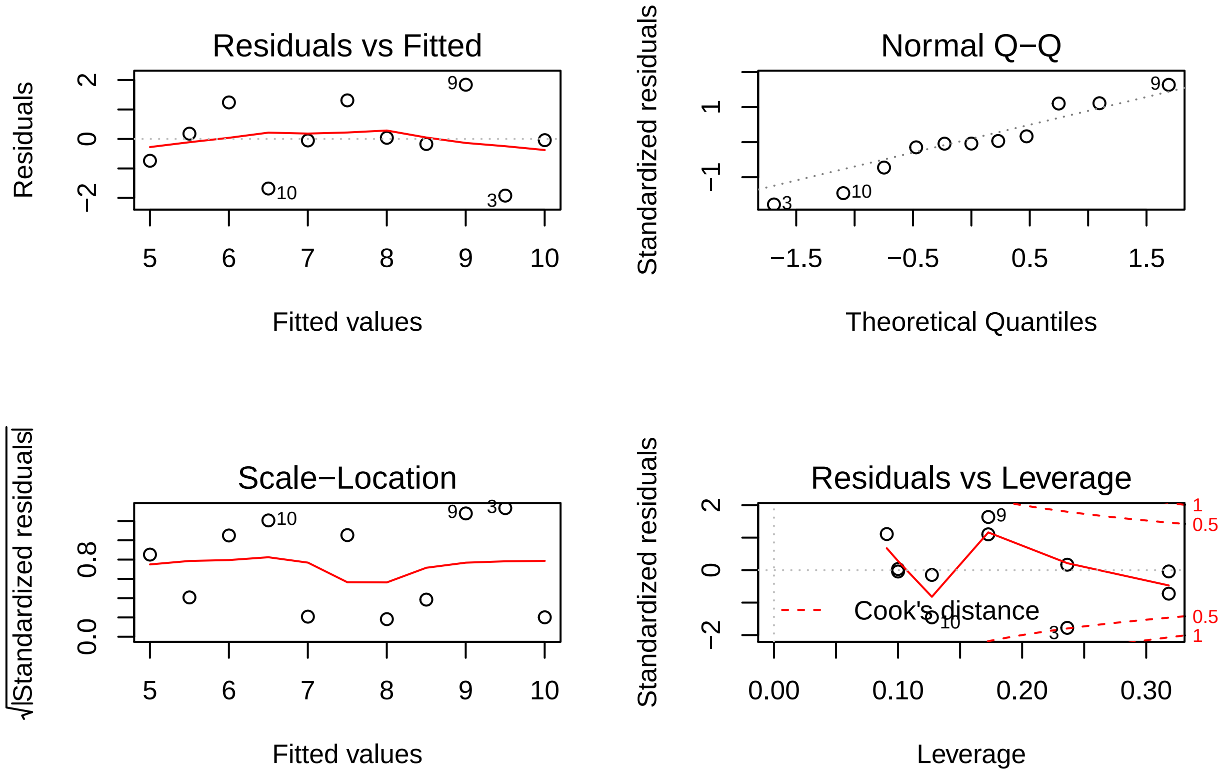
G. For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be as specific as to why for each pair!

The assumptions for using a linear regression are:

- linear relationship between x and y
- independent data points
- residuals show no patterns
- homogeneity of variance

