Preyasi Gaur

705704939

Stats 10

Lab 3

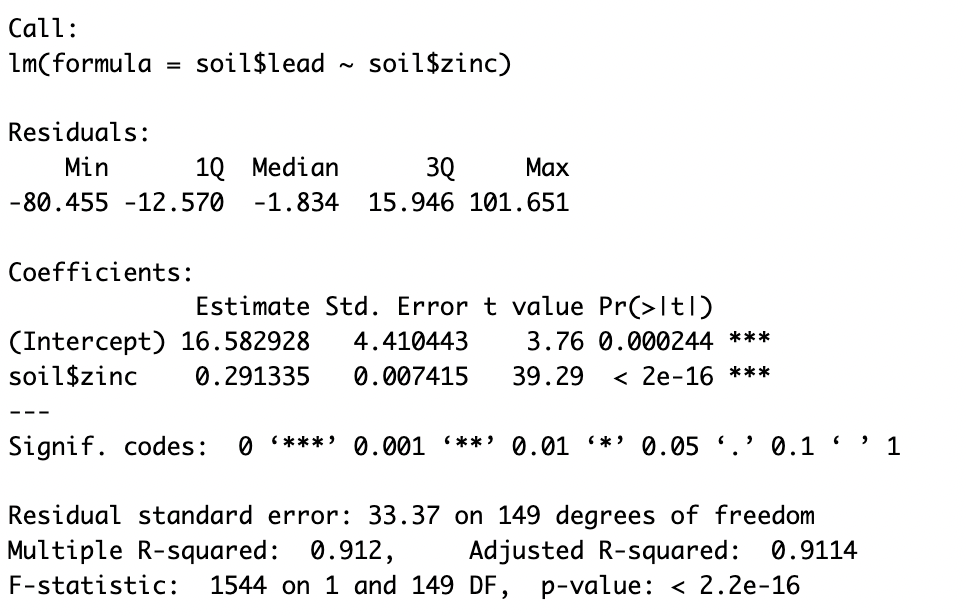
**Exercise 1**

1. Question:   
   Run a linear regression of lead against zinc concentrations (treat lead as the response variable). Use the summary function just like in the example above and paste the output into your report.

Prompt Code:   
soil <- read.table("/Users/preyasigaur/Desktop/soil\_complete.txt", header = TRUE)

soil\_model <- lm(soil$lead ~ soil$zinc)

summary(soil\_model)

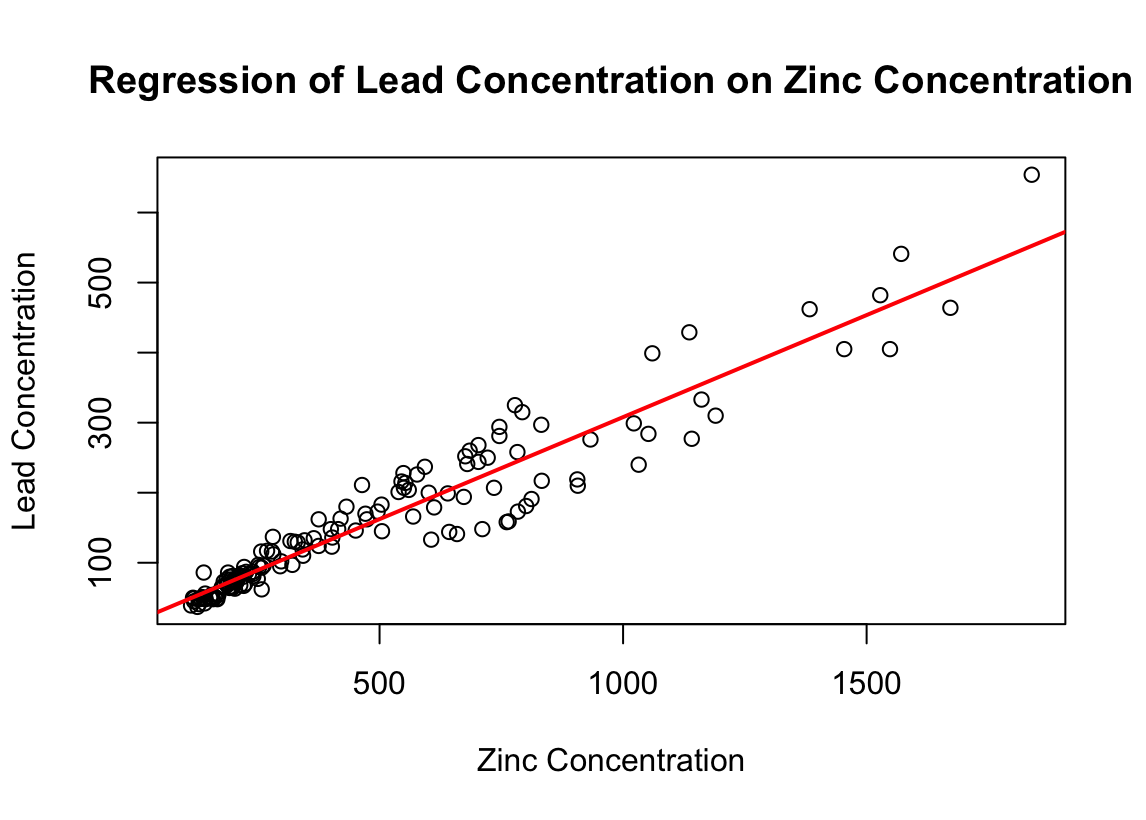
Output:   


