STATS 10 ASSIGNMENT 3

PART I

Exercise 1

We will be working with some soil mining data and are interested in looking at some of the relationships between metal concentrations (in ppm). Download the data 'soil_complete.txt' from the course website and read it into R. When you read in the data, name your object "soil".

Code:

```
soil <- read.table("soil_complete.txt", header=TRUE)
#as its txt so we have to use read.table.
#header=true means that first row of the dataset contains header
#linear regression
```

Output:

```
> soil <- read.table("soil_complete.txt", header=TRUE)
>
```

a. 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.

Code:

```
linear_model <- lm(lead ~ zinc, data = soil)
###in the above the first variable is the response variable and then you tell it where
##it creates a list which contains info like coefficients ressiduals
summary(linear_model)
```

Output:

```
> linear_model <- lm(lead ~ zinc, data = soil)
> ###in the above the first variable is the response variable and then you
tell it where the data is from
> ##it creates a list which contains info like coefficients ressiduals
> summary(linear_model)

Call:
lm(formula = lead ~ zinc, data = soil)

Residuals:
    Min      1Q  Median      3Q     Max
-80.455 -12.570  -1.834  15.946 101.651

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.582928   4.410443    3.76 0.000244 ***
```

```
zinc         0.291335   0.007415   39.29  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.37 on 149 degrees of freedom
Multiple R-squared:  0.912,    Adjusted R-squared:  0.9114
F-statistic:  1544 on 1 and 149 DF,  p-value: < 2.2e-16
```

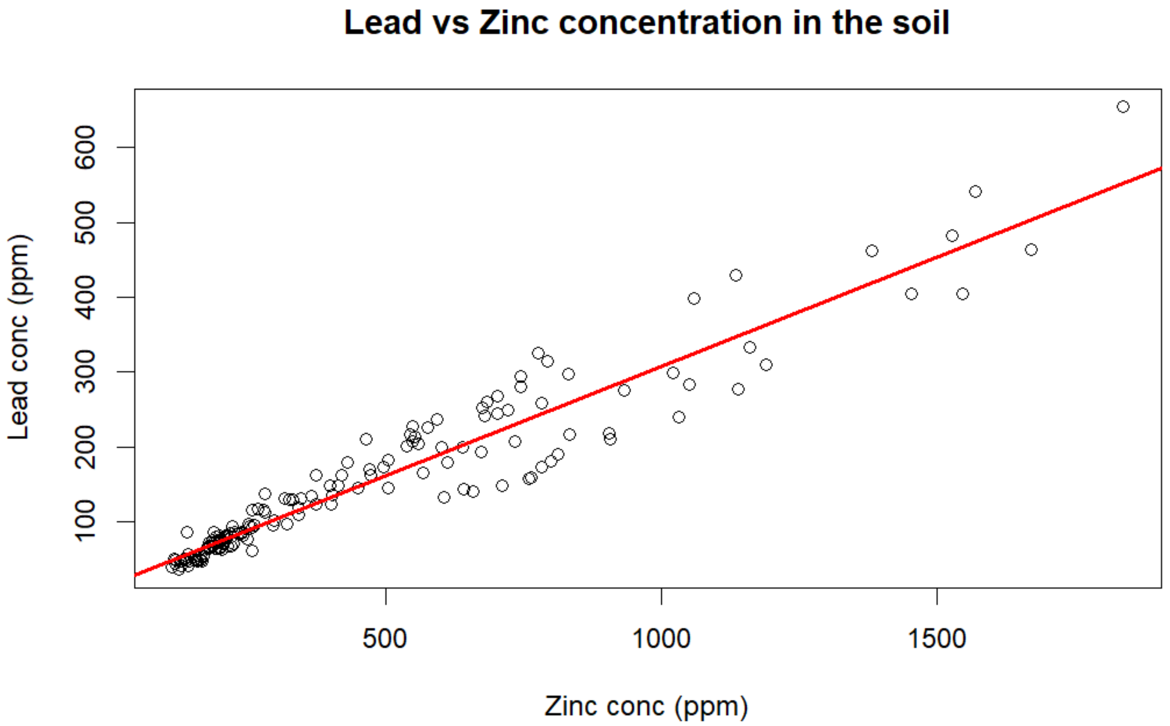
b. Plot the lead and zinc data, then use the abline() function to overlay the regression line onto
the data.

Code:

```
#Plot the lead and zinc data, then use the abline() function
##
plot(lead ~ zinc, data=soil, xlab="Zinc conc (ppm)", ylab= "Lead conc (ppm)",
     main="Lead vs Zinc concentration in the soil")
     abline(linear_model, col="red", lw=2 )
## we use abline(a,b) as a func to plot y= a+bx where a is intercept and b is slope
```

Output:

```
> plot(lead ~ zinc, data=soil, xlab="Zinc conc (ppm)", ylab= "Lead conc (p
pm)", main="Lead vs Zinc concentration in the soil")
>      abline(linear_model, col="red", lw=2 )
```



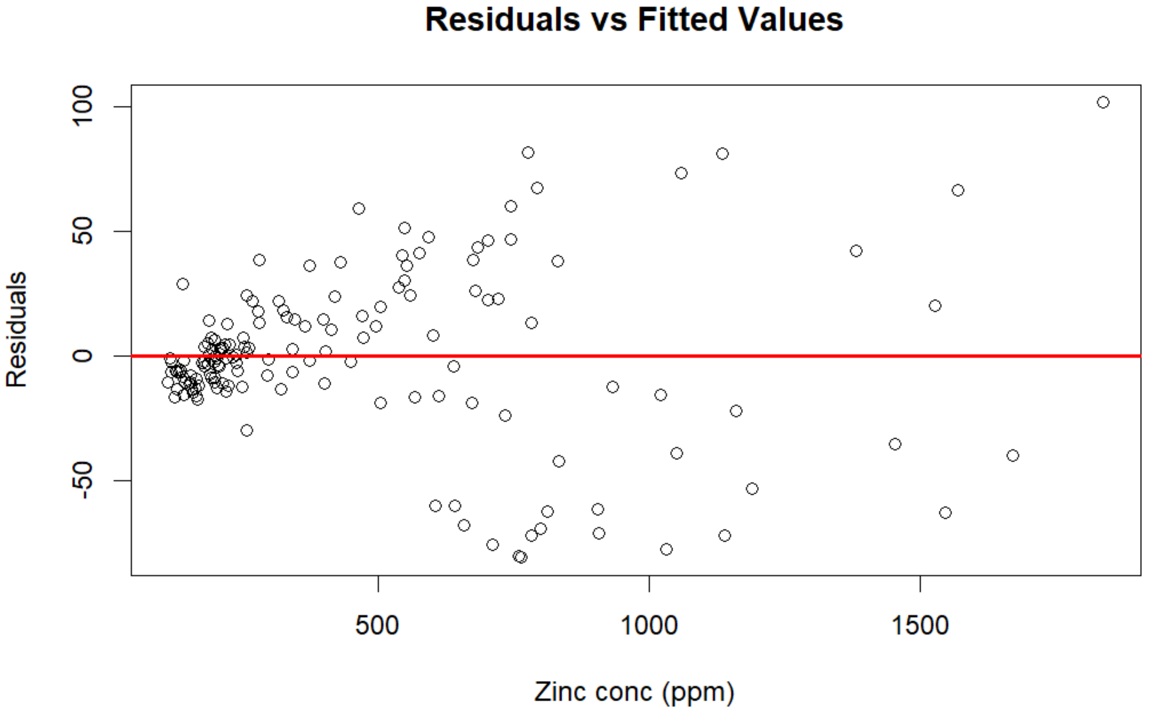
c. In a separate plot, plot the residuals of the regression from (a), and again use the abline() function to overlay a horizontal line.