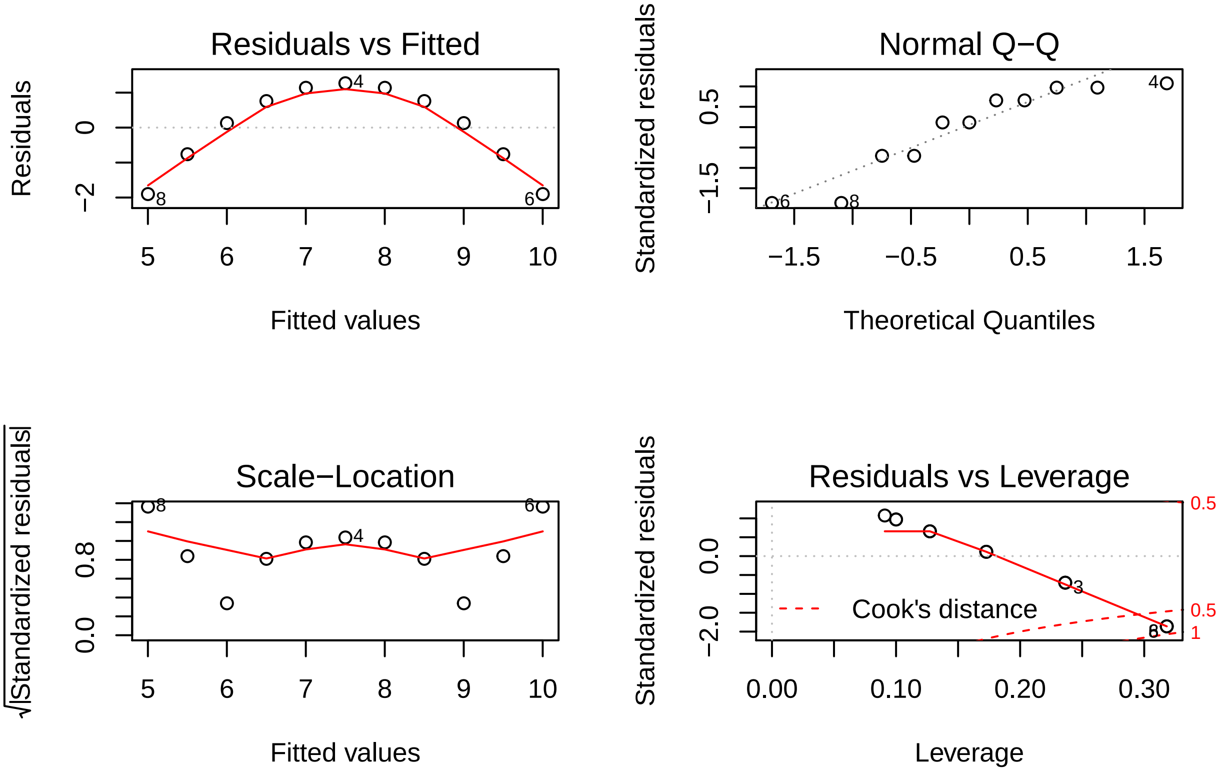
data1: Independence is assumed. The top left graph indicates that data1 is linear and variance is homegenous because of the random nature. The top right graph shows a slightly curvely line. This means that while not perfect, the residual check passes.

```r
par(mfrow = c(2, 2))
ab1 <- lm(data$y1 ~ data$x1)
plot(ab1)
```



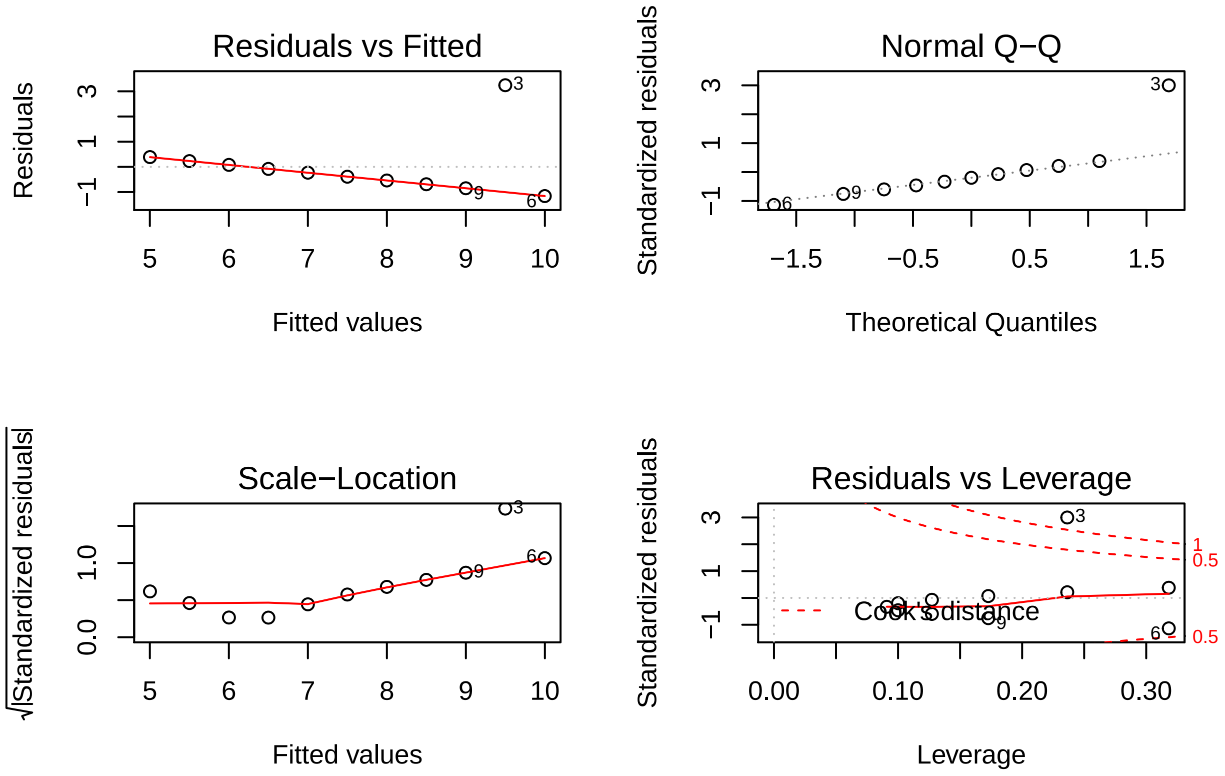
data2: Independence is assumed. The top left graph shows that the relationship is not linear. The check fails, other tests aren't needed.

```r
par(mfrow = c(2, 2))
ab2 <- lm(data$y2 ~ data$x2)
plot(ab2)
```