1. Question: Plot the lead and zinc data, then use the abline() function to overlay the regression line onto the data.

Prompt Code:   
plot(soil$lead ~ soil$zinc, xlab = "Zinc Concentration", ylab = "Lead Concentration", main = "Regression of Lead Concentration on Zinc Concentration")

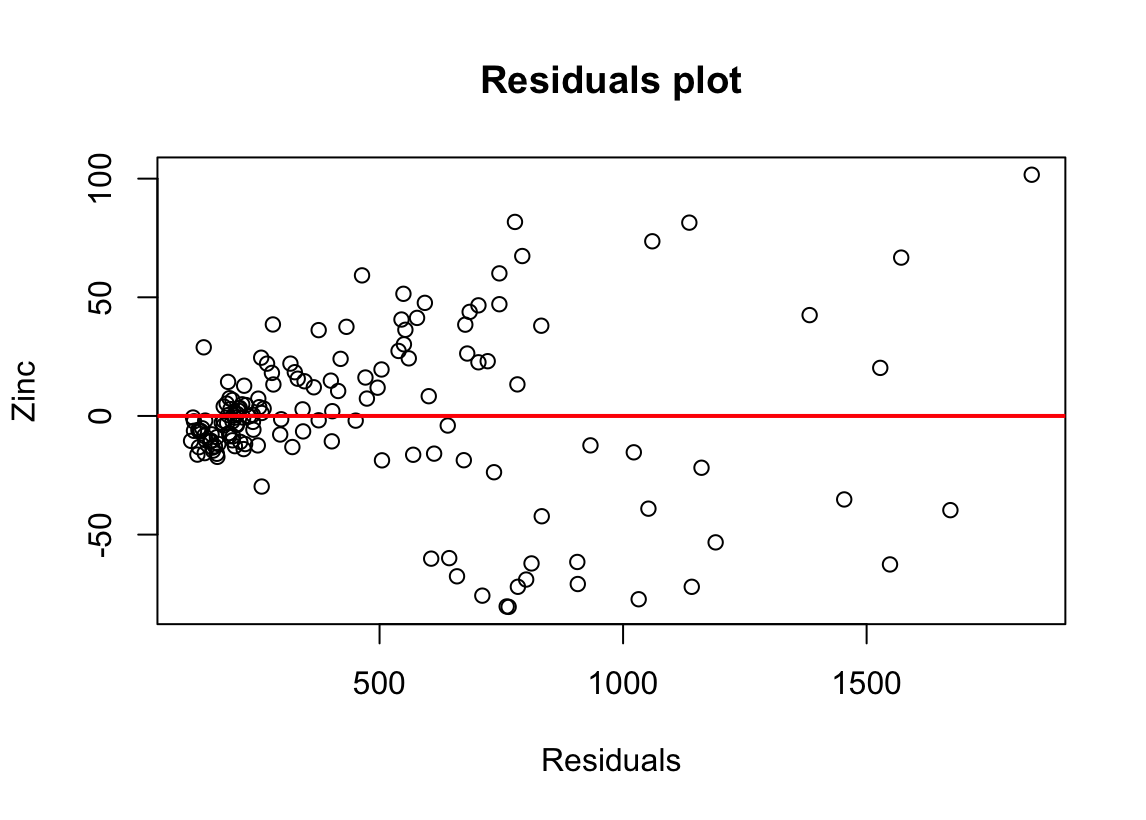
abline(soil\_model, col = "red", lwd = 2)

