DA 606 Final Exam

Dan Fanelli

May 14, 2016

Part 1

Considering the given information:

[a] Describe the two distributions.

These 2 distributions have basically the same mean. They do not have similar standard deviations. A is skewed right, while B does not have such a skew. As stated in the question, A is the distribution of an observed variable, and B is the distribution of the mean of that distribution A.

[b] Explain why the means of these two distributions are similar but the standard deviations are not.

The means are similar because of the Central Limit Theorum (stated below). The Standard Deviations are not similar because they are standard deviations of 2 different statistics: **A** is the standard deviation of the values of the random sample, while **B** is the standard deviation of the MEAN of those values given 500 random samples of sample size 30.

[c] What is the statistical principal that describes this phenomenon:

The **Central Limit Theorem (for normal data)** is the statistical principal that describes this phenomenon:

"The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size."

Part 2

The Input:

```
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,

data2 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5), y=c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.2

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,

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,
```

Take a quick look at them:

```
data1
```

```
##     x    y
## 1  10  8.0
## 2   8  7.0
## 3  13  7.6
## 4   9  8.8
## 5  11  8.3
## 6  14 10.0
## 7   6  7.2
## 8   4  4.3
## 9  12 10.8
## 10  7  4.8
## 11  5  5.7
```

```
data2
```

```
##     x   y
## 1  10 9.1
## 2   8 8.1
## 3  13 8.7
## 4   9 8.8
## 5  11 9.3
## 6  14 8.1
## 7   6 6.1
## 8   4 3.1
## 9  12 9.1
## 10  7 7.3
## 11  5 4.7
```

```
data3
```

```
##     x    y
## 1  10  7.5
## 2   8  6.8
## 3  13 12.7
## 4   9  7.1
## 5  11  7.8
## 6  14  8.8
## 7   6  6.1
## 8   4  5.4
## 9  12  8.2
## 10  7  6.4
## 11  5  5.7
```

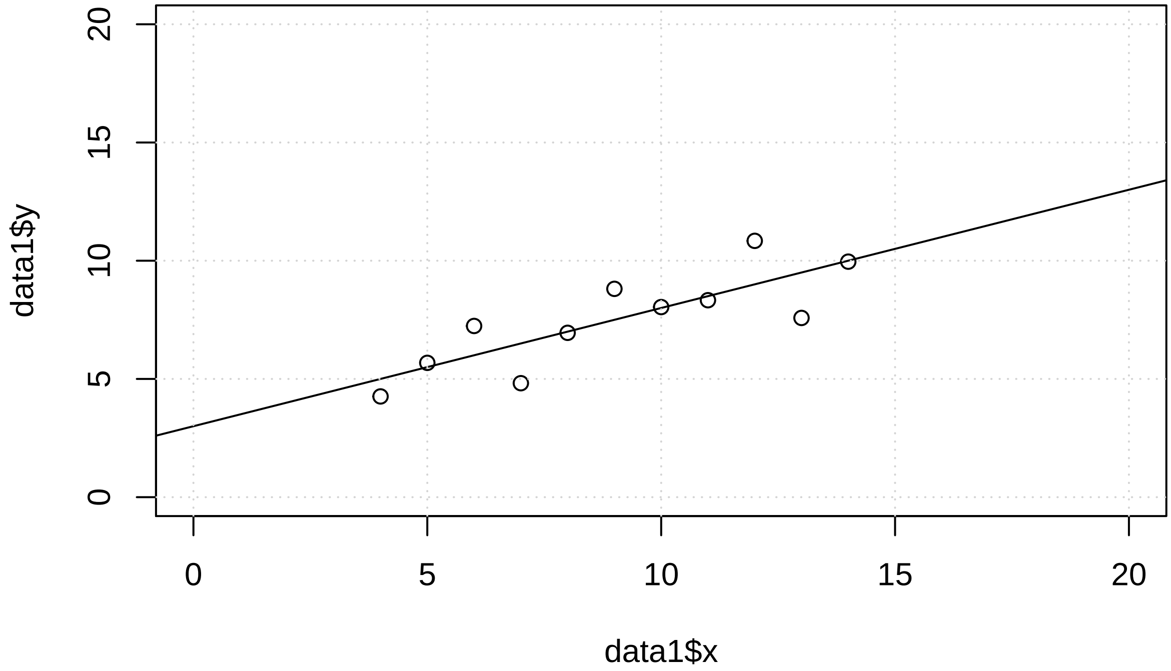
```
data4
```

```
##     x    y
## 1   8  6.6
## 2   8  5.8
## 3   8  7.7
## 4   8  8.8
## 5   8  8.5
## 6   8  7.0
## 7   8  5.2
## 8  19 12.5
## 9   8  5.6
## 10  8  7.9
## 11  8  6.9
```