Code:

```
plot(linear_model$residuals ~ soil$zinc,
     xlab="Zinc conc (ppm)", ylab= "Residuals",
     main= "Residuals vs Fitted Values")
abline(a=0, b=0, col= "red", lw=2)#plots horizontal line helps check if residual is symmetric across
```

Output:

```
> plot(linear_model$residuals ~ soil$zinc,xlab="Zinc conc (ppm)", ylab= "R
esiduals", main= "Residuals vs Fitted Values")
> abline(a=0, b=0, col= "red", lw=2)
```



Parts d-h can be answered by hand, using a calculator, or any R functions of your choice.

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

```
#1d

summary(linear_model)
```

Output:

```
> summary(linear_model)

Call:
lm(formula = lead ~ zinc, data = soil)

Residuals:
    Min      1Q  Median      3Q     Max 
-80.455 -12.570  -1.834  15.946 101.651 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.582928   4.410443    3.76 0.000244 ***
```

```
zinc         0.291335   0.007415   39.29  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.37 on 149 degrees of freedom
Multiple R-squared:  0.912,    Adjusted R-squared:  0.9114
F-statistic:  1544 on 1 and 149 DF,  p-value: < 2.2e-16
```

Intercept estimate 16.582928
zinc slope estimate 0.291335
The equation is lead= 16.582928 + 0.291335 * zinc
The equation is lead= 16.58 + 0.29 * zinc

e. 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?

Plug the new data point into the regression equation found in the previous part:
16.582928 + 0.291335 * 1000
The equation is lead= 16.582928 + 0.291335 * zinc

=307.9179

The prediction for lead concentration is 307.9179 ppm

f. 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?

There are data points at two locations
lead_A= 16.582928 + 0.291335 * zinc_A
lead_B= 16.582928 + 0.291335 * zinc_B

lead_A= 16.582928 + 0.291335 * (zinc_B + 100)
lead_B= 16.582928 + 0.291335 * zinc_B

lead_A - lead_B= 16.582928 + 0.291335 * (zinc_B + 100) - (16.582928 + 0.291335 * zinc_B)
lead_A - lead_B= 0.291335 *(zinc_B +100)- 0.291335 * zinc_B
lead_A - lead_B= 29.1335

We expect the lead concentration at site A to be 29.1335 ppm higher than the concentration at site B.

g. Report the R-squared value and explain in words what it means in context.

Code:

```
#1g
summary(linear_model)
```

Output:

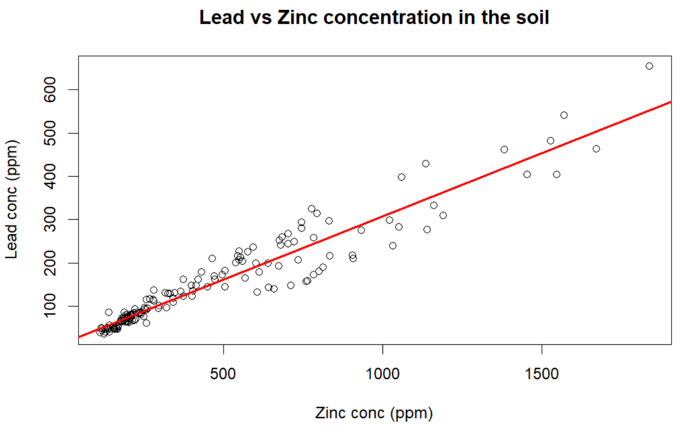
```
> #1g
> summary(linear_model)

Call:
lm(formula = lead ~ zinc, data = soil)

Residuals:
    Min      1Q  Median      3Q     Max 
-80.455 -12.570  -1.834  15.946 101.651 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.582928   4.410443    3.76 0.000244 ***
zinc         0.291335   0.007415   39.29  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.37 on 149 degrees of freedom
Multiple R-squared:  0.912,	Adjusted R-squared:  0.9114 
F-statistic:  1544 on 1 and 149 DF,  p-value: < 2.2e-16
```

R-squared value (from the information provided using summary func) = 0.912

It describes the proportion of variance of the response variable that is explained by the explanatory variable. This means that 91.2% of the variation in lead concentration (response variable) can be explained by the zinc concentration (explanatory variable).

h. 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.

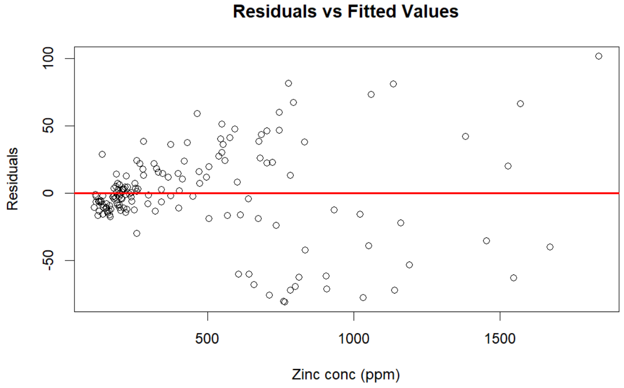
Linearity: I believe that the linearity condition is met. Based on the lead vs zinc plot, the two variables have a linear relationship since most of the data is close to the regression line. Answered using the graph below:



Symmetry: Based on the residual plot, it seems like the residuals are scattered symmetrically across the x-axis (positive and negative residuals line up with one another).

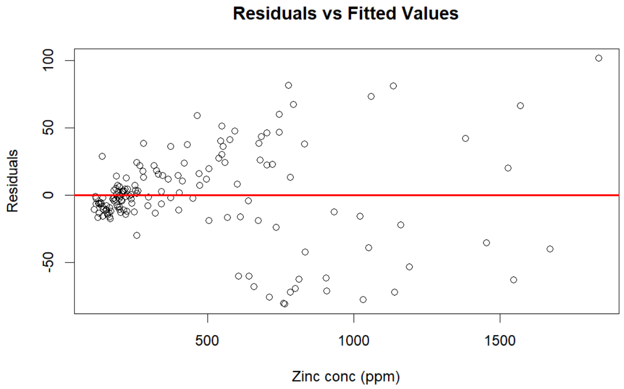
Thus, the symmetry condition has been reasonably met since roughly the same number of positive and negative residuals are on both sides of the line.

Answered using the graph:



Equal variance: Looking at the lead vs zinc plot, it looks like the points in the bottom left corner are more closely clustered than the points in the top right. Also, in the residual plot we can see that the residuals are close to zero on the left and more spread out on the right. The x-value is directly proportional to variance. As we traverse along the x-axis, we see an increase in residual variance. This indicates that the variance is not equal across all values of the explanatory variable.