Output:



1. Question:   
   In a separate plot, plot the residuals of the regression from (a), and again use the abline() function to overlay a horizontal line.  
     
   Prompt Code:   
   > plot(soil\_model$residuals ~ soil$zinc, xlab = "Residuals", ylab = "Zinc", main = "Residuals plot")

> abline(a = 0, b = 0, col = "red", lwd = 2)

Output:  


1. Question:   
   Based on the output from (a), what is the equation of the linear regression line?

Answer:  
 Lead Concentration = 16.582928 + 0.291335 \* Zinc Concentration

1. Question:   
   Imagine we have a new data point. We find out that the zinc concentration at this point is 1,000 ppm. What would we expect the lead concentration at this point to be?

Answer:   
16.582928 + 0.291335 \* 1000 = 307.917928 ppm

1. Question:   
   Imagine two locations (A and B) for which we only observe zinc concentrations. Location A contains 100ppm higher concentration of zinc than location B. How much higher would we expect the lead concentration to be in location A compared to location B?

Answer:  
zinc\_A = 100 + zinc\_B  
lead\_A - lead\_B = (16.582928 + 0.291335 \* zinc\_A) - (16.582928 + 0.291335 \* zinc\_B)  
 = 0.291335 \* (zinc\_A - zinc\_B)  
 = 0.291335 \* 100

= 0.291335 \* 100

= 29.1335 ppm

1. Question:   
    Report the R-squared value and explain in words what it means in context.

Answer:   
The R-squared value is 0.912. This is extremely high and means that 91.2% of the variation in lead concentration can be explained by the zinc concentration.

1. Question:  
   Comment on whether you believe the three main assumptions (linearity, symmetry, equal variance) for linear regression are met for this data. List any concerns you have.

Answer:

Linearity Assumption: Since most of the data is close to the regression line, the linearity assumption holds well.

Symmetry Assumption: Although relatively subjective, the symmetry is good as roughly there is the same number of positive and negative residuals on both the sides of the regression line.

Variance Assumption: This is violated, as the x-value is directly proportional to the variance. As we traverse along the x-axis, we see a clear increase in the residual variance.