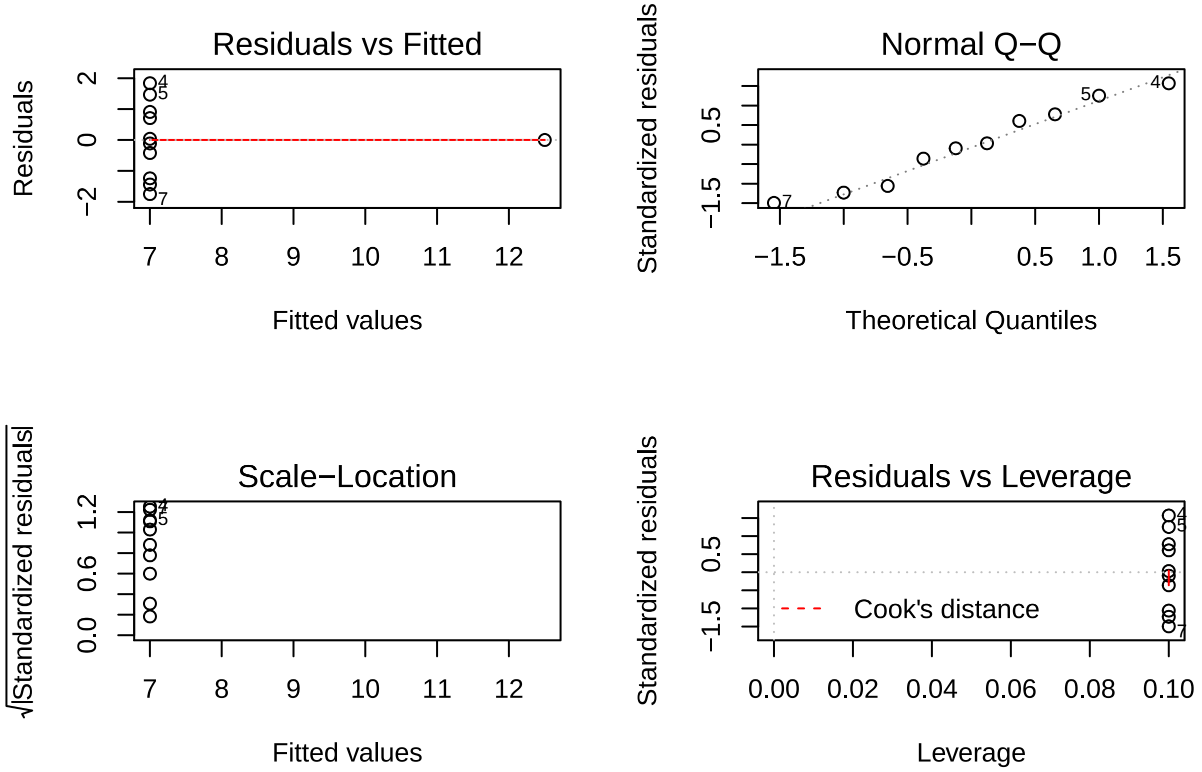
data3: Independence is assumed. The top left plot shows a pattern, so the check fails due to variance not being homogeneous.

```r
par(mfrow = c(2, 2))
ab3 <- lm(data$y3 ~ data$x3)
plot(ab3)
```



data4: Independence is assumed. The top left graph is not random, so the check fails due to variance not being homogeneous, and not being linear.

```r
par(mfrow = c(2, 2))
ab4 <- lm(data$y4 ~ data$x4)
plot(ab4)
```



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

In the examples from above, it is possible to instantly tell whether or not they pass linear regression assumptions. With purely numbers, it is still possible to check, but not as clearly.

More generally, visualizations help put the numbers into perspective. It can mean the difference between finding a new trend or the data remaining obscure. Visualizations are also essential in explaining findings to others who have not intimately worked with the data.