Answered using the graph:



Exercise 2

Our next data set is what is known as a time series, or data in time. It contains the measurements
via satellite imagery of sea ice extent in millions of square kilometers for each month from 1988
to 2011. Please download the “sea_ice” data from the course website and read it into R. If you
have your working directory properly set, you can use the line below:
ice <- read.csv("sea_ice.csv", header = TRUE)

Note that currently R does not know what class the Date column is. We need to convert the Date
column into class “date” using the following line:
ice$Date <- as.Date(ice$Date, "%m/%d/%Y")

a. Produce a summary of a linear model of sea ice extent against time.
Code:

```
#2a
ice <- read.csv("sea_ice.csv", header = TRUE)

ice$Date <- as.Date(ice$Date, "%m/%d/%Y")
#we need to do the above as R does not know what the date column is
#we convert the Datecolumn into class “date”

##Produce a summary of a linear model of sea ice extent against time.
linear_model_ice <- lm(Extent ~ Date, data = ice)
#sea ice extent is the response variable and time is the explanatory variable.
summary (linear_model_ice)
```

Output:

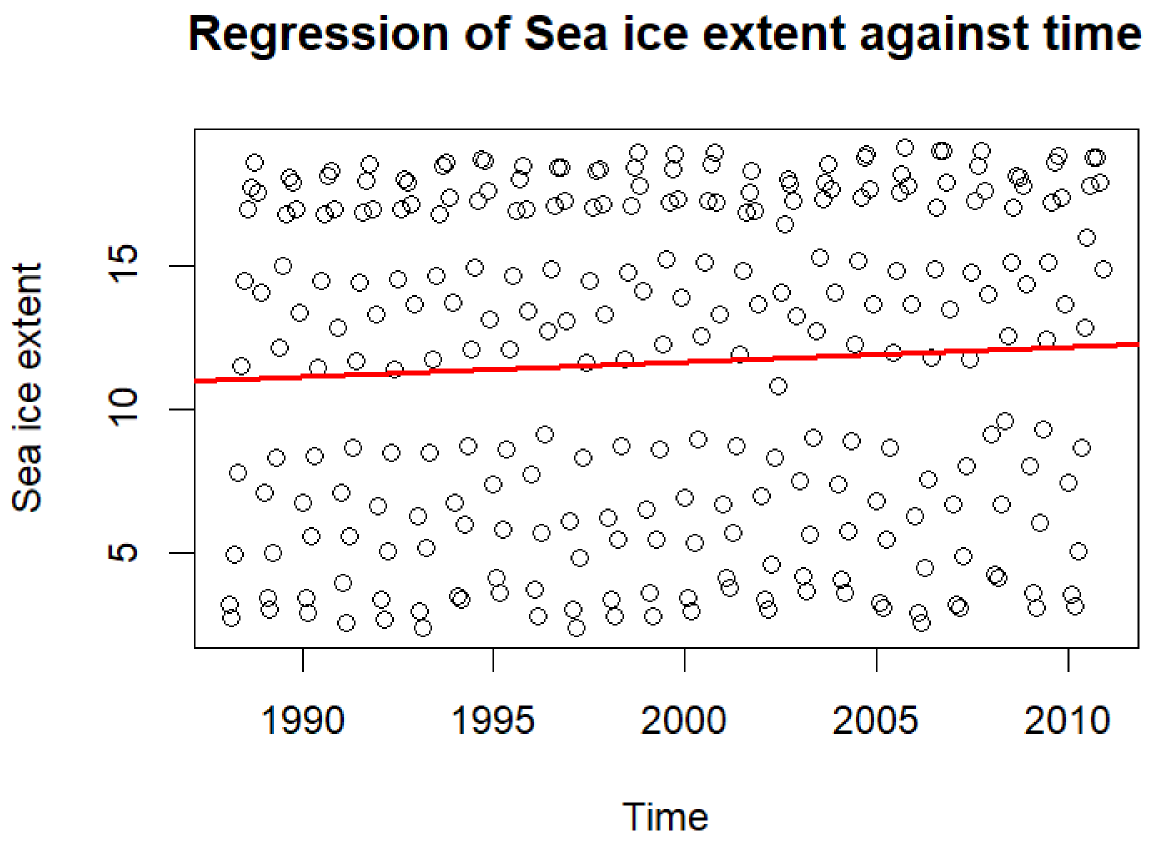
```
> ice <- read.csv("sea_ice.csv", header = TRUE)
>
> ice$Date <- as.Date(ice$Date, "%m/%d/%Y")
> ##Produce a summary of a linear model of sea ice extent against time.
> linear_model_ice <- lm(Extent ~ Date, data = ice)
> #sea ice extent is the response variable and time is the explanatory var
iable.
> summary (linear_model_ice)

Call:
lm(formula = Extent ~ Date, data = ice)

Residuals:
   Min     1Q Median     3Q    Max
-9.445 -5.439  1.442  5.599  7.564

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.011e+01  1.558e+00   6.486 4.11e-10 ***
Date        1.438e-04  1.411e-04   1.019    0.309
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.654 on 273 degrees of freedom
Multiple R-squared:  0.003787, Adjusted R-squared:  0.0001377
F-statistic: 1.038 on 1 and 273 DF,  p-value: 0.3093
```

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

```
#2b

##Plot the data and overlay the regression line.
plot(Extent ~ Date, data=ice, xlab="Time", ylab= "Sea ice extent",
      main="Regression of Sea ice extent against time")
abline(linear_model_ice, col="red", lw=2 )
```

Output:

```
> plot(Extent ~ Date, data=ice, xlab="Time", ylab= "Sea ice extent",
+       main="Regression of Sea ice extent against time")
> abline(linear_model_ice, col="red", lw=2 )
```



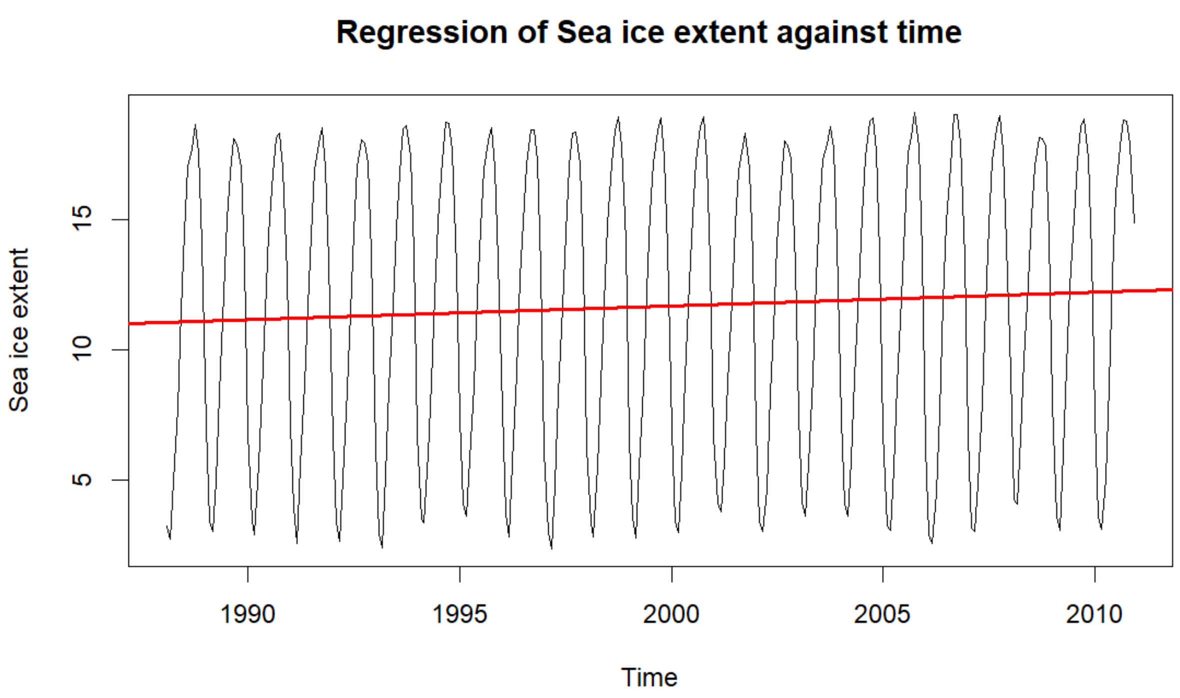
Trend:

Code:

```
##Plot the data and overlay the regression line.
plot(Extent ~ Date, data=ice, xlab="Time", ylab= "Sea ice extent",
      main="Regression of Sea ice extent against time", type= "l")
abline(linear_model_ice, col="red", lw=2 )
```

Output:

```
> plot(Extent ~ Date, data=ice, xlab="Time", ylab= "Sea ice extent",
+       main="Regression of Sea ice extent against time", type= "l")
> abline(linear_model_ice, col="red", lw=2 )
```



As it is a time series data, from the scatter plot we can't observe a trend so we plot another graph in which we connect the dots in order to see if there is a trend. From the above plot we can see that there is a sinusoidal pattern. We can observe a seasonal trend, during winters the sea ice extent is greater than during summers.

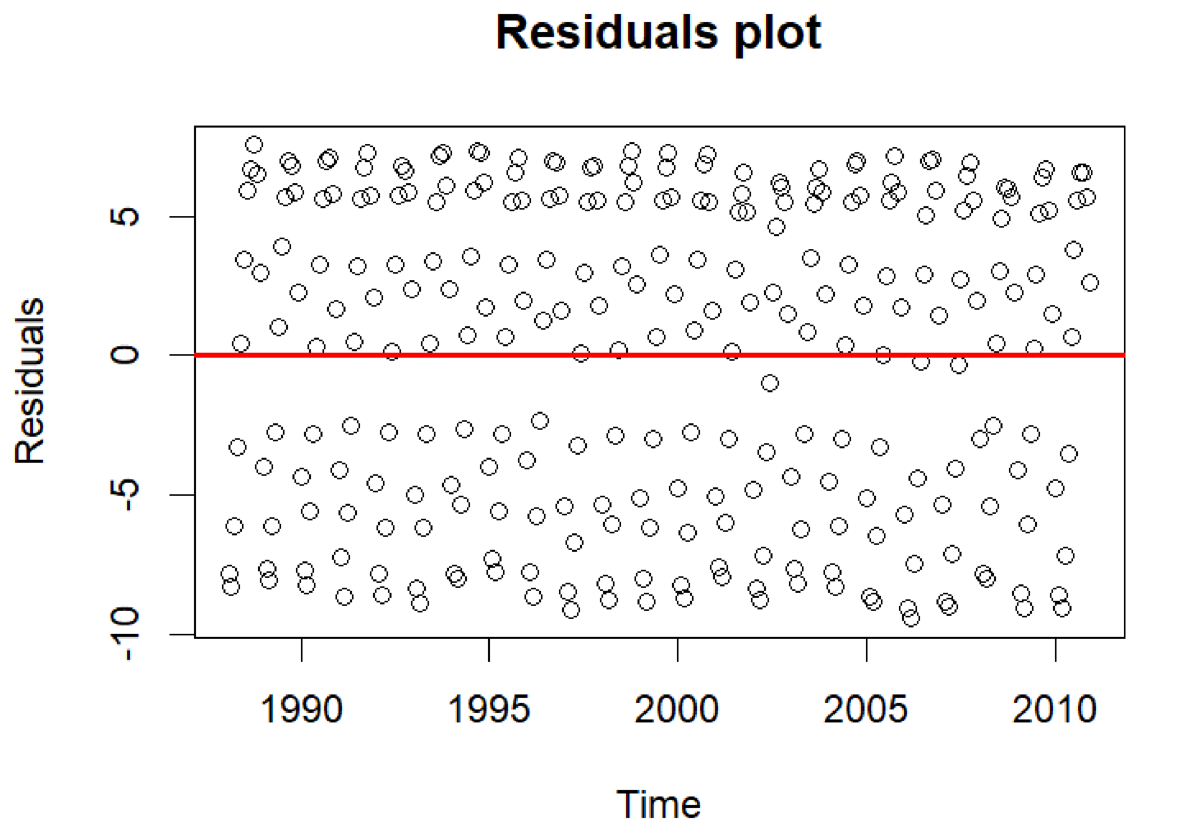
c. Plot the residuals of the model over time and include a horizontal line. What assumption(s) about the linear model should we be concerned about?

Code:

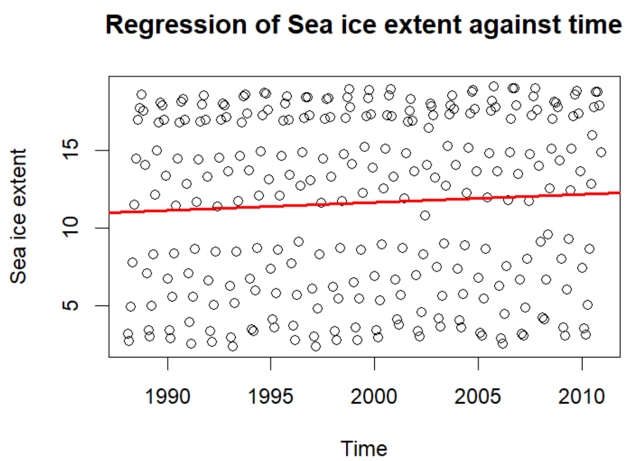
```
#2C
##Plot the residuals of the model over time and include a horizontal line.
plot(linear_model_ice$residuals ~ ice$Date,xlab="Time", ylab= "Residuals",
     main= "Residuals plot")
abline(a=0, b=0, col= "red", lw=2)
```

Output:

```
> plot(linear_model_ice$residuals ~ ice$Date,xlab="Time", ylab= "Residuals
", main= "Residuals plot")
> abline(a=0, b=0, col= "red", lw=2
```

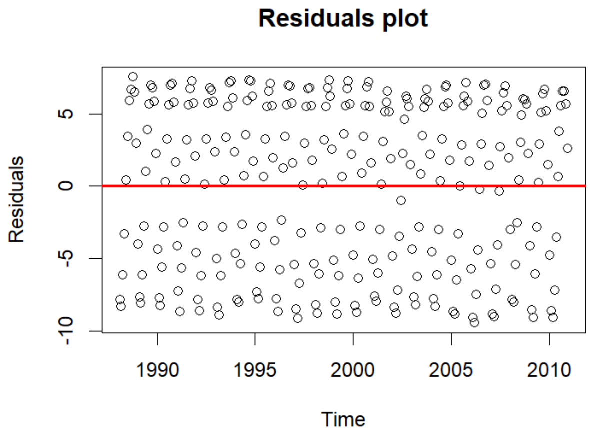


Linearity: I believe that the linearity condition is not met. Based on the sea ice extent vs time plot, the two variables do not have a linear relationship since most of the data is spread away from the regression line.
Answered using the graph below:

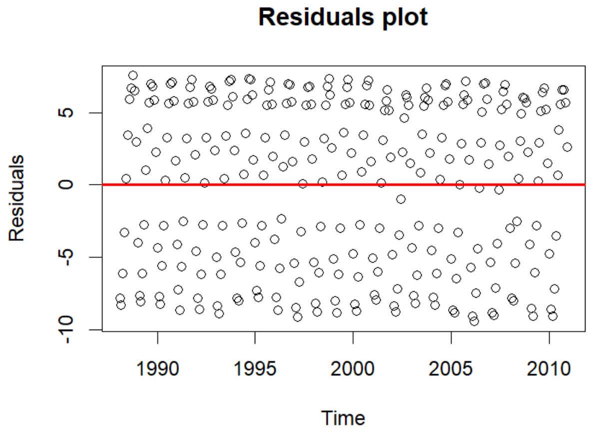


Symmetry: Based on the residual plot, it seems like the residuals are not scattered symmetrically across the x-axis (positive and negative residuals do not line up with one another). Thus, the symmetry condition has not been reasonably met.