**Exercise 2**

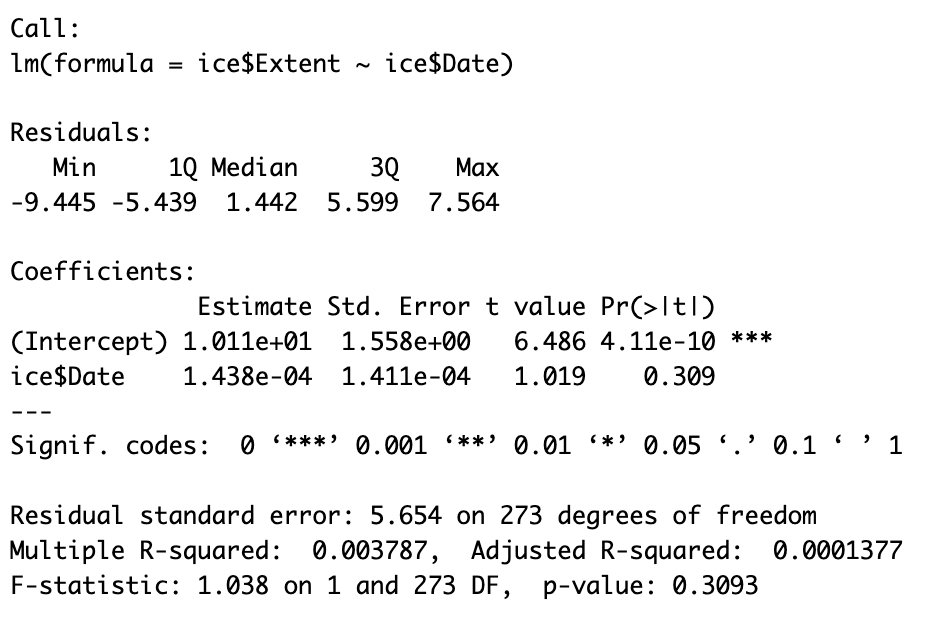
1. Question:   
   Produce a summary of a linear model of sea ice extent against time.  
     
   Prompt Code:   
   ice <- read.csv("/Users/preyasigaur/Desktop/sea\_ice.csv", header = TRUE)

ice$Date <- as.Date(ice$Date, "%m/%d/%Y")

ice\_model <- lm(ice$Extent ~ ice$Date)

summary(ice\_model)

Output:



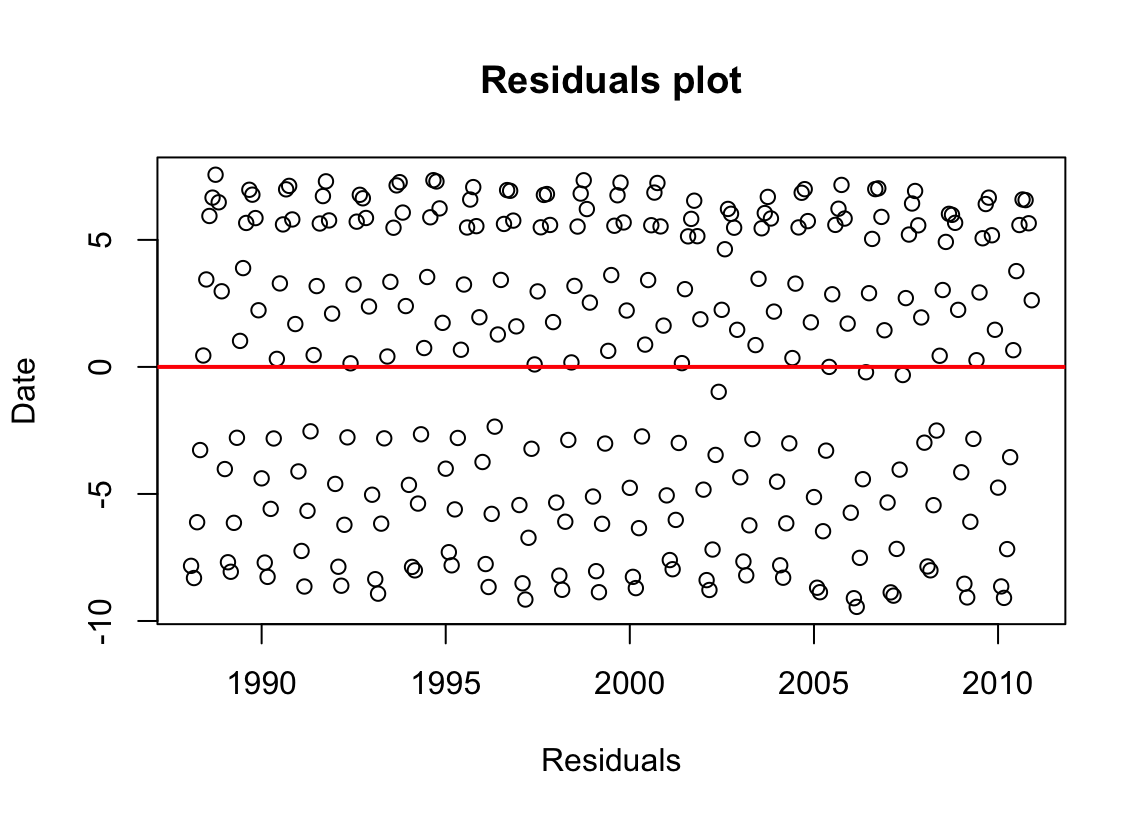
1. Question:   
   Plot the data and overlay the regression line. Does there seem to be a trend in this data?