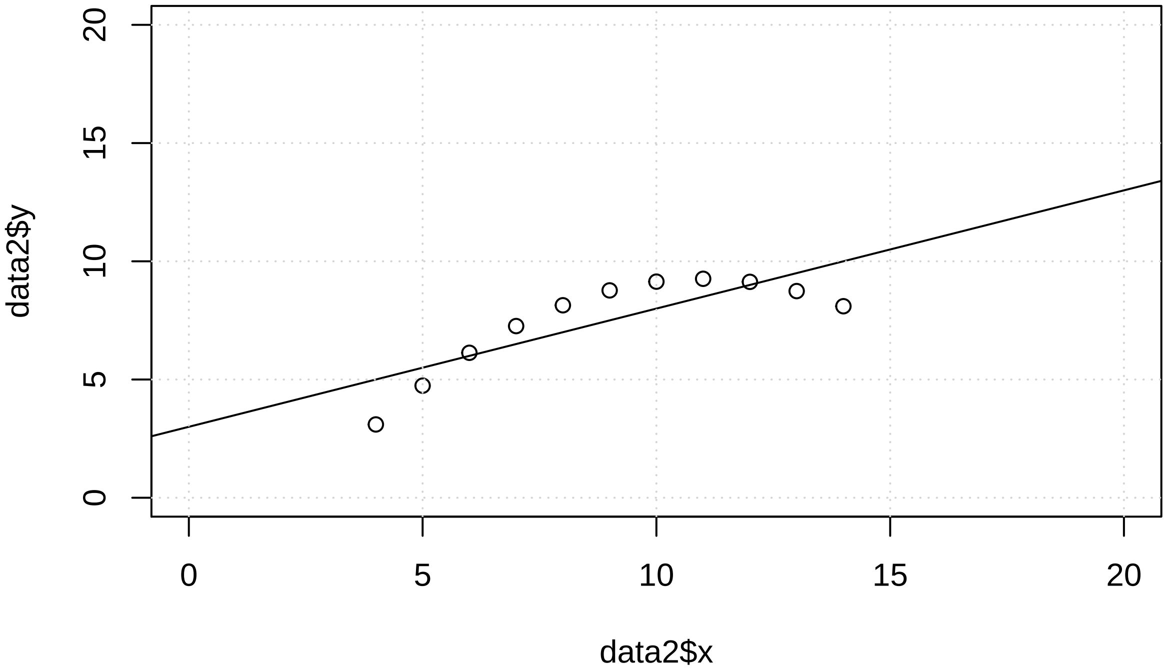
```
# confirm mins and maxes and plot them all with the same bounds for comparison:
max(data1$x,data2$x,data3$x,data4$x,data1$y,data2$y,data3$y,data4$y)
```

```
## [1] 19
```

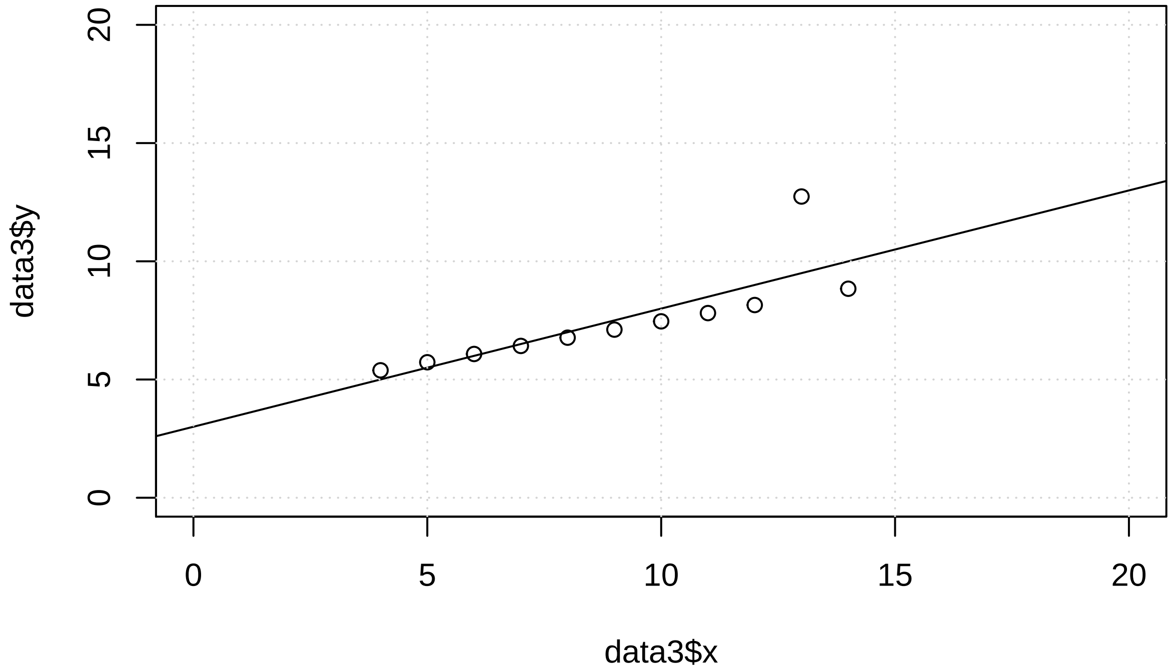
```
# above max was 19, so we'll go with 20x20
lims <- c(0, 20)

plot(data1$x, data1$y, xlim=lims, ylim=lims)
fit1 <- lm(data1$y ~ data1$x)
abline(fit1)
grid(NULL, NULL)
```



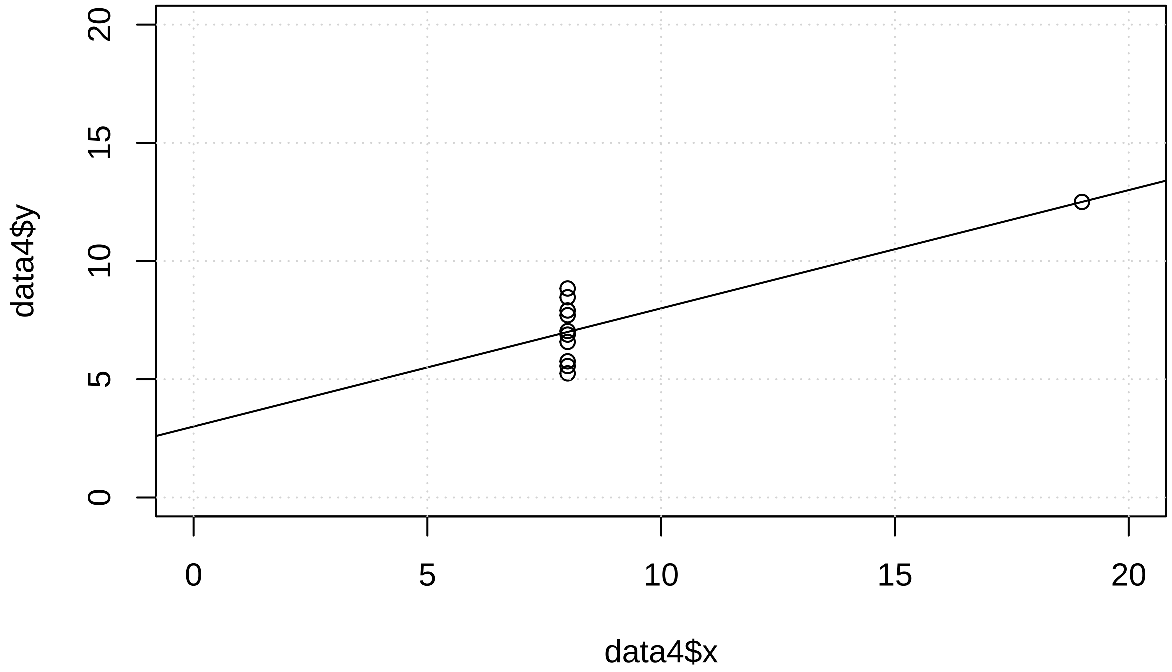
```
plot(data2$x, data2$y, xlim=lims, ylim=lims)
fit2 <- lm(data2$y ~ data2$x)
abline(fit2)
grid(NULL, NULL)
```



```r
plot(data3$x, data3$y, xlim=lims, ylim=lims)
fit3 <- lm(data3$y ~ data3$x)
abline(fit3)
grid(NULL, NULL)
```



```
plot(data4$x, data4$y, xlim=lims, ylim=lims)
fit4 <- lm(data4$y ~ data4$x)
abline(fit4)
grid(NULL, NULL)
```



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

```
mean(data1$x)
```

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

```
## [1] 9
```

```
mean(data1$y)
```

```
## [1] 7.5
```

```
mean(data2$x)
```

```
## [1] 9
```

```
mean(data2$y)
```

```
## [1] 7.5
```

```
mean(data3$x)
```

```
## [1] 9
```

```
mean(data3$y)
```

```
## [1] 7.5
```

```
mean(data4$x)
```

```
## [1] 9
```

```
mean(data4$y)
```

```
## [1] 7.5
```

```


median(data1$x)
```

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

```
## [1] 9
```

```
median(data1$y)
```

```
## [1] 7.6
```

```
median(data2$x)
```

```
## [1] 9
```

```
median(data2$y)
```

```
## [1] 8.1
```

```
median(data3$x)
```

```
## [1] 9
```

```
median(data3$y)
```

```
## [1] 7.1
```

```
median(data4$x)
```

```
## [1] 8
```

```
median(data4$y)
```

```
## [1] 7
```

```


sd(data1$x)
```

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

```
## [1] 3.3
```

```
sd(data1$y)
```

```
## [1] 2
```

```
sd(data2$x)
```

```
## [1] 3.3
```

```
sd(data2$y)
```

```
## [1] 2
```

```
sd(data3$x)
```

```
## [1] 3.3
```

```
sd(data3$y)
```

```
## [1] 2
```

```
sd(data4$x)
```

```
## [1] 3.3
```

```
sd(data4$y)
```

```
## [1] 2
```

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

```
cor(data1$x, data1$y)
```

d. The correlation (1 pt).

```
## [1] 0.82
```

```
cor(data2$x, data2$y)
```

```
## [1] 0.82
```

```
cor(data3$x, data3$y)
```

```
## [1] 0.82
```

```
cor(data4$x, data4$y)
```

```
## [1] 0.82
```

e. Linear regression equation (2 pts). Equation 1: y = 3 + 0.5*x

```
fit1
```

```
##
## Call:
## lm(formula = data1$y ~ data1$x)
##
## Coefficients:
## (Intercept)      data1$x
##         3.0          0.5
```

Equation 2: y = 3 + 0.5*x

```
fit2
```

```
##
## Call:
## lm(formula = data2$y ~ data2$x)
##
## Coefficients:
## (Intercept)      data2$x
##         3.0          0.5
```

Equation 3: y = 3 + 0.5*x

```
fit3
```

```
##
## Call:
## lm(formula = data3$y ~ data3$x)
##
## Coefficients:
## (Intercept)      data3$x
##         3.0          0.5
```

Equation 4: y = 3 + 0.5*x

```
fit4
```

```
##
## Call:
## lm(formula = data4$y ~ data4$x)
##
## Coefficients:
## (Intercept)      data4$x
##         3.0          0.5
```

```
summary(fit1)$r.squared
```

f. R-Squared (2 pts).

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

```
summary(fit2)$r.squared
```

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

```
summary(fit3)$r.squared
```

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

```
summary(fit4)$r.squared
```

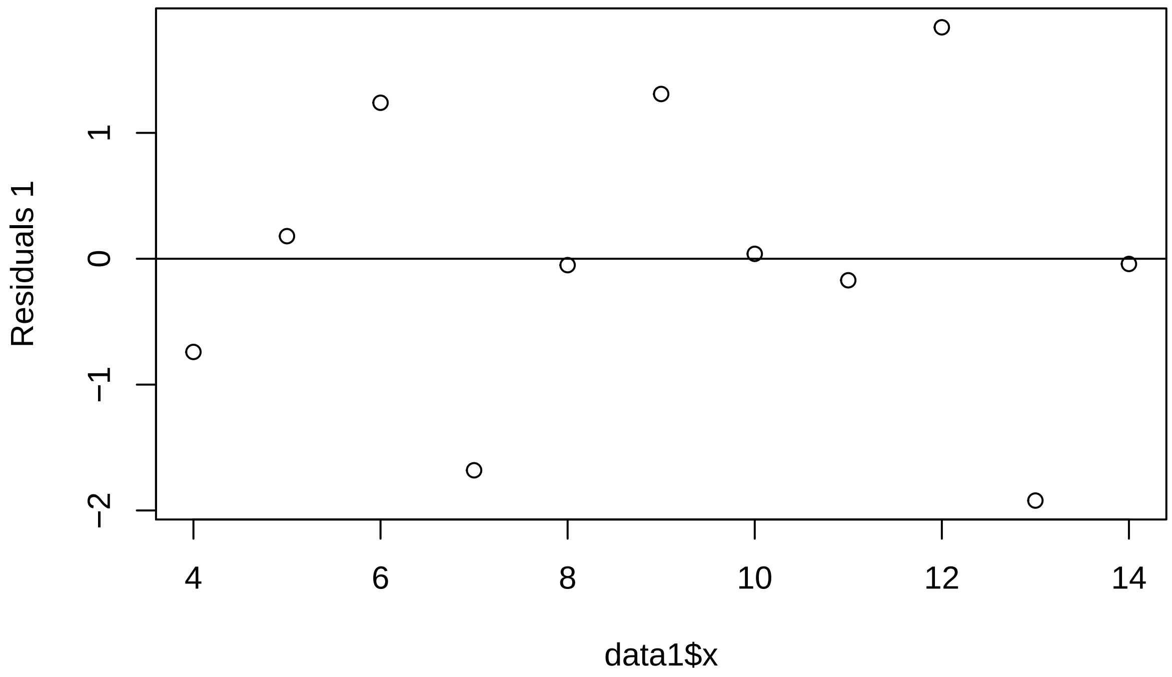
```
## [1] 0.67
```

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) Sources (such as http://stattrek.com/regression/linear-regression.aspx) tell us that Simple linear regression is appropriate when:

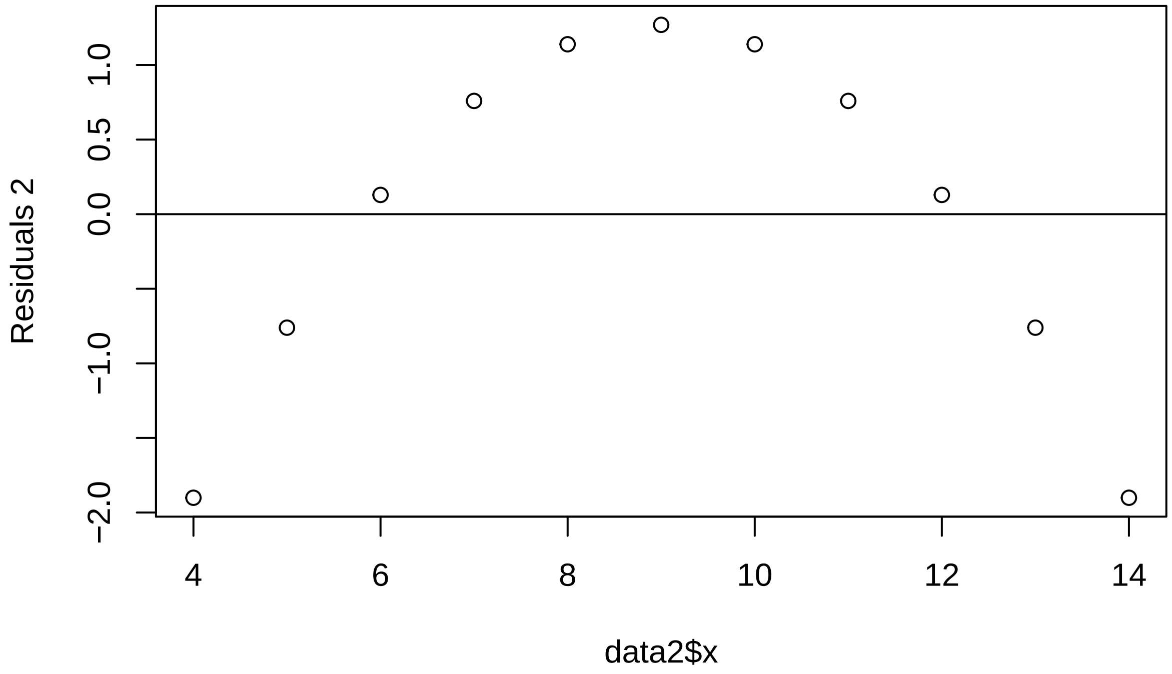
1. The dependent variable Y has a linear relationship to the independent variable X. To check this, make sure that the XY scatterplot is linear and that the residual plot shows a random pattern.
2. For each value of X, the probability distribution of Y has the same standard deviation ??. When this condition is satisfied, the variability of the residuals will be relatively constant across all values of X, which is easily checked in a residual plot

3. For any given value of X, The Y values are independent, as indicated by a random pattern on the residual plot.
4. For any given value of X, The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large. A histogram or a dotplot will show the shape of the distribution.

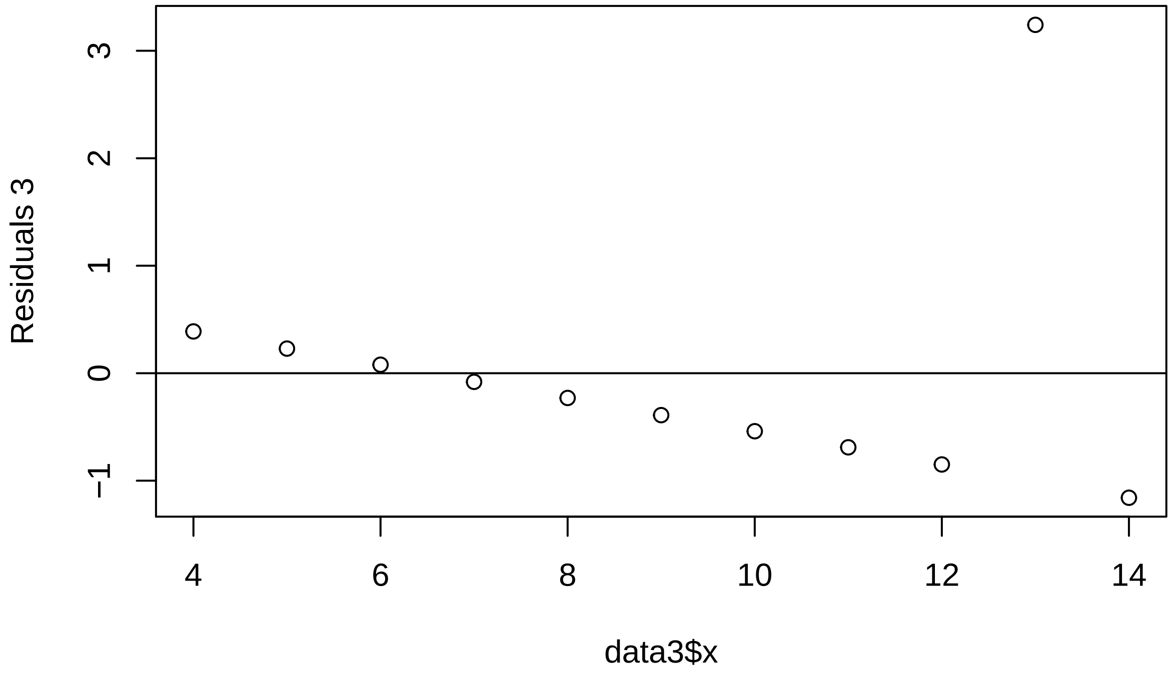
```
resid1 = resid(fit1)
plot(data1$x, resid1, ylab="Residuals 1")
abline(0, 0)
```



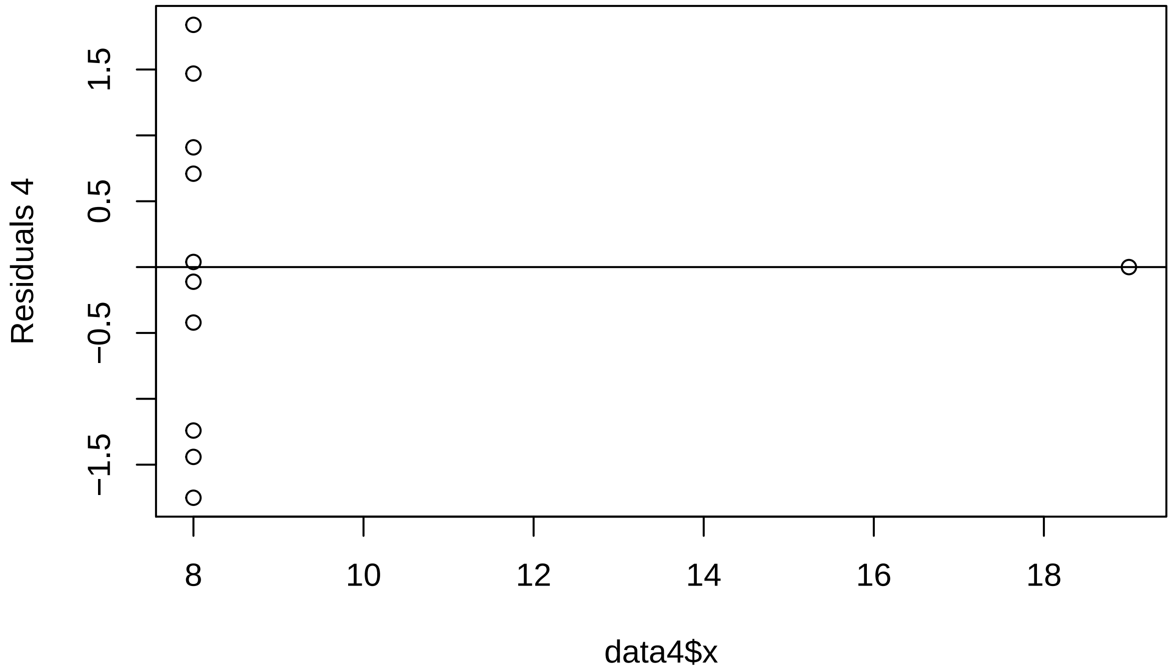
```
resid2 = resid(fit2)
plot(data2$x, resid2, ylab="Residuals 2")
abline(0, 0)
```



```
resid3 = resid(fit3)
plot(data3$x, resid3, ylab="Residuals 3")
abline(0, 0)
```

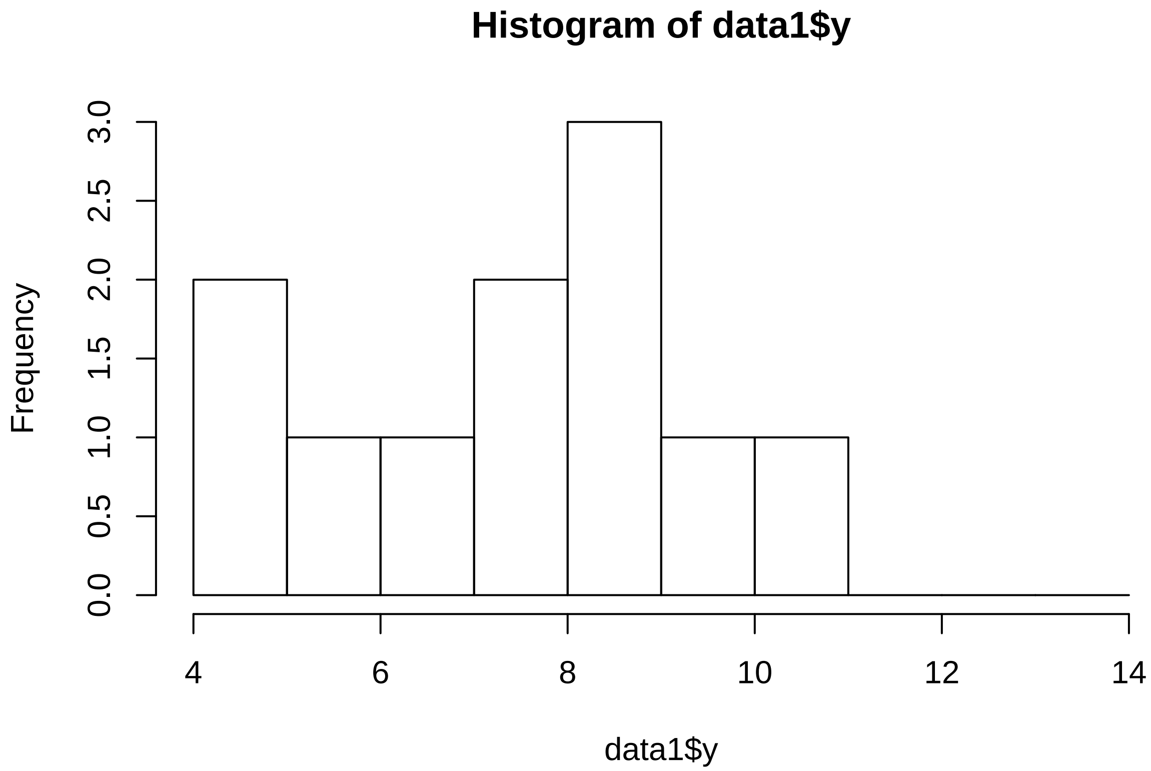


```
resid4 = resid(fit4)
plot(data4$x, resid4, ylab="Residuals 4")
abline(0, 0)
```



- **ONLY** data1's residuals show no pattern, so data2, data3, and data4 should not use a linear regresison model. As for data1, #1, #2, and #4 also hold true, so it is a match for a linear model.
- For **data1**, (1) The XY scatterplot is linear and that the residual plot shows a random pattern, (2) the variability of the residuals will be relatively constant across all values of X, (3) Y's are independent as the residual plot shows, and (4) The Y's are **roughly** normally distributed (see histogram below)

```r
hist(data1$y, data1$x)
```



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

- In all problem solving, a picture is worth 1000 words
- Specifically in these problems, even if it were possible to visualize uneven histogram distributions, skews, and other attributes that point towards problems using particular strategies, it is very difficult to tell by the numbers when there is a pattern in the residuals. Visualizations, on the other hand, make patterns as hidden as the residuals very apparent.
- Included visuals (see above)