Answered using the graph:



Equal variance: Looking at the sea ice extent vs time plot, it looks like the points are randomly scattered. Also, in the residual plot, we can see that the residuals spread out throughout. As we traverse along the x-axis, we see that mostly the residual variance remains similar. This indicates that the variance is mostly equal across all values of the explanatory variable.



Exercise 3

One of Adam's favorite casino games is called "Craps". In the first round of this game, two fair 6-sided dice are rolled. If the sum of the two dice equal 7 or 11, Adam doubles his money! If a 2, 3, or 12 are rolled, Adam loses all the money he bets.

a. 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?

When you roll 2 dice, there are 36 unique combinations.

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

Possible Outcomes that add up to 7:

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

Possible Outcomes that add up to 11:

(5,6), (6,5): 2 combinations

In total:

$$\frac{2+6}{36} = \frac{8}{36} = \frac{2}{9}$$

The probability that Adam will double his money in the first round of the game is 2/9= 0.222.

Possible Outcomes that add up to 2:

(1,1): 1 combination

Possible Outcomes that add up to 3:

(1,2), (2,1): 2 combinations

Possible Outcomes that add up to 12:

(6,6): 1 combination

In total:

$$\frac{1+2+1}{36} = \frac{4}{36} = \frac{1}{9}$$

The probability that Adam will lose his money in the first round of the game is 1/9= 0.111.

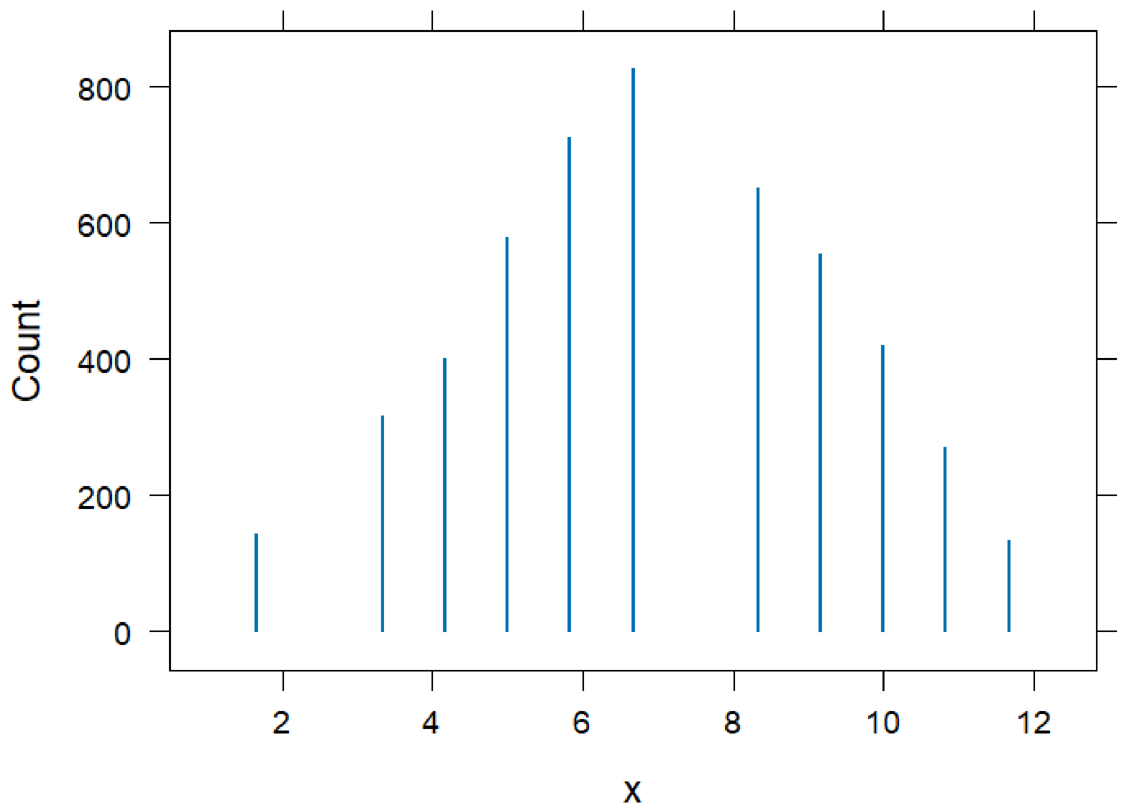
b. Let's now approximate the results in (a) by simulation. First, set the seed to 123. Then, create an object that contains 5,000 sample first round Craps outcomes (simulate the sum of 2 dice, 5,000 times). Use the appropriate function to visualize the distribution of these outcomes (hint: are the outcomes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

Code:

```
#3b
die_vals <-c(1,2,3,4,5,6)
set.seed(123)
round_dice <- replicate(5000, sample(die_vals, 2, replace= TRUE))
round_sums <- colSums(round_dice)

#use any plot that is appropriate for the data
dotPlot(round_sums, cex=5)
```

Output:

```
> #3b
> die_vals <-c(1,2,3,4,5,6)
> set.seed(123)
> round_dice <- replicate(5000, sample(die_vals, 2, replace= TRUE))
> round_sums <- colSums(round_dice)
>
> #use any plot that is appropriate for the data
> dotPlot(round_sums, cex=5)
```



The data is discrete.

c. 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.
code:

```
round_sums %in% c(7,11)

#if we take a mean of true or false vector it will give the proportion
#now we find the proportion of TRUEs to get the empirical probability

mean (round_sums %in% c(7,11))
```

Output:

```
round_sums %in% c(7,11)
   [1] FALSE FALSE FALSE FALSE FALSE
   [6]  TRUE FALSE FALSE FALSE FALSE
  [11] FALSE FALSE FALSE FALSE FALSE
  [16] FALSE FALSE  TRUE FALSE FALSE
  [21] FALSE FALSE  TRUE  TRUE FALSE
  [26]  TRUE FALSE FALSE  TRUE FALSE
  [31] FALSE FALSE FALSE  TRUE FALSE
  [36] FALSE FALSE  TRUE FALSE FALSE
  [41]  TRUE FALSE FALSE  TRUE FALSE
  [46]  TRUE FALSE  TRUE FALSE FALSE
  [51] FALSE FALSE FALSE FALSE  TRUE
  [56] FALSE FALSE FALSE FALSE  TRUE
  [61] FALSE FALSE  TRUE FALSE FALSE
  [66] FALSE FALSE  TRUE FALSE  TRUE
  [71]  TRUE FALSE FALSE FALSE FALSE
  [76]  TRUE FALSE FALSE  TRUE FALSE
  [81]  TRUE  TRUE FALSE FALSE  TRUE
  [86]  TRUE FALSE FALSE  TRUE FALSE
  [91] FALSE FALSE FALSE FALSE FALSE
  [96] FALSE FALSE FALSE FALSE FALSE
 [101]  TRUE FALSE FALSE FALSE FALSE
 [106] FALSE FALSE FALSE FALSE FALSE
 [111] FALSE FALSE FALSE  TRUE FALSE
 [116] FALSE FALSE FALSE FALSE FALSE
 [121] FALSE  TRUE  TRUE FALSE FALSE
 [126] FALSE FALSE FALSE  TRUE FALSE
 [131] FALSE FALSE FALSE FALSE FALSE
 [136] FALSE FALSE FALSE FALSE FALSE
 [141]  TRUE FALSE  TRUE  TRUE FALSE
 [146] FALSE  TRUE FALSE FALSE FALSE
 [151] FALSE  TRUE  TRUE FALSE FALSE
 [156] FALSE FALSE FALSE FALSE FALSE
 [161]  TRUE  TRUE FALSE FALSE  TRUE
 [166] FALSE FALSE FALSE  TRUE FALSE
 [171]  TRUE FALSE FALSE FALSE FALSE
 [176] FALSE FALSE FALSE FALSE FALSE
 [181]  TRUE FALSE FALSE FALSE FALSE
 [186] FALSE FALSE FALSE FALSE FALSE
 [191] FALSE FALSE FALSE FALSE  TRUE
 [196] FALSE FALSE FALSE FALSE FALSE
 [201] FALSE FALSE FALSE FALSE FALSE
 [206] FALSE FALSE FALSE FALSE FALSE
 [211] FALSE FALSE FALSE FALSE FALSE
 [216] FALSE FALSE FALSE FALSE FALSE
 [221] FALSE FALSE FALSE FALSE FALSE
 [226] FALSE FALSE  TRUE  TRUE FALSE
 [231] FALSE FALSE FALSE FALSE FALSE
 [236] FALSE FALSE FALSE FALSE FALSE
 [241] FALSE FALSE FALSE  TRUE FALSE
 [246] FALSE FALSE FALSE  TRUE FALSE
```

```
[251] FALSE FALSE  TRUE FALSE FALSE
[256]  TRUE FALSE FALSE  TRUE FALSE
[261] FALSE  TRUE FALSE FALSE FALSE
[266] FALSE FALSE  TRUE FALSE FALSE
[271] FALSE FALSE FALSE FALSE  TRUE
[276] FALSE  TRUE  TRUE FALSE FALSE
[281] FALSE  TRUE FALSE FALSE FALSE
[286] FALSE FALSE FALSE FALSE FALSE
[291] FALSE FALSE FALSE FALSE FALSE
[296] FALSE FALSE FALSE FALSE FALSE
[301] FALSE FALSE FALSE FALSE  TRUE
[306]  TRUE  TRUE FALSE FALSE FALSE
[311] FALSE FALSE FALSE FALSE FALSE
[316] FALSE  TRUE  TRUE FALSE FALSE
[321] FALSE FALSE FALSE FALSE FALSE
[326] FALSE FALSE FALSE FALSE FALSE
[331] FALSE FALSE  TRUE FALSE FALSE
[336] FALSE  TRUE FALSE FALSE  TRUE
[341]  TRUE FALSE FALSE FALSE  TRUE
[346] FALSE  TRUE  TRUE FALSE FALSE
[351] FALSE FALSE  TRUE FALSE FALSE
[356] FALSE FALSE FALSE  TRUE  TRUE
[361]  TRUE FALSE  TRUE FALSE FALSE
[366] FALSE FALSE FALSE  TRUE  TRUE
[371] FALSE FALSE FALSE FALSE FALSE
[376] FALSE FALSE FALSE  TRUE FALSE
[381] FALSE FALSE FALSE FALSE FALSE
[386] FALSE  TRUE FALSE FALSE FALSE
[391] FALSE FALSE FALSE  TRUE FALSE
[396] FALSE  TRUE  TRUE FALSE  TRUE
[401] FALSE FALSE FALSE FALSE FALSE
[406]  TRUE FALSE FALSE  TRUE  TRUE
[411] FALSE FALSE  TRUE FALSE FALSE
[416] FALSE FALSE FALSE  TRUE FALSE
[421] FALSE FALSE FALSE FALSE FALSE
[426] FALSE  TRUE FALSE FALSE FALSE
[431] FALSE FALSE FALSE FALSE FALSE
[436]  TRUE FALSE  TRUE FALSE FALSE
[441] FALSE FALSE FALSE FALSE FALSE
[446] FALSE FALSE FALSE FALSE FALSE
[451] FALSE FALSE FALSE FALSE FALSE
[456] FALSE FALSE FALSE FALSE FALSE
[461] FALSE  TRUE  TRUE FALSE FALSE
[466] FALSE FALSE FALSE  TRUE FALSE
[471] FALSE FALSE FALSE  TRUE FALSE
[476] FALSE FALSE FALSE FALSE FALSE
[481] FALSE FALSE  TRUE  TRUE  TRUE
[486] FALSE FALSE FALSE FALSE FALSE
[491] FALSE FALSE  TRUE FALSE FALSE
[496] FALSE FALSE FALSE FALSE FALSE
[501]  TRUE  TRUE  TRUE FALSE FALSE
[506] FALSE FALSE FALSE FALSE FALSE
[511] FALSE FALSE  TRUE FALSE FALSE
[516] FALSE  TRUE FALSE FALSE FALSE
[521]  TRUE FALSE  TRUE FALSE  TRUE
[526] FALSE FALSE  TRUE FALSE FALSE
[531]  TRUE FALSE FALSE  TRUE FALSE
[536]  TRUE FALSE FALSE FALSE FALSE
[541] FALSE FALSE  TRUE FALSE  TRUE
[546] FALSE FALSE  TRUE FALSE FALSE
[551] FALSE FALSE FALSE FALSE FALSE
[556] FALSE FALSE FALSE FALSE  TRUE
[561] FALSE FALSE FALSE FALSE FALSE
[566]  TRUE  TRUE FALSE FALSE FALSE
[571] FALSE FALSE FALSE FALSE FALSE
[576] FALSE  TRUE FALSE FALSE FALSE
[581]  TRUE FALSE FALSE FALSE  TRUE
[586]  TRUE  TRUE  TRUE FALSE FALSE
[591] FALSE FALSE  TRUE FALSE FALSE
```

```
[596] FALSE  TRUE FALSE FALSE  TRUE
[601] FALSE FALSE  TRUE FALSE FALSE
[606] FALSE FALSE FALSE FALSE FALSE
[611] FALSE FALSE  TRUE  TRUE FALSE
[616] FALSE FALSE FALSE  TRUE FALSE
[621] FALSE FALSE FALSE FALSE FALSE
[626]  TRUE FALSE FALSE FALSE FALSE
[631] FALSE FALSE FALSE  TRUE  TRUE
[636] FALSE FALSE FALSE FALSE FALSE
[641] FALSE  TRUE FALSE FALSE FALSE
[646] FALSE FALSE  TRUE FALSE  TRUE
[651] FALSE FALSE FALSE FALSE FALSE
[656] FALSE FALSE  TRUE  TRUE FALSE
[661] FALSE FALSE FALSE  TRUE FALSE
[666] FALSE  TRUE FALSE FALSE FALSE
[671] FALSE FALSE FALSE FALSE FALSE
[676]  TRUE FALSE  TRUE FALSE FALSE
[681] FALSE FALSE  TRUE FALSE  TRUE
[686] FALSE FALSE FALSE FALSE FALSE
[691]  TRUE FALSE FALSE FALSE  TRUE
[696] FALSE FALSE FALSE FALSE FALSE
[701] FALSE FALSE FALSE  TRUE FALSE
[706] FALSE FALSE FALSE FALSE  TRUE
[711] FALSE FALSE  TRUE  TRUE FALSE
[716]  TRUE FALSE FALSE  TRUE FALSE
[721] FALSE  TRUE FALSE FALSE FALSE
[726] FALSE FALSE FALSE FALSE FALSE
[731] FALSE FALSE FALSE  TRUE  TRUE
[736] FALSE FALSE  TRUE FALSE FALSE
[741]  TRUE FALSE FALSE FALSE FALSE
[746] FALSE FALSE  TRUE FALSE FALSE
[751] FALSE  TRUE FALSE FALSE FALSE
[756]  TRUE FALSE FALSE FALSE FALSE
[761] FALSE FALSE FALSE FALSE FALSE
[766] FALSE FALSE  TRUE FALSE FALSE
[771] FALSE FALSE FALSE  TRUE FALSE
[776] FALSE FALSE FALSE FALSE FALSE
[781]  TRUE FALSE FALSE FALSE FALSE
[786] FALSE FALSE FALSE FALSE FALSE
[791] FALSE FALSE FALSE FALSE  TRUE
[796] FALSE FALSE FALSE FALSE FALSE
[801]  TRUE FALSE FALSE FALSE FALSE
[806] FALSE FALSE FALSE FALSE FALSE
[811] FALSE  TRUE  TRUE FALSE  TRUE
[816]  TRUE FALSE FALSE FALSE FALSE
[821] FALSE  TRUE FALSE FALSE FALSE
[826] FALSE FALSE FALSE FALSE  TRUE
[831]  TRUE  TRUE FALSE FALSE  TRUE
[836]  TRUE FALSE FALSE FALSE FALSE
[841] FALSE FALSE  TRUE FALSE FALSE
[846] FALSE FALSE FALSE FALSE FALSE
[851] FALSE FALSE FALSE FALSE FALSE
[856] FALSE FALSE FALSE FALSE FALSE
[861]  TRUE FALSE FALSE FALSE FALSE
[866] FALSE  TRUE FALSE FALSE FALSE
[871]  TRUE FALSE FALSE FALSE  TRUE
[876] FALSE FALSE FALSE FALSE FALSE
[881] FALSE  TRUE FALSE FALSE FALSE
[886] FALSE FALSE FALSE FALSE FALSE
[891] FALSE FALSE  TRUE  TRUE FALSE
[896]  TRUE FALSE FALSE FALSE FALSE
[901] FALSE  TRUE FALSE FALSE FALSE
[906] FALSE FALSE FALSE FALSE  TRUE
[911]  TRUE  TRUE  TRUE FALSE FALSE
[916]  TRUE FALSE FALSE FALSE FALSE
[921]  TRUE FALSE FALSE FALSE FALSE
[926]  TRUE FALSE FALSE FALSE FALSE
[931]  TRUE FALSE  TRUE FALSE FALSE
[936] FALSE FALSE  TRUE FALSE FALSE
```

```
 [941] FALSE FALSE FALSE FALSE  TRUE
 [946] FALSE FALSE FALSE FALSE FALSE
 [951] FALSE FALSE FALSE FALSE  TRUE
 [956]  TRUE FALSE FALSE FALSE FALSE
 [961] FALSE FALSE FALSE FALSE FALSE
 [966]  TRUE FALSE FALSE FALSE FALSE
 [971] FALSE FALSE  TRUE FALSE FALSE
 [976] FALSE FALSE FALSE FALSE FALSE
 [981] FALSE FALSE FALSE  TRUE FALSE
 [986]  TRUE FALSE FALSE FALSE FALSE
 [991] FALSE FALSE FALSE FALSE FALSE
 [996]  TRUE FALSE FALSE  TRUE  TRUE
 [ reached getOption("max.print") -- omitted 4000 entries ]
> mean (round_sums %in% c(7,11))
[1] 0.2188
```

The percentage of time Adam doubled his money: 21.88%

Now to calculate the percentage of time Adam lost his money:

Code:

```
round_sums %in% c(2,3,12)
mean (round_sums %in% c(2,3,12))
```

Output:

```
> round_sums %in% c(2,3,12)
   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE  TRUE
  [14] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
  [27]  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE F
ALSE FALSE
  [40] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
  [53] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE F
ALSE FALSE
  [66] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE F
ALSE FALSE
  [79] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE  TRUE
  [92]  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [105]  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE F
ALSE  TRUE
 [118] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [131] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE F
ALSE FALSE
 [144] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
TRUE FALSE
 [170] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [183] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
TRUE FALSE
 [196] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [209] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE F
ALSE FALSE
```

```
 [222]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE
TRUE FALSE
 [235] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [248] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [261] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
TRUE FALSE
 [274] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE F
ALSE  TRUE
 [287]  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [300] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE F
ALSE  TRUE
 [313] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE F
ALSE FALSE
 [326] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [339] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [352] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE  TRUE
 [365] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE F
ALSE  TRUE
 [378] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [391] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [404] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE F
ALSE FALSE
 [417] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [430]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [443] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
TRUE FALSE
 [456] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
TRUE FALSE
 [469] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [482]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [495] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [508] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE F
ALSE  TRUE
 [521] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
TRUE FALSE
 [534] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [547] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE F
ALSE FALSE
 [560] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE F
ALSE FALSE
 [573] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE F
ALSE FALSE
 [586] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [599] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [612]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
TRUE  TRUE
 [625] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE F
ALSE FALSE
 [638]  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [651] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
TRUE FALSE
```

```
 [664] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [677] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [690] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
TRUE FALSE
 [703] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE F
ALSE  TRUE
 [716] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE F
ALSE FALSE
 [729] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE F
ALSE FALSE
 [742] FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [755] FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE F
ALSE  TRUE
 [768] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
TRUE FALSE
 [781] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [794] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [807] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [820] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [833] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [846] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [859] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [872] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE F
ALSE FALSE
 [885] FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE F
ALSE FALSE
 [898] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE F
ALSE FALSE
 [911] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE F
ALSE FALSE
 [924] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [937] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE F
ALSE FALSE
 [950] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE FALSE
 [963] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE  TRUE
 [976] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE F
ALSE  TRUE
 [989] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE F
ALSE
 [ reached getOption("max.print") -- omitted 4000 entries ]
> mean (round_sums %in% c(2,3,12))
[1] 0.1172
```

The percentage of time Adam lost his money: 11.72 %

d. Adam winning money and Adam losing money can both be considered events. Are these two events independent, disjoint, or both? Explain why.

If two events are independent, then information about the first event will not impact our belief in whether or not the second event will occur. Since winning money and losing money cannot happen simultaneously for Adam, the aforementioned events are not independent.
Let P(A) be the Probability of winning.
Let P(B) be the probability of losing.
Two events are independent if P(B|A) = P(B).
Here, P(B|A) = 0, while P(B) > 0. Since P(B|A) != P(B), these two events are not independent. They are not independent events since the occurrence of one event affects the chances of the occurrence of the other event.

Two events are called disjoint or mutually exclusive if P(B|A)= 0= P(A|B). We know that it is impossible for Adam to lose and win his money at the same time as the dice cannot sum up to two different numbers at the same time, so P(A and B)= 0 which implies: P(B|A)= 0= P(A|B). Hence, the events are disjoint.

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

Let P(A) be the Probability of winning.
Let P(B) be the probability of losing.
If two events are independent, P(A and B) = P(A) * P(B)
From d we know, P(A and B) = 0
P(A) * P(B) = 2/9 * 1/9 = 2/81
P(A and B) != P(A) * P(B)
Therefore, these two events are not independent.

PART II

Stats 10- Assignment 3.

Exercise 1

$P(A) = 0.32$ $P(B) = 0.21$ $P(C) = 0.23$

a) The probability of getting A or B can be found by adding the invidual probabilities of A and B as these events are mutually exclusive. This means that the student can not get both A and B at the same time.

$$P(A \cup B) = P(A) + P(B)$$
$$= 0.32 + 0.21$$
$$= 0.53.$$

$\therefore$ the probability that the student will get A or a B is 0.53.

b) Similar to part A, as the events as mutually exclusive we can find the probability by adding the individual probabilities.

$\therefore$ Probability that a student will get an A or a B or a C is:

$$P(A) + P(B) + P(C)$$
$$= 0.32 + 0.21 + 0.23$$
$$= 0.76.$$

c) We can find this by taking a complement of getting a C or higher. We subtract it from 1 because sum of all possible outcomes should equal 1.

$\therefore$ probability that a student will get a grade lower than a C is : ~~1-0~~

$$1 - 0.76$$
$$= 0.24$$

Exercise 2)

a) Let E be the event of getting at least one six. To find ~~the~~ it we can first find the probability of getting no six which is its complement and then subtract it from one as the sum of probability of all possible outcomes should be 1.

*

- Probability of getting no six in four rolls of a single die:

$$\frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$$

$$= \frac{625}{1296} = 0.4822253$$

- Probability of getting atleast one six ~~#~~ in four rolls of a single die:

$$P(E) = 1 - \frac{625}{1296} = \frac{671}{1296}$$

$$= \sim 0.518$$

$P(E) = 1 - P(E^c) = 1 - P(A1 \text{ and } A2 \text{ and } A3 \text{ and } A4)$

$= 1 - (5/6)^{\wedge 4}$

Let A_i be the event of getting no six in ith roll.

$P(A_i) = \frac{5}{6}$. We know that rolls are independent so using the multiplication rule.

b) let F be the event of getting atleast one doublesix in 24 throws of a pair of dice.

let Ai be the event - no double six is thrown on the ith roll.

$$Ai = \frac{35}{36}$$

F = A1 and A2 and A3. . . . and A24

$P(F) = 1 - P(F^c)$

$P(F^c)$ is the event of no double six in 24 throws.

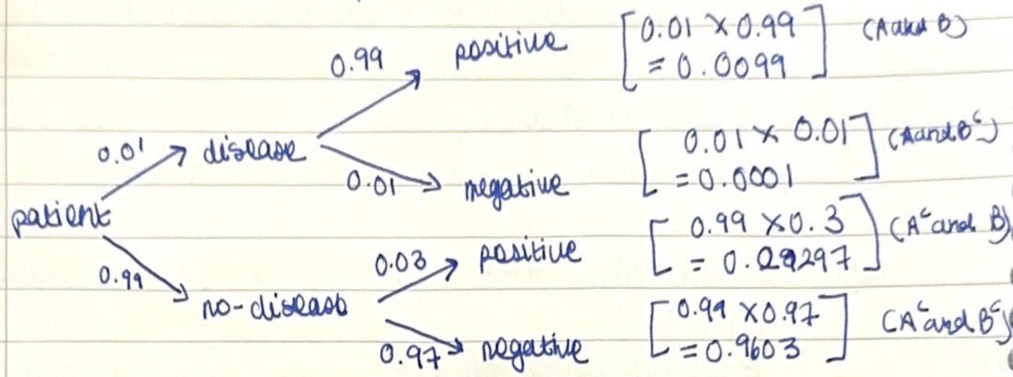
$$P(F^c) = \frac{35}{36}^{\wedge 24}$$

$$P(F) = 1 - \frac{35}{36}^{\wedge 24}$$

$$= \sim 0.491$$

$P(B)$ = patient test +ve
$P(B^c)$ = Patient test -ve.
$P(A)$ patient has disease
$P(A^c)$ patient does not have disease

<u>Exercise 8</u>

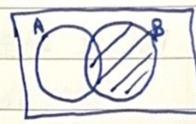


we need to find $P(A|B)$

$P(\text{positive}) = P(\text{disease} + \text{positive}) + P(\text{non disease} + \text{positive})$

$P(B) = P(A \text{ and } B) + P(A^c \text{ and } B) = 0.0099 + 0.0297$

$= 0.0396$



$P(A|B)$ ~~P(disease & positive)~~ $= \dfrac{P(\text{disease} + \text{positives})}{P(\text{positives})}$

$\boxed{P(A|B) = \dfrac{P(A \cap B)}{P(B)}}$

$= \dfrac{0.0099}{0.0396}$

$= 0.25 = 25\%$

∴, if the test result comes back positive then the probability the patient actually has the disease is 25%.

Exercise 4)

a) Theoretical probability of getting heads:

as there are 2 equally likely outcomes:

$$P(\text{head}) = \frac{1}{2} = 0.5$$

empirical probability of getting heads:

$$\frac{\text{~~proportion~~ number of heads observed in the experiment}}{\text{total num of outcoms}}$$

$$= \frac{58}{100} = 0.58$$

b) Theoretical probability of getting tails:

similar to part a,

$$P(\text{tails}) = \frac{1}{2} = 0.5$$

empirical probability of getting tails:

$$\frac{\text{num of tails observed}}{\text{total number of outcomes}} = \frac{42}{100} = 0.42$$

c) If we were to flip the coin 1000 times and record the proportion of times we get head, we would expect the empirical probability to be close to theoretical probability of 0.5 or $\frac{1}{2}$ or 50% due to law of large numbers. This is because as the number of coin flips inc, the empirical probability tends to converge to the theoretical probability. So the observed probability would approach 0.5 as we perform more and more trials in line with the true theoretical probability of the fair coin.

d) Empirical probabilities are frequently used in various real life situations. eg in weather forecasting, meteorologists use historical weather data to calculate the empirical probabilities for events like rain, snow or temp patterns. Empirical probabilities are valuable in decision making when historical data is available to inform future expectations.

Exercise 5)

empirical probability is $= \frac{\text{possible outcomes}}{\text{total number of outcomes}}$

we consider empirical probability as the probability found from the experiment is empiracal.

a) For the first 20 trials, the probabity of rolling a 4 is:

From the table the num of possible outcomes is 2

so ~~P~~ P(rolling a 4 for 20 trials) $= \frac{2}{20} = \frac{1}{10} = 0.1$

b) For the ~~first~~ 100 trials:

From the table the num of possible outcoms is 20 so

P(rolling a 4 for 100 trials) $= \frac{20}{100} = \frac{1}{5} = 0.2$

c) For 1000 trials

from the table the num of possible outcomes is 166 so

P(rolling a 4 for 1000 trials) $= \frac{166}{1000} = 0.166$

d) theoretical probability of rolling a 4 with a fair six sided die:

~~we~~ using the equally likely outcomes formula:

$P(A) = \frac{\text{number of outcomes in A}}{\text{number of all possible outcomes in S}}$

$= \frac{1}{6} = 0.167$

e) In this experiment we can observe that as the total number of trials increase the empirical probability of becomes closer to theoretical probability of a fair dice 0.167 as stated in law of large numbers. This is because as the number of trials increase the empirical probability tends to converge to theoretical probability. ~~For 20 trials~~ the empirical probability for 20 trials was 0.1, for 100 trial was 0.2 and for 1000 trials was 0.166 which clearly supports the above facts