DATA 606 Fall 2016 - Final Exam

Ahsanul Choudhury

December 9, 2016

Part I

Figure A below represents the distribution of an observed variable. Figure B below represents the distribution of the mean from 500 random samples of size 30 from A. The mean of A is 5.05 and the mean of B is 5.04. The standard deviations of A and B are 3.22 and 0.58, respectively.

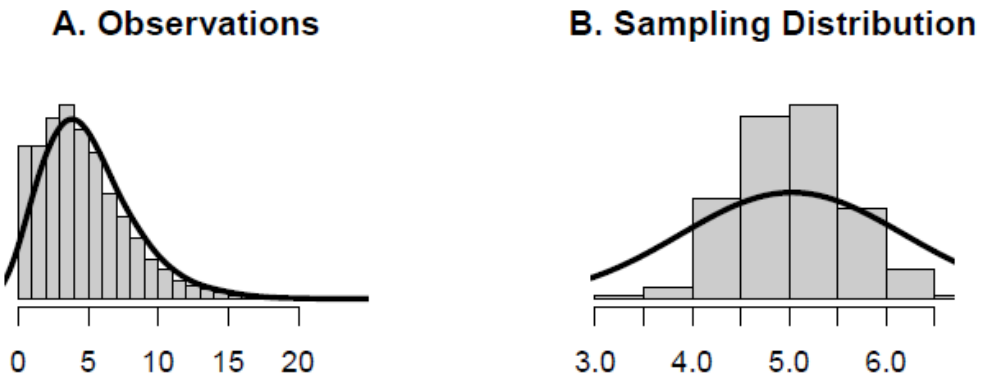


Figure 1:

a. Describe the two distributions (2 pts).

The distribution in figure A (observed variable) is unimodal, highly skewed to the right and ranges from 0 to 20. The distribution in figure B (mean of random samples) is also unimodal but near normal and centered at the true mean of the population.

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

Figure B represents the distribution of the mean from 500 random samples of size 30 from A (population) therefore the mean approximates the true population mean. The standard deviation of sample means is always smaller than population mean and can be obtained by the following formula:

$$SE = \frac{\sigma}{\sqrt{n}}$$

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

The statistical principal that describes this phenomenon is **central limit theorem**. Central limit theorem states if the sample size is large enough (30+), the sampling distribution of any independent, random variable will be normal or nearly normal and the sample mean will approximate the population mean.

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).

Mean for each column

```
df1 <- data.frame(x = c(mean(data1$x), mean(data2$x), mean(data3$x), mean(data4$x)),
                  y = c(mean(data1$y), mean(data2$y), mean(data3$y), mean(data4$y)),
                  row.names = c("data1", "data2", "data3", "data4"))
knitr::kable(format(df1, digits=2, nsmall=2))
```

	x	y
data1	9.00	7.50
data2	9.00	7.50
data3	9.00	7.50
data4	9.00	7.50

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

Median for each column

```
df2 <- data.frame(x = c(median(data1$x), median(data2$x), median(data3$x), median(data4$x)),
                  y = c(median(data1$y), median(data2$y), median(data3$y), median(data4$y)),
                  row.names = c("data1", "data2", "data3", "data4"))
knitr::kable(format(df2, digits=2, nsmall=2))
```

	x	y
data1	9.00	7.58
data2	9.00	8.14
data3	9.00	7.11
data4	8.00	7.04

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

Standard deviation for each column

```
df3 <- data.frame(x = c(sd(data1$x), sd(data2$x), sd(data3$x), sd(data4$x)),
                  y = c(sd(data1$y), sd(data2$y), sd(data3$y), sd(data4$y)),
                  row.names = c("data1", "data2", "data3", "data4"))
knitr::kable(format(df3, digits=2, nsmall=2))
```

	x	y
data1	3.32	2.03
data2	3.32	2.03
data3	3.32	2.03
data4	3.32	2.03

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

d. The correlation (1 pt).

Correlation

```
df4 <- data.frame(data1 = cor(data1$x, data1$y),
                  data2 = cor(data2$x, data2$y),
                  data3 = cor(data3$x, data3$y),
                  data4 = cor(data4$x, data4$y),
                  row.names = c("correlation"))

knitr::kable(format(df4, digits=2, nsmall=2))
```

	data1	data2	data3	data4
correlation	0.82	0.82	0.82	0.82

e. Linear regression equation (2 pts).

```
m1 <- lm(y ~ x, data = data1)
m2 <- lm(y ~ x, data = data2)
m3 <- lm(y ~ x, data = data3)
m4 <- lm(y ~ x, data = data4)
```

equation:

$$\hat{y} = \beta_0 + \beta_1 * x$$

data1:

$$\hat{y} = 3.00 + 0.50 * x$$

data2:

$$\hat{y} = 3.00 + 0.50 * x$$

data3:

$$\hat{y} = 3.00 + 0.50 * x$$

data4:

$$\hat{y} = 3.00 + 0.50 * x$$

f. R-Squared (2 pts).

data1 $r^2 = 0.67$

data2 $r^2 = 0.67$

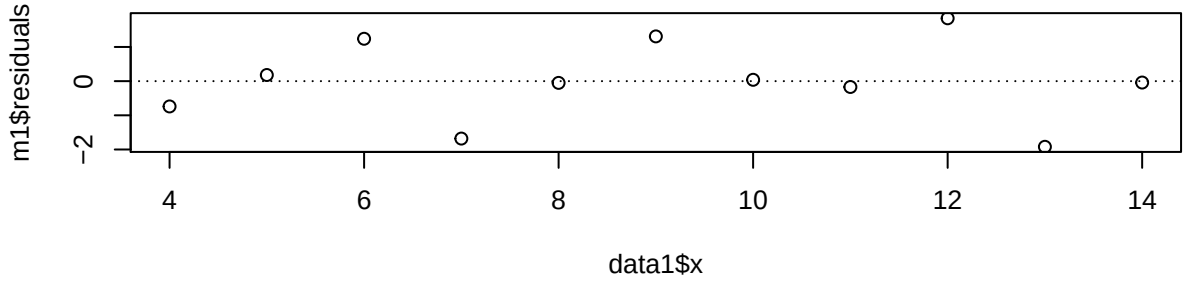
data3 $r^2 = 0.67$

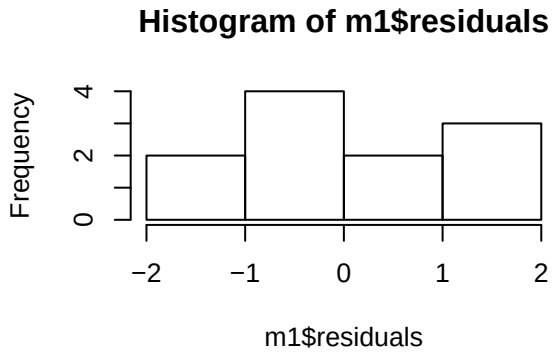
data4 $r^2 = 0.67$

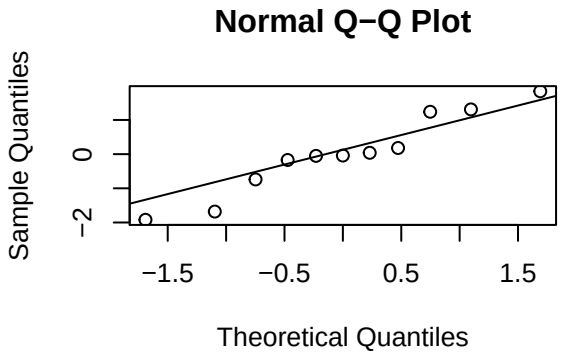
For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

data1 visualizations:

```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))

plot(data1$x, m1$residuals)
abline(h = 0, lty = 3)

hist(m1$residuals)

qqnorm(m1$residuals)
qqline(m1$residuals)
```





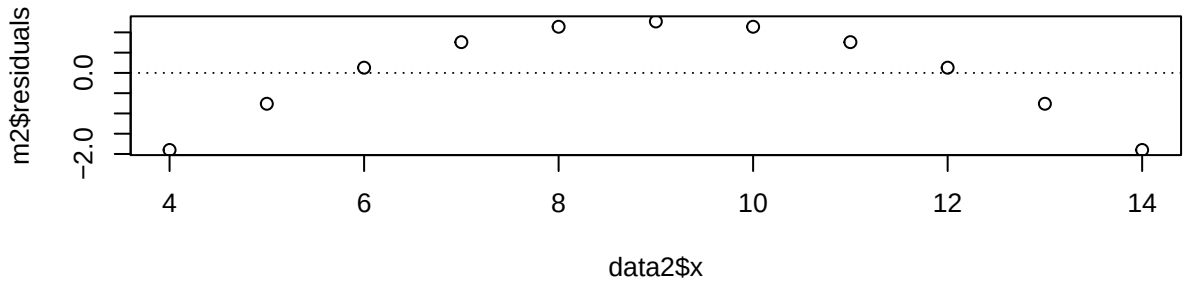


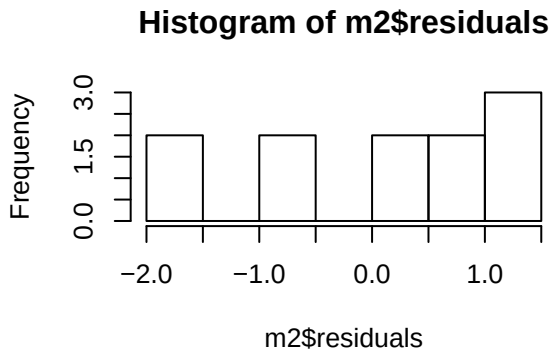
data1 evaluation

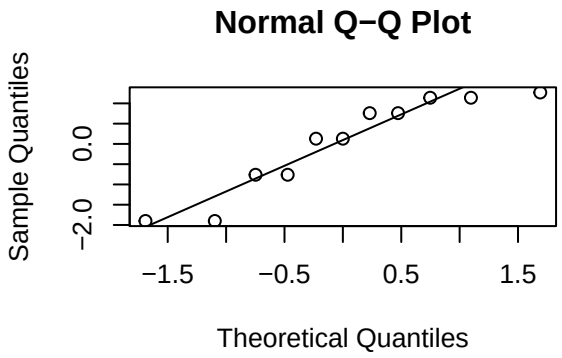
There is no apparent pattern in the residual scatterplot, the histogram shows hint of normality and the Q-Q plot shows near constant variability. In the case of data1 linear regression model is appropriate.

data2 visualizations:

```r
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))

plot(data2$x, m2$residuals)
abline(h = 0, lty = 3)

hist(m2$residuals)

qqnorm(m2$residuals)
qqline(m2$residuals)
```





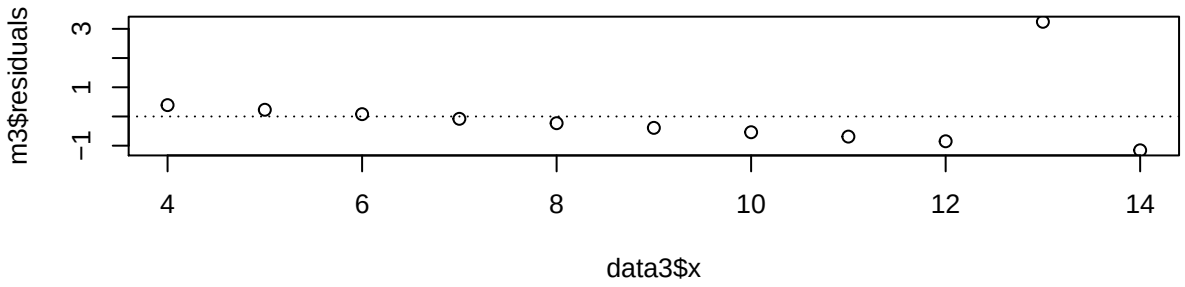


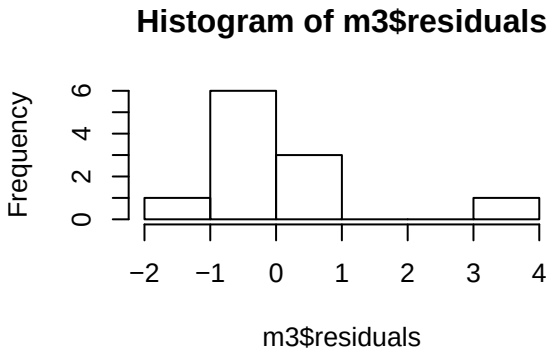
data2 evaluation

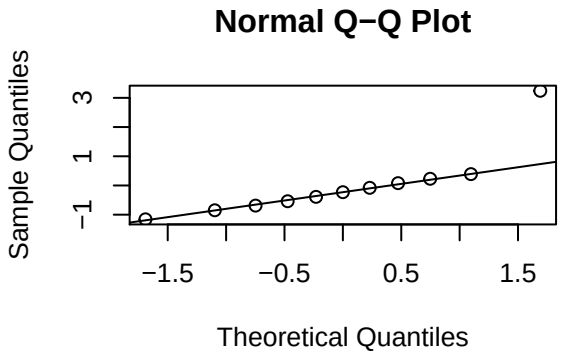
There seem to be a pattern in the residual scatterplot, it looks like a strong non linear relationship. The histogram shows no normality. In the case of data2 linear regression model is not appropriate.

data3 visualizations:

```r
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))

plot(data3$x, m3$residuals)
abline(h = 0, lty = 3)

hist(m3$residuals)

qqnorm(m3$residuals)
qqline(m3$residuals)
```





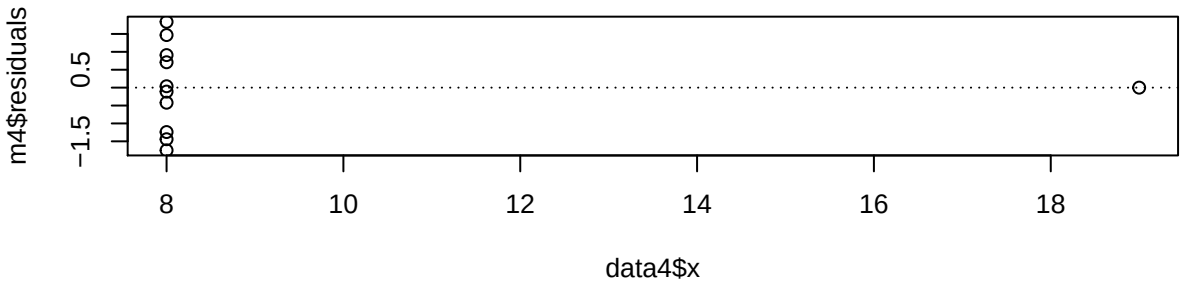


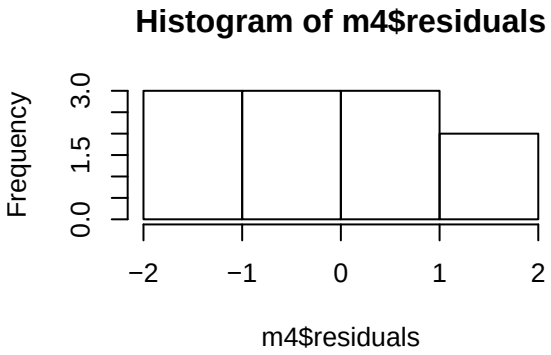
data3 evaluation

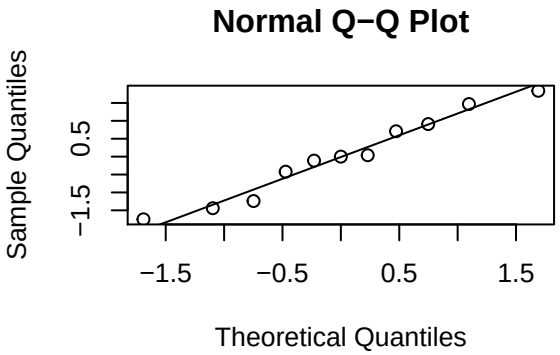
From the scatter plot we can see the varibility is not constant and there is an extreme outlier, for data3 regressionn model is not appropriate.

data4 visualizations:

```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))

plot(data4$x, m4$residuals)
abline(h = 0, lty = 3)

hist(m4$residuals)

qqnorm(m4$residuals)
qqline(m4$residuals)
```





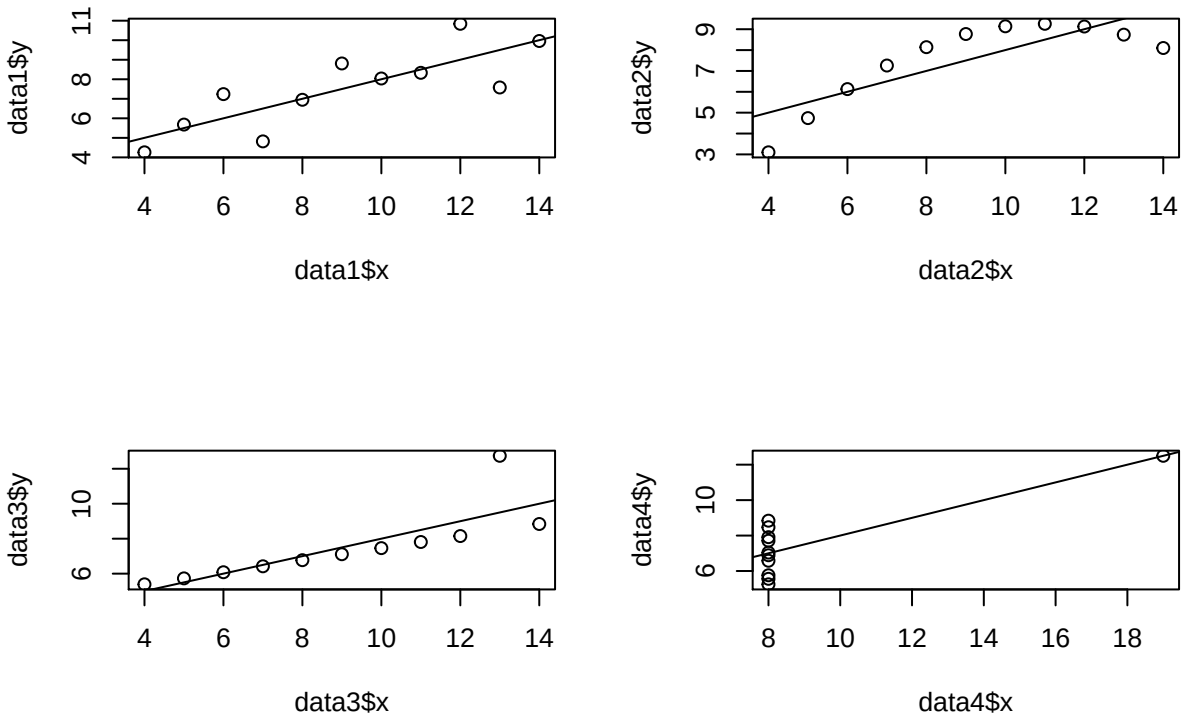


data4 evaluation

Apart from an extreme oulier there is clear pattern in the residual scatterplot, therefore the observations might not be independent. In the case of data4 linear regression model is not appropriate.

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

```
par(mfrow=c(2,2))

plot(data1$y ~ data1$x)
abline(m1)

plot(data2$y ~ data2$x)
abline(m2)

plot(data3$y ~ data3$x)
abline(m3)


plot(data4$y ~ data4$x)
abline(m4)
```



The four data sets have seemingly identical summary statistics apart from the median of y, if we take two decimal point all four data sets have same mean, corrolation and even r^2 value. When we plot the data than the difference becomes apparent, visualization helps us identify linearity, distribution, constant variability and independence of data, which can be observerd from the plots above.