Prompt Code:   
plot(ice$Extent ~ ice$Date, xlab = "Date", ylab = "Extent", main = "Regression of Extent against Date")  
abline(ice\_model, col = "red", lwd = 2)  
  
Output:   
  
Though the slope is very small, there is a positive trend. It also happens to be oscillating in its nature.

1. Question:   
   Plot the residuals of the model over time and include a horizontal line. What assumption(s) about the linear model should we be concerned about?

Prompt Code:

plot(ice\_model$residuals ~ ice$Date, xlab = "Residuals", ylab = "Date", main = "Residuals plot")  
abline(a = 0, b = 0, col = "red", lwd = 2)

Output:   


Linearity Assumption: As noticeable, most of the data is far from the regression line and thus the linearity assumption is not met

Symmetry Assumption: The symmetry is not met as there is a huge gap below the line.

Variance Assumption: The variation assumption is met, as the data points are randomly distributed around y = 0 for this plot.

**Exercise 3**

1. Question:  
   Based on your lecture notes, what is the chance Adam will double his money in the first round of the game? What is the chance Adam will lose his money in the first round of the game?  
     
   Answer:   
   Notice that in this situation,  
   Total possible cases => 36 cases

(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)

(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6)

(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)

(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)

(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6)

(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)  
  
Cases where money is doubled (sum = 7 or 11) => 8 cases  
(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3), (5, 6), (6, 5)

Cases where the money is lost (sum = 2, 3, or 12) => 4 cases

(1, 1), (1, 2), (2, 1), (6, 6)

P(doubling money) = =   
P(losing money) = =

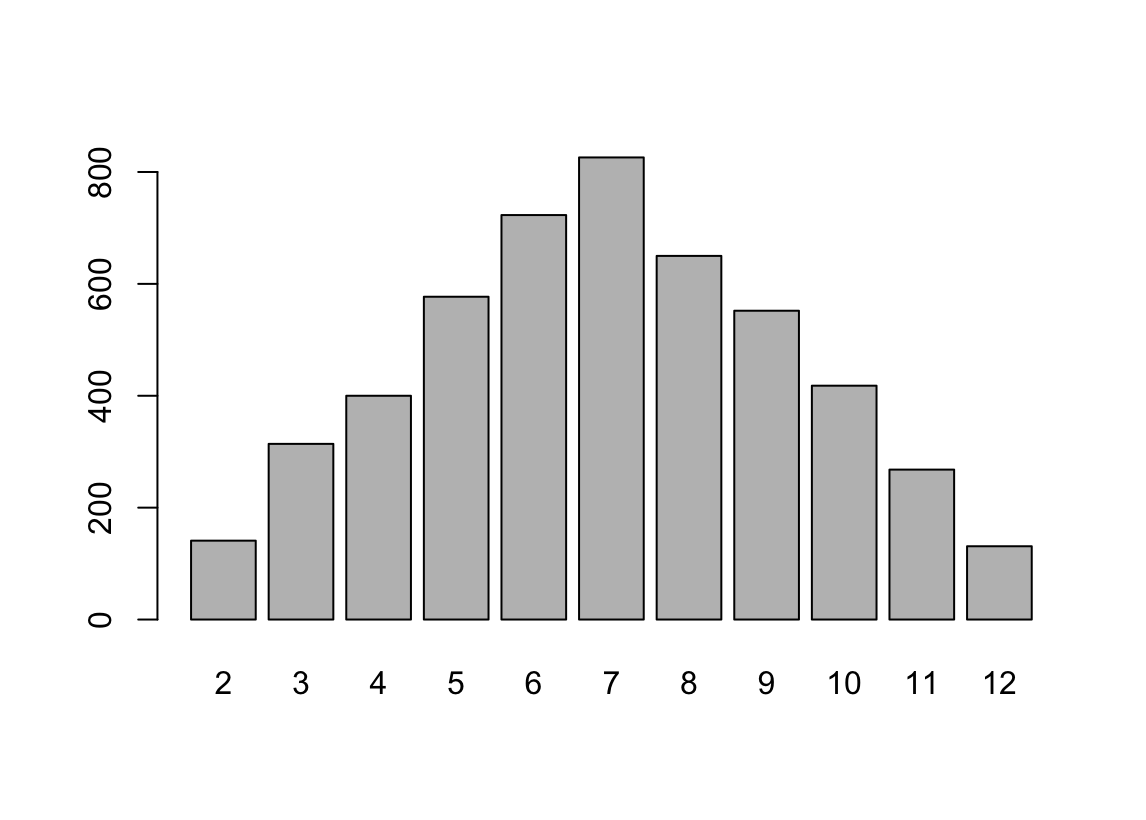
1. Question:   
    Let’s now approximate the results in (a) by simulation. First, set the seed to 123. Then, create an object that contains 5,000 sample first round Craps outcomes (simulate the sum of 2 dice, 5,000 times). Use the appropriate function to visualize the distribution of these outcomes (*hint: are the outcomes discrete or continuous?*)  
     
   Prompt Code:

set.seed(123)

numbers = 1:6

rand\_draws = replicate(5000, sample(numbers, 2, replace = TRUE))

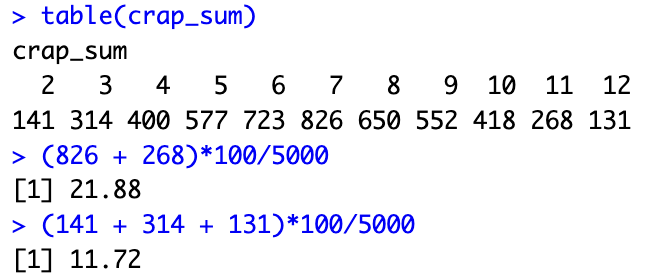
crap\_sum = colSums(rand\_draws)

barplot(table(crap\_sum))  
  
Output:  


1. Question:   
   Imagine these sample results happened in real life for Adam. Using R functions of your choice, calculate the percentage of time Adam doubled his money. Calculate the percentage of time Adam lost his money.

Prompt Code:

table(crap\_sum)  
(826 + 268)\*100/5000  
(141 + 314 + 131)\*100/5000

Output:   


Thus Adam doubled his money 21.88% of the time and lost his money 11.72% of the time.

1. Question:   
    Adam winning money and Adam losing money can both be considered events. Are these two events independent, disjoint, or both? Explain why.  
     
   Answer:   
   As Adam cannot simultaneously win and lose money, these events are disjoint, i.e there is no intersection between them.   
   We also know that two events, A and B are independent if P(B | A) = P(B)

Notice that in our case, P(B | A) = 0, but P(B) > 0. Thus, in this case P(B | A) ≠ P(B) and thus, the events are not independent.

Therefore the two events, Adam winning money and Adam losing money are disjoint but not independent.

1. Question:   
    Quickly mathematically verify by calculator if those events are independent using part (a) and what you learned in lecture. Show work.  
     
   Answer:

We know that two events, A and B are independent if P(A ∩ B) = P(A)\*P(B)

Notice that in our case,   
P(A ∩ B) = 0  
P(A) \* P(B) = \* = ≠ 0.   
Thus, in this case P(A ∩ B) ≠ P(A)\*P(B) and the events are not independent.