Assignment 1 Solutions

Abhishek Devarajan

2023-11-19

Part 1

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".

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.

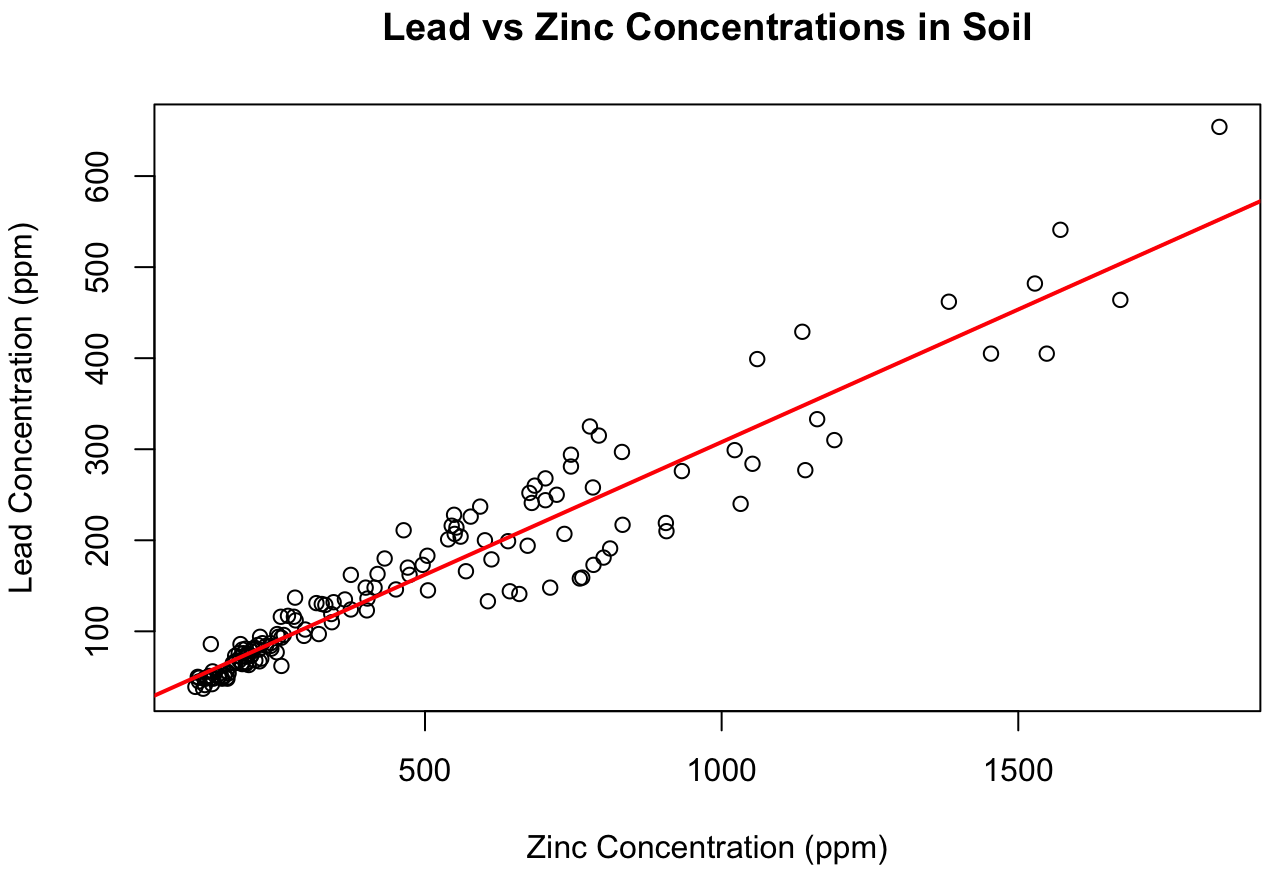
```
#The file path will be different depending on your working directory
soil <- read.table("../soil_complete.txt", header = TRUE)

lead_zinc_model <- lm(soil$lead ~ soil$zinc)
summary(lead_zinc_model)
```

```
##
## Call:
## lm(formula = soil$lead ~ soil$zinc)
##
## 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 ***
## soil$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
```

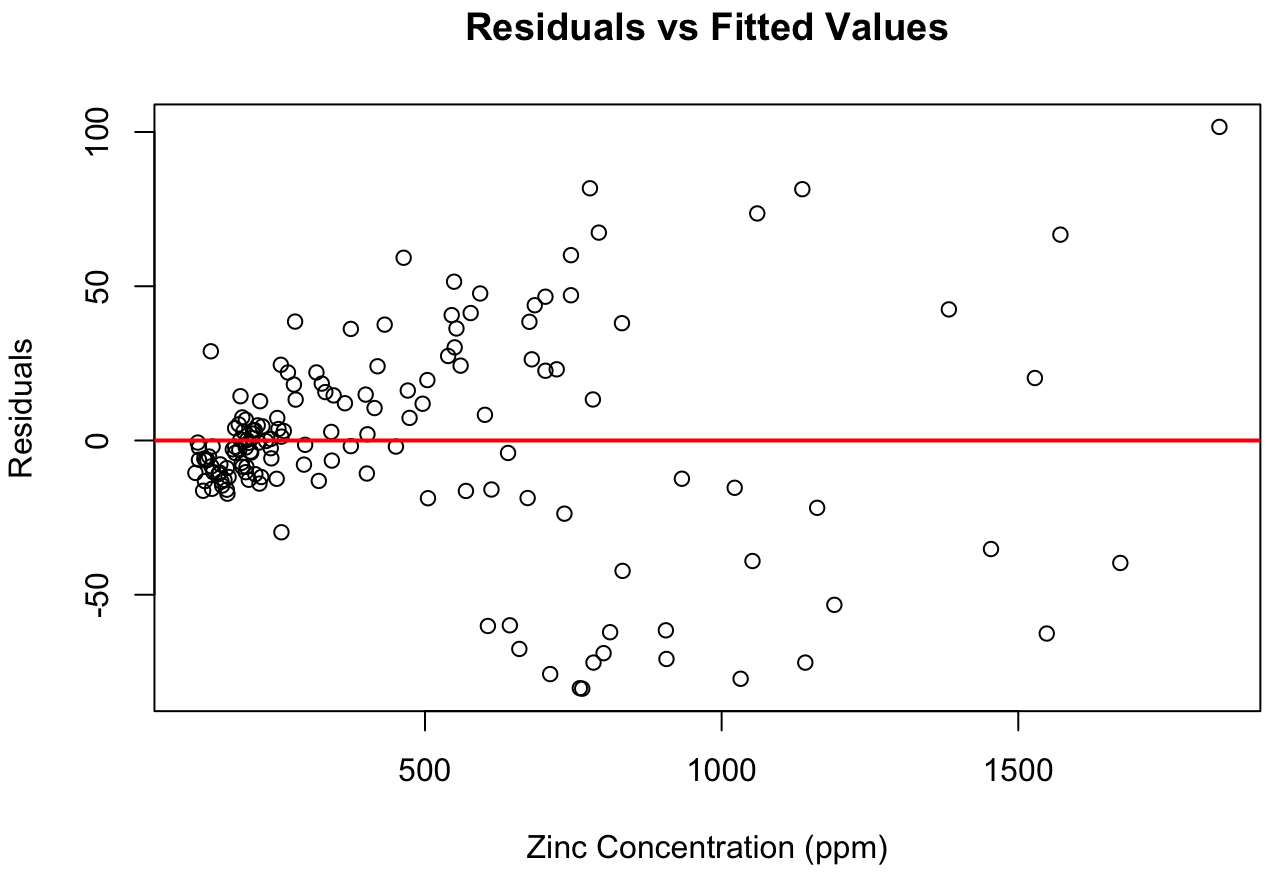
b) Plot the lead and zinc data, then use the abline() function to overlay the regression line onto the data.

```
plot(lead ~ zinc, data = soil,
     xlab = "Zinc Concentration (ppm)",
     ylab = "Lead Concentration (ppm)",
     main = "Lead vs Zinc Concentrations in Soil")
abline(lead_zinc_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.

```
plot(lead_zinc_model$residuals ~ soil$zinc,
     xlab = "Zinc Concentration (ppm)",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values")
abline(0,0, col = "red", lw = 2)
```



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

In the summary from part (a), we can see that the intercept value is approximately 16.582928. Additionally, the slope estimate for zinc is approximately 0.291335. Plugging these values into the equation of a line gives us the following equation for the regression line:

$$\text{lead} = 16.582928 + 0.291335 * \text{zinc}.$$

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?

We can plug the zinc level of this new data point into the regression equation to get the estimated lead level.

$$\begin{aligned}\text{lead} &= 16.582928 + 0.291335 * \text{zinc} \\ &= 16.582928 + 0.291335 * 1000 \\ &= 16.582928 + 291.335 \\ &= 307.9179.\end{aligned}$$

Therefore, we would expect the lead concentration at this new data point to be about 307.9 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?

First, we can set up the regression equations for site A and site B

$$\begin{aligned}\text{lead}_A &= 16.582928 + 0.291335 * \text{zinc}_A \\ \text{lead}_B &= 16.582928 + 0.291335 * \text{zinc}_B.\end{aligned}$$

Now, we can use the fact that the zinc at A is 100 ppm higher than the zinc at B to get

$$\begin{aligned}\text{lead}_A &= 16.582928 + 0.291335 * (\text{zinc}_B + 100) \\ \text{lead}_B &= 16.582928 + 0.291335 * \text{zinc}_B.\end{aligned}$$

Finally, we can subtract the two equations to get the difference in expected lead levels

$$\begin{aligned}\text{lead}_A - \text{lead}_B &= 16.582928 + 0.291335 * (\text{zinc}_B + 100) - (16.582928 + 0.291335 * \text{zinc}_B) \\ &= 0.291335 * (\text{zinc}_B + 100) - (0.291335 * \text{zinc}_B) \\ &= 0.291335 * 100 \\ &= 29.1335.\end{aligned}$$

Based on the calculations above, we would expect the lead concentration at site A to be 29.1335 ppm higher than the lead concentration at site B.

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

Based on the summary in part (a), the R^2 value is 0.912. This means that 91.2% of the variation in soil lead concentration can be explained by the zinc concentration in the soil.

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.

Based on the plot of the data in part (b), it is pretty clear that lead and zinc have a linear relationship. As zinc increases, lead increases at a constant rate, which means that the linearity assumption is met. When we look at the residual plot in part (c), it seems like the residuals are roughly symmetric for lower values of zinc. At higher zinc levels, it seems like there are bigger positive residuals than negative ones. Overall, the residuals are mostly symmetric so I believe this assumption has been met. Finally, looking at the residual plot in part (c) shows us that the residuals are very clustered around 0 for low zinc levels, and they gradually fan outwards from the x-axis as zinc increases. This “fan” pattern indicates that the variance is not constant for all values of zinc. For this reason, I believe the equal variance assumption has not been met.

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.

```
ice <- read.csv("../sea_ice.csv", header = TRUE)

#Changing the date column from characters to dates

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

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

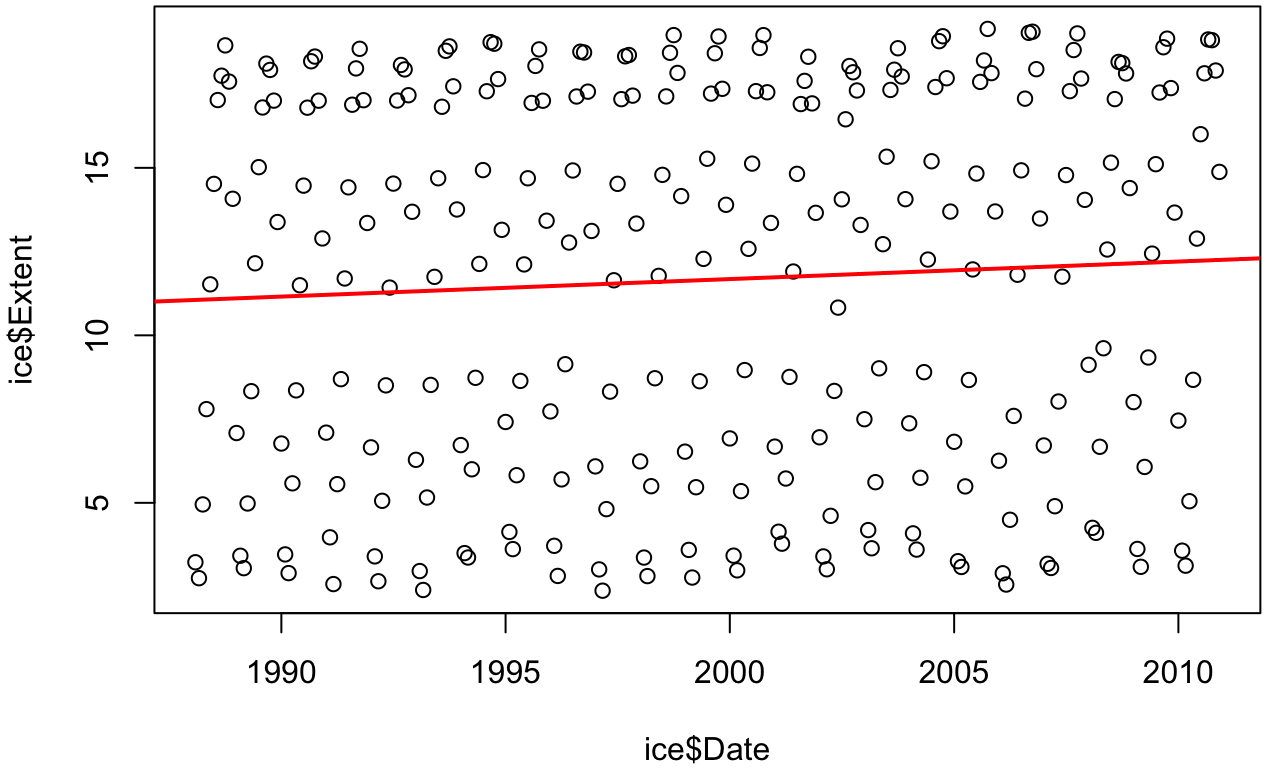
```
ice_model <- lm(Extent ~ Date, data = ice)

summary(ice_model)
```

```
##
## 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?

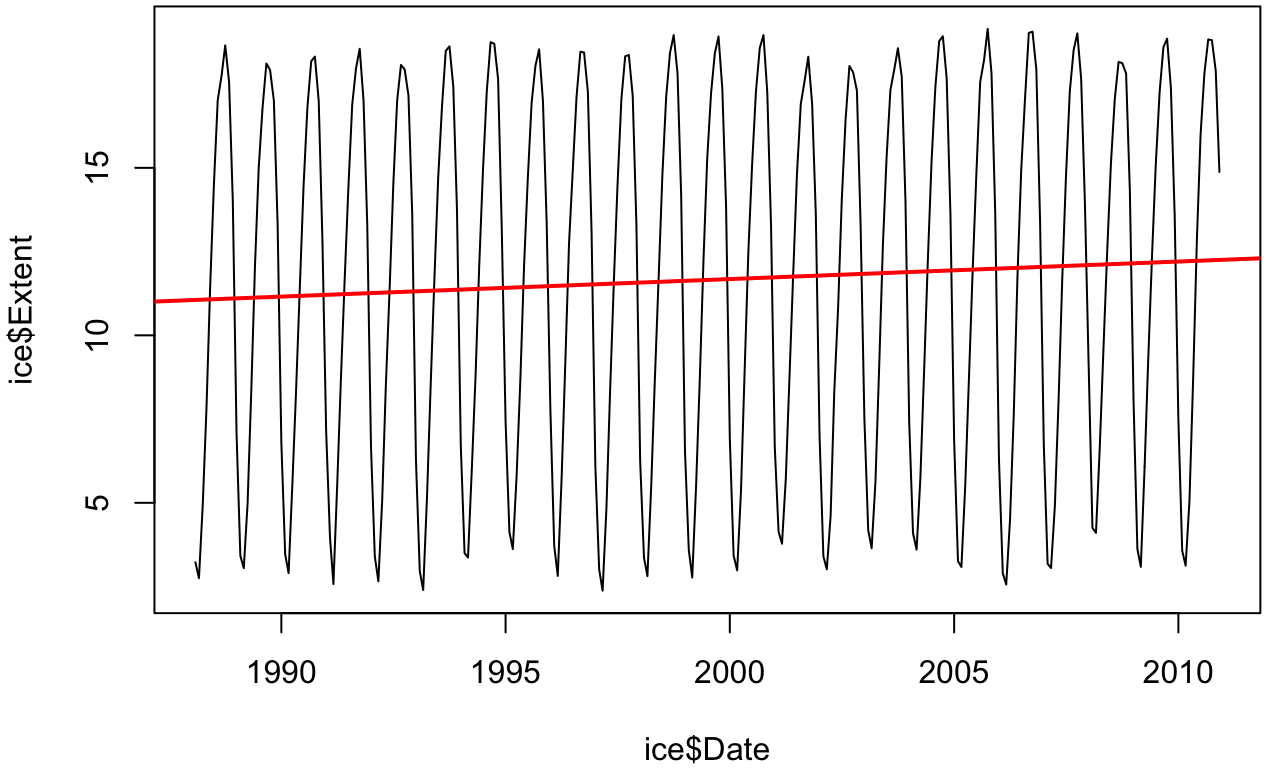
```
plot(ice$Date, ice$Extent)
abline(ice_model, lw = 2, col = "red")
```



Based on the scatter plot above, it is unclear if there are any trends in the data. In order to see the trend in the data, we have to change from a scatter plot to a line plot. This is because the data we have is sequential (the data is ordered by time) so the order of the data points matters. When we connect the dots sequentially, the pattern appears to be cyclical. The graph below shows this pattern clearly.

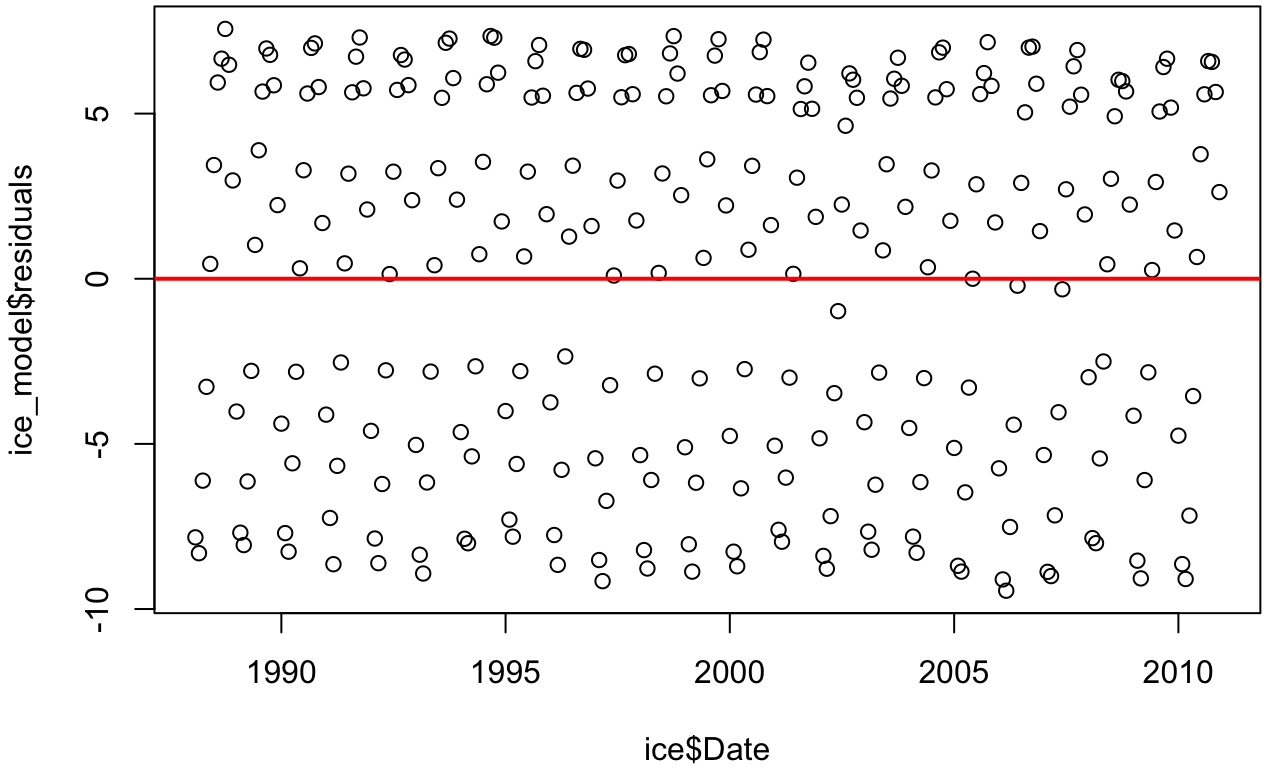
Note: It is okay to say that the pattern is seasonal, cyclical, sinusoidal or any other synonym.

```
plot(ice$Date, ice$Extent, type = "l")
abline(ice_model, lw = 2, col = "red")
```



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?

```
plot(ice_model$residuals ~ ice$Date)
abline(0,0,col="red",lw=2)
```



Based on the residuals, it looks like the symmetry and equal variance assumptions are met fairly well. The residuals are not clustered or fanned out at any points and they are not overly skewed towards the negative or positive side. However, based on the plot in part (b) it is clear that the data does not follow a linear relationship. Thus, the linearity assumption is not met.

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?

Let's start with the probability that Adam doubles his money. When we toss 2 dice, there are 36 possible combinations that we can get. Of those 36 combinations, the following combinations add up to 7:

$$\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}.$$

Similarly, the following combinations add up to 11:

$$\{(5,6),(6,5)\}.$$

In total, there are 8 possible combinations that result in Adam winning the round. Thus, the probability that he doubles his money is $\frac{8}{36} = \frac{2}{9} \approx 0.222$.

Now, we can look at the probability that Adam loses his money. There is only one combination that adds up to 2:

$$\{(1,1)\}.$$

For 3, the possible combinations are

$$\{(1,2),(2,1)\}.$$

Finally, there is only one combination that adds up to 12:

$$\{(6,6)\}.$$

In total, there are 4 combinations that result in Adam losing his money. Thus, the probability that he loses his money is $\frac{4}{36} = \frac{1}{9} \approx 0.111$.

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?).

```
set.seed(123)
die_vals <- c(1,2,3,4,5,6)

dice <- replicate(5000,sample(die_vals, 2, replace = TRUE))
sums <- colSums(dice)

### Use any plot that is appropriate for the data
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

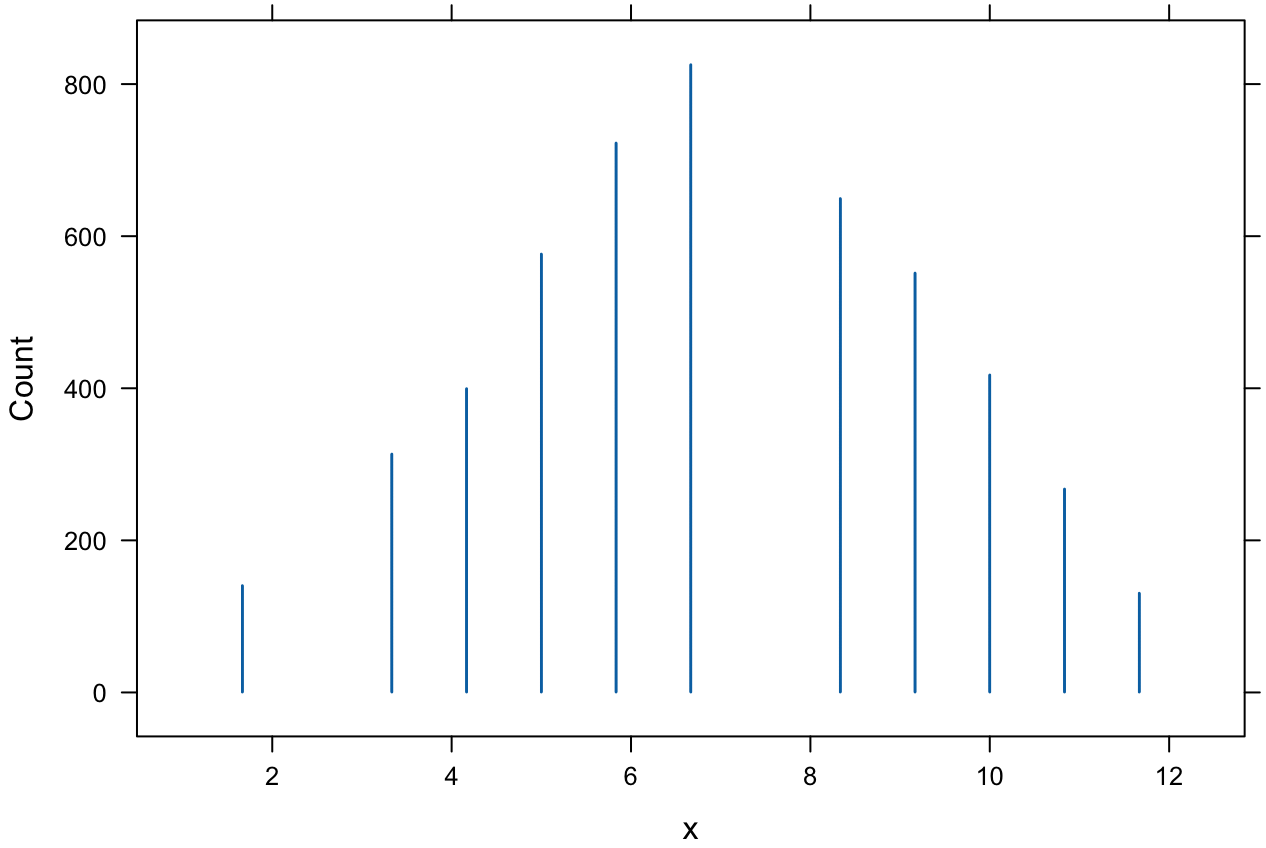
```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
dotPlot(sums, cex = 5)
```



Note: You can use any graph that is appropriate for discrete numerical data. If you use a histogram, you need to make sure that the bins are appropriately sized so there are no decimals.

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.

```
# Option 1: Check if each value in sums is in the set {7,11} or in the set {2,3,12}

## Probability of winning
mean(sums %in% c(7,11))
```

```
## [1] 0.2188
```

```
##Probability of losing
mean(sums %in% c(2,3,12))
```

```
## [1] 0.1172
```

```
# Option 2: Use an or-statement
## Probability of winning
mean(sums == 7 | sums == 11)
```

```
## [1] 0.2188
```

```
##Probability of losing
mean(sums == 2 | sums == 3 | sums == 12)
```

```
## [1] 0.1172
```

Based on the output, the empirical probability of Adam winning money is 0.2188, and the empirical probability of Adam losing money is 0.1172.

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

Let's start by determining if winning money and losing money are independent events. Intuitively, if two events are independent, then information about the first event will not impact your belief in whether or not the second event will occur. If we know that Adam doubled his money, we instantly know that he could not have lost any money. Thus, the events are not independent.

Now, let's determine whether the two events are disjoint. Two events are disjoint if they cannot both occur at the same time. In this case, the set of winning values, {7,11}, is completely distinct from the set of losing values, {2,3,12}. This means that Adam cannot win and lose at the same time. Thus, doubling his money and losing his money are disjoint events.

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

From part (d), we know that Adam winning money and Adam losing money are disjoint events. Let's label these events as A and B respectively. We know that, for disjoint events, $\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = 0$.

From part (a), we know that $\mathrm{P}(\mathrm{A}) = \frac{2}{9}$ and $\mathrm{P}(\mathrm{B}) = \frac{1}{9}$. If we multiply these values, we get $\mathrm{P}(\mathrm{A}) \cdot \mathrm{P}(\mathrm{B}) = \frac{2}{81}$. This means that

$$\mathrm{P}(\mathrm{A}) \cdot \mathrm{P}(\mathrm{B}) \neq \mathrm{P}(\mathrm{A} \cap \mathrm{B}),$$

which proves that A and B are not independent.

Part 2

Exercise 1: Assume the grades possible in a history course are A, B, C, or lower than C. The probability that a randomly selected student will get an A in the course is 0.32, the probability that a student will get a B in the course is 0.21, and the probability that a student will get a C in the course is 0.23.

a) What is the probability that a student will get an A OR a B?

Let A, B, C,and D denote the events where a student receives an A, B, C or less than a C respectively. Since a student can only receive one letter grade for a class, each of these events are disjoint.

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) \\ &= 0.32 + 0.21 \\ &= 0.53. \end{aligned}$$

Thus, the probability of getting an A or a B is 0.53.

b) What is the probability that a student will get an A OR a B OR a C?

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C) \\ &= P(A) + P(B) + P(C) \\ &= 0.32 + 0.21 + 0.23 \\ &= 0.76. \end{aligned}$$

Thus, the probability of getting an A, B, or C is 0.76.

c) What is the probability that a student will get a grade lower than a C?

The probability that a student gets a grade lower than a C is the complement of the event that a student gets an A, B, or C. Thus, we can just subtract our answer from part (b) from 1 to get the answer:

$$\begin{aligned} P(D) &= 1 - P(A \cup B \cup C) \\ &= 1 - 0.76 \\ &= 0.24. \end{aligned}$$

Thus, the probability that a student got less than a C is 0.24.

Exercise 2: De Mere's Dice Problem

a) Let E be the event of getting at least one six in four rolls of a single die. Find P(E).

The event E occurs when you roll 1,2,3, or 4 sixes. Since this is a complicated thing to calculate, we can look at the complementary event. The complement, E^c occurs when you roll no sixes. For a single roll of the die, there's a $\frac{5}{6}$ chance that you roll something other than a 6. Since each roll is independent, we can multiply $\frac{5}{6}$ by itself four times to get $P(E^c)$. Then, we can subtract that from 1 to get $P(E)$.

$$\begin{aligned} P(E) &= 1 - P(E^c) \\ &= 1 - \left(\frac{5}{6}\right)^4 \\ &\approx 0.518. \end{aligned}$$

Note: You can calcualte the probability of E directly, but you must show all relevent work and you must have the same final answer.

b) Let F be the event of getting at least one double six in 24 throws of 2 dice. Find P(F).

Again, we can break this event up by first looking at the complement of F. F^c is the event where you don't roll double sixes for any of the 24 rolls. In one roll, there are 36 possible combinations for the 2 dice. Out of those, 35 of the combinations are something other than a double six. Since each roll is independent, we know that $P(F^c) = \left(\frac{35}{36}\right)^{24}$. We can subtract that from 1 to get $P(F)$.

$$\begin{aligned} P(F) &= 1 - P(F^c) \\ &= 1 - \left(\frac{35}{36}\right)^{24} \\ &\approx 0.491. \end{aligned}$$

Exercise 3: A patient is displaying some symptoms and received a disease screening test. The test comes back positive 99% of the time for people who have the disease, and comes back negative 97% of the time for people who do not have the disease. The doctor knows that the disease affects 1 in 100 people in the country. Suppose the test result for the patient came back positive, what is the probability that the patient actually has the disease?

First, let's write out what we know using the conditional probability notation. For simplicity, I will denote D+ as the event that the patient has the disease and D- as the event where the patient does not have the disease. Analogously, I will use T+ to denote the event where the patient's test is positive, and T- for the event where the patient's test is negative.

$$\begin{aligned} \mathrm{P}(\mathrm{D}+) &= 0.01 \\ \mathrm{P}(\mathrm{T}+|\mathrm{D}+) &= 0.99 \\ \mathrm{P}(\mathrm{T}-|\mathrm{D}-) &= 0.96. \end{aligned}$$

The quantity we wish to find is $\mathrm{P}(\mathrm{D}+|\mathrm{T}+)$. Using Bayes' Rule, we know that

$$\begin{aligned} \mathrm{P}(\mathrm{T}+|\mathrm{D}+) &= \frac{\mathrm{P}(\mathrm{D}+ \cap \mathrm{T}+)}{\mathrm{P}(\mathrm{D}+)} \\ \implies \mathrm{P}(\mathrm{D}+ \cap \mathrm{T}+) &= \mathrm{P}(\mathrm{T}+|\mathrm{D}+) \cdot \mathrm{P}(\mathrm{D}+). \end{aligned}$$

Using Bayes' Rule again, we can write $\mathrm{P}(\mathrm{D}+|\mathrm{T}+)$ in terms of $\mathrm{P}(\mathrm{T}+|\mathrm{D}+)$:

$$\begin{aligned} \mathrm{P}(\mathrm{D}+|\mathrm{T}+) &= \frac{\mathrm{P}(\mathrm{D}+ \cap \mathrm{T}+)}{\mathrm{P}(\mathrm{T}+)} \\ &= \frac{\mathrm{P}(\mathrm{T}+|\mathrm{D}+) \cdot \mathrm{P}(\mathrm{D}+)}{\mathrm{P}(\mathrm{T}+)}. \end{aligned}$$

Now, we have all of the quantities we need except for $\mathrm{P}(\mathrm{T}+)$. To find this probability, we can use the Law of Total Probability:

$$\mathrm{P}(\mathrm{T}+) = \mathrm{P}(\mathrm{T}+|\mathrm{D}+) \cdot \mathrm{P}(\mathrm{D}+) + \mathrm{P}(\mathrm{T}+|\mathrm{D}-) \cdot \mathrm{P}(\mathrm{D}-) = \mathrm{P}(\mathrm{T}+|\mathrm{D}+) \cdot \mathrm{P}(\mathrm{D}+) + (1 - \mathrm{P}(\mathrm{T}-|\mathrm{D}-)) \cdot (1 - \mathrm{P}(\mathrm{D}+)).$$

Plugging in all of the quantites we know gives us

$$\begin{aligned} \mathrm{P}(\mathrm{T}+) &= 0.99 \cdot (0.01) + (1 - 0.97) \cdot (1 - 0.01) \\ &= 0.0396. \end{aligned}$$

Finally, we can solve for $\mathrm{P}(\mathrm{D}+|\mathrm{T}+)$:

$$\begin{aligned} \mathrm{P}(\mathrm{D}+|\mathrm{T}+) &= \frac{\mathrm{P}(\mathrm{T}+|\mathrm{D}+) \cdot \mathrm{P}(\mathrm{D}+)}{\mathrm{P}(\mathrm{T}+)} \\ &= \frac{0.99 \cdot (0.01)}{0.0396} \\ &= 0.25. \end{aligned}$$

Thus, the probability that a patient has the disease, given that their test came back positive is 0.25.

Exercise 4: Suppose you flip a fair coin 100 times and record the results. You get 58 heads and 42 tails.

a) Find both the theoretical probability and the empirical probability of getting heads.

Theoretically, the probability of getting heads is 0.5. This is because getting a heads or tails on a fair coin is equally likely. Empirically, we have 58 heads out of 100 tosses. This means our empirical probability of getting a heads is 0.58.

b) Find both the theoretical probability and the empirical probability of getting tails.

As in part (a), the theoretical probