Problem Set #2: MVIS 5301

C. Todd Lombardo

16 January 2015

This work can also be found online at: http://rpubs.com/ctodd/PS2

---

1. Suppose the amount of money earned by all MICA graduates from speaking engagements is normally distributed with a mean of $4,040 and a standard deviation of $510. What percentage of MICA graduates earn between $4,000 and $4,500 on speaking engagements?

To find this, we will calculate the Z values at $4,000 and $4,500, then use the z-lookup tables to get the area under the curve between those Z values.
The formula to calculate `Z = (Xi - Xbar) / s` We have been given s, Xi and Xbar

```
# Z = (Xi - Xbar) / s
s = 510
x4000 = 4000
x4500 = 4500
Xbar = 4040

Z4000 = (x4000-Xbar)/s
Z4500 = (x4500-Xbar)/s

Z4000

## [1] -0.07843137

Z4500

## [1] 0.9019608
```

The area of a normal distribution curve between z = 0.902 and z = -0.0784 is 0.8159 (@ Z= 0.90) minus 0.4681 (@Z = -0.08).

```
0.8159 - 0.4681

## [1] 0.3478
```

Therefore, an estimated 34.8% of MICA graduates earn between $4,000 and $4,500 for speaking engagements.

---

2. Find the confidence intervals for the following.

```
A --> X = 20, n = 36, s = 3, confidence interval = 95%
B --> X = 25, n = 36, s = 3, confidence interval = 95%
C --> X = 30, n = 25, s = 4, confidence interval = 90%
D --> X = 50, n = 16, s = 5, confidence interval = 99%
```

Formula for Confidence Interval: CI = X ± Z(s/sqrt(n))
90% CI Z value = 1.645
95% CI Z value = 1.96
99% CI Z value = 2.575

```
# A -- X = 20, n = 36, s = 3, confidence interval = 95%
20 + 1.96 * (3/sqrt(36))

## [1] 20.98

20 - 1.96 * (3/sqrt(36))

## [1] 19.02

# B --> X = 25, n = 36, s = 3, confidence interval = 95%
25 + 1.96 * (3/sqrt(36))

## [1] 25.98

25 - 1.96 * (3/sqrt(36))

## [1] 24.02

# C --> X = 30, n = 25, s = 4, confidence interval = 90%
30 + 1.645 * (4/sqrt(25))

## [1] 31.316

30 - 1.645 * (4/sqrt(25))

## [1] 28.684

# D --> X = 50, n = 16, s = 5, confidence interval = 99%
50 + 2.575 * (5/sqrt(16))

## [1] 53.21875

50 - 2.575 * (5/sqrt(16))

## [1] 46.78125
```

3. This question asks you to think about the directions of causality. Indicate the possible direction(s) of causality for the relationship between unemployment and the crime rate. Discuss your reasoning (a few sentences will suffice).

A correlation between two variables does not indicate which variable is the dependent variable. Does lower crime cause lower unemployment or does lower unemployment cause lower crime? Directions of causality imply that one occurs as a precursor to the other. Employment can be considered an exchange of a person's time for money, so crimes where money is a motivating factor may result in higher causal nature, since the outcome of either act, crime or employment, results in a persons gain of money (legally, or illeglly). Further, since employment takes up someone's time, there may be less time available for people to commit crimes.
It is important to note that crime rate is an overarching term that encompasses many aspects: violent crimes such as murder and rape, property crime such as theft and burglary, statutory crimes such as driving under the influence, and other crimes such as conspiracy, or prostituion. The severity is also another factor, would this be felonies, or heavy crime? Or the more "tame" misdemeanor crimes? Another assumption is that all of these crimes are taken as a whole and that employment rate would be an effect on all of these crimes.

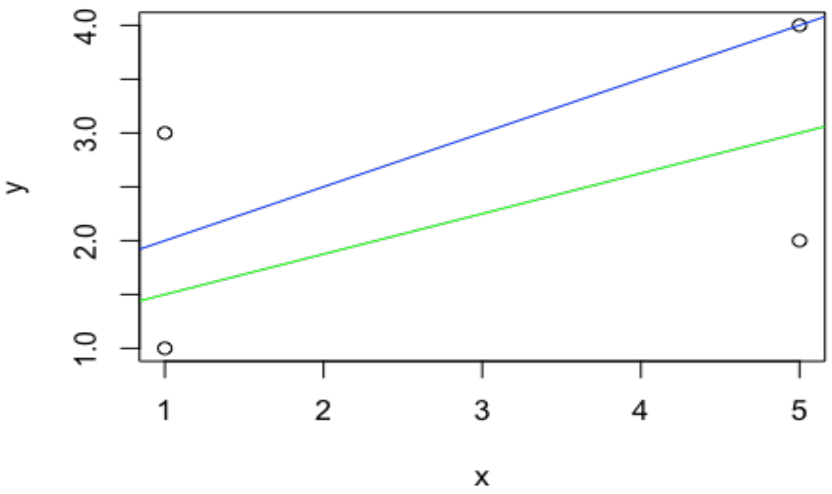
4. There are 3 parts to this question:

x = c(1,1,5,5) y = c(1,3,2,4)

Line A: y = 1.5 + 0.5x
Line B: y = 1.125 + 0.375x

a. Graph the linear equations and data points

```
x = c(1,1,5,5)          # Add the data to a vector
y = c(1,3,2,4)
plot (x,y)                         # Plot the data
abline(1.5,0.5, col = 'blue')      # Line A in Blue
abline(1.125,0.375, col = 'green') # Line B in Green
```



b. Construct tables for x, y, yHat, e, and e2;

Line A: y = 1.5 + 0.5x **Line B: y = 1.125 + 0.375x**

x	y	yHat	e	e2
1	1	2	-1	1
1	3	2	1	1
5	2	4	-2	4
5	4	4	0	0

x	y	yHat	e	e2
1	1.0	1.5	-0.5	0.25
1	3.0	1.5	1.5	2.25
5	2.0	3.0	-1.0	1
5	4	3.0	1.0	1

c. Determine which line fits the set of data points better, according to the least-squares criterion

For Line A the sum of e2 = 6 and for Line B, the sum of e2 = 4.5, therefore Line B is considered the better fitting line for this dataset.

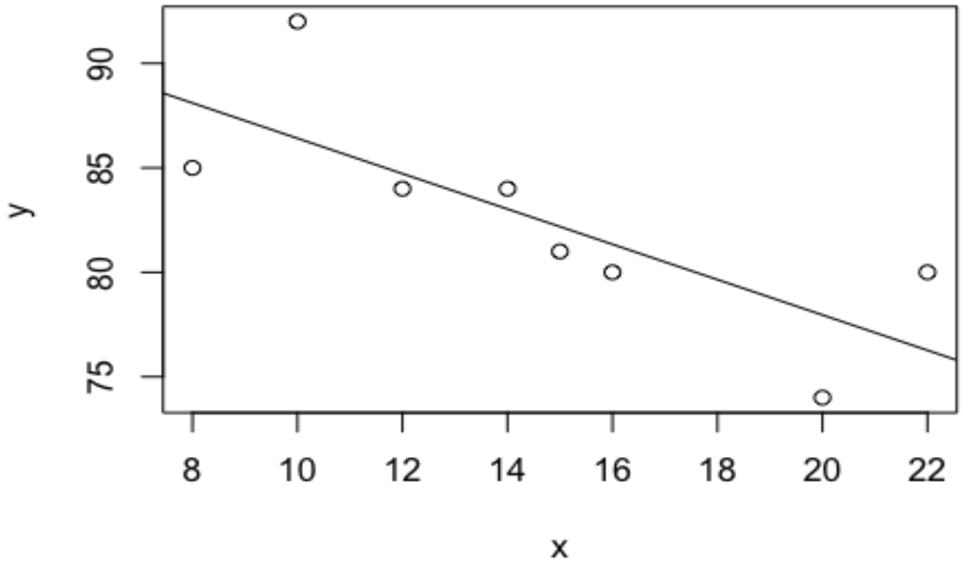
5. An instructor at Arizona State University asked a random sample of eight students to record their study times in a beginning calculus course. She then made a table for total hours studied (x) over 2 weeks and the test score (y) at the end of the 2 weeks. Here are the results.

x 10 15 12 20 8 16 14 22
y 92 81 84 74 85 80 84 80

The regression equation for these data is: yHat = 94.86698 – 0.84561x.

a. Graph the regression equation and the data points

```
x = c(10,15,12,20,8,16,14,22)          # Put data in to Vectors
y = c(92,81,84,74,85,80,84,80)
plot (x,y)                             # Plot the Data
regLine=lm(y~x)                        # R's Linear Model function: s
hould be the same regression data given
regLine                                # Call regLine to ensure slope
and y-intercept are same

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     94.8670      -0.8456

abline(regLine)                        # Plot the line on the graph
```



b. Describe the apparent relationship between the two variables under consideration.

There appears to be a negative relationship between the variables: Longer hours of study is correlated with exam grade decrease.

c. Interpret the slope of the regression line.

Slope is in a negative direction; -0.84561. for every 1 hour increase in reported study time, a drop in 0.8456 exam points could be observed. One should note that 0 reported hours would result in a 94.87 score (y-intercept, where x = 0) on the exam, certainly an interesting finding.

d. What is the predicted exam score for a student who studies for 15 hours?

```
 # y = 94.86698 - 0.84561x
 # x = 15
94.86698 - 0.84561*(15)  # Thinking there should be a STD range call
out here? since this is a predicted score?

## [1] 82.18283
```

---

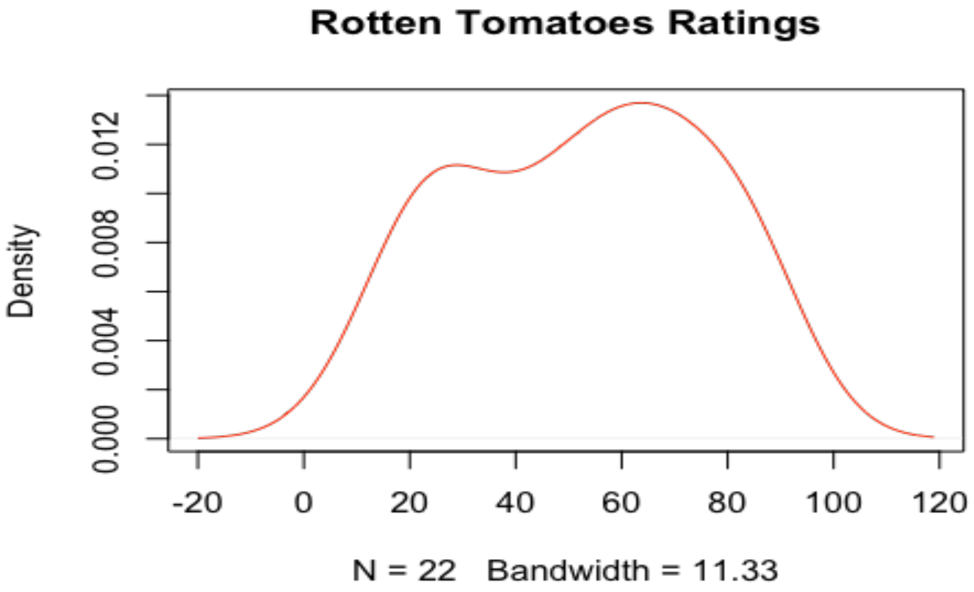
6. Read this article from Nate Silver at 538 http://fivethirtyeight.com/datalab/killing-the-interview-could-cost-sony-100-million/ and write a paragraph about the potential sources of bias and error in the data and regression methodology.

Souces of error: The time of year the film is released, the film is loosely based on some real-world news about North Korea, so there is a news-media cycle that could influence the revenue. Rotten Tomatoes is a consumer based site and, while they take steps to control for abuse, certain films may have inflated ratings. The data sources are cited as IMDB.com, also a crowd-sourced site, which may nor may not be the most accurate in terms of reported information in the movie industry. The assumption that the average rating would be the rating for "The Interview" could also be invalid. The predictions fall within the standard deviations of the dataset. In calculating Y-hat based on a regression of ratings and revenue, the predicted value is $115m, witihin the standard error but higher than the value in the article.

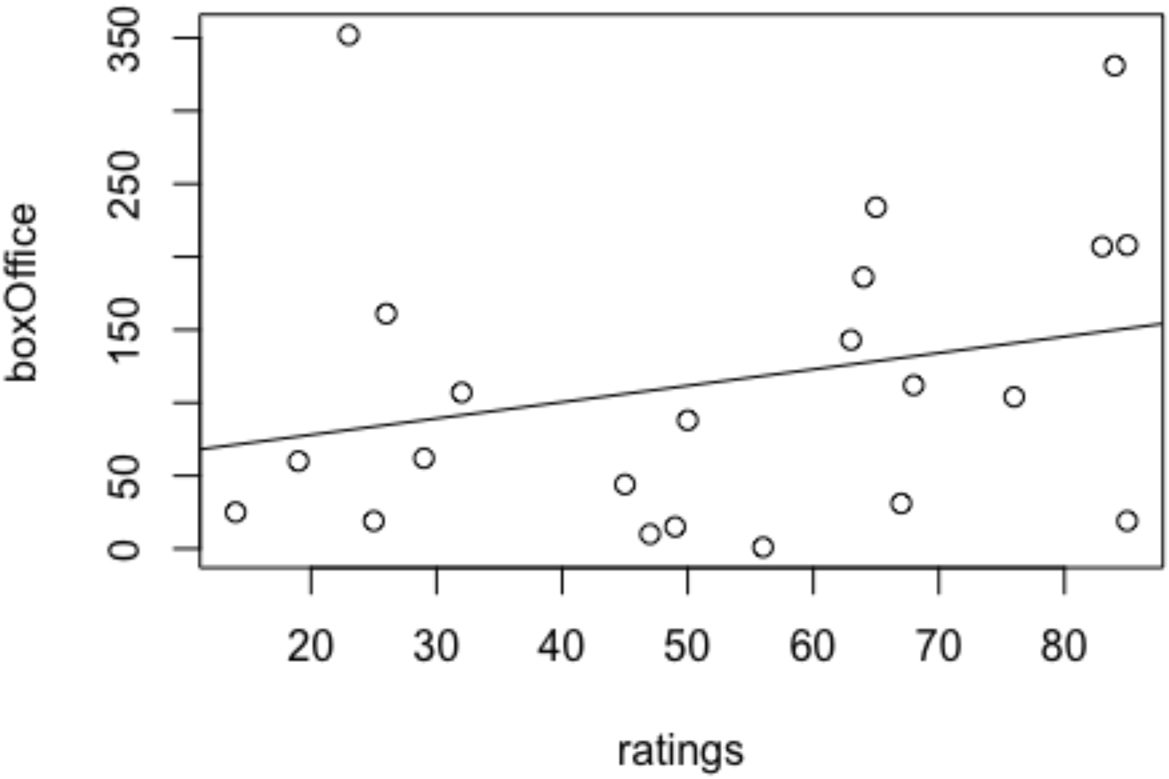
Some calculations to play with the article's dataset

```
ratings = c(23,84,65,85,83,64,26,63,68,32,76,50,29,19,45,67,14,25,85
,49,47,56)
boxOffice = c(352,331,234,208,207,186,161,143,112,107,104,88,62,60,4
4,31,25,19,19,15,10,1)
summary(ratings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   14.00   29.75   53.00   52.50   67.75   85.00

sd(ratings)

## [1] 23.36002

plot(density(ratings), main="Rotten Tomatoes Ratings", col= "red")
```



```
plot(ratings, boxOffice)
regLine1=lm(boxOffice~ratings)
abline(regLine1)
```



```
regLine1

## 
## Call:
## lm(formula = boxOffice ~ ratings)
## 
## Coefficients:
## (Intercept)      ratings  
##      55.572        1.122

# Y = 55.572 + 1.122x
55.572 + 1.122*53

## [1] 115.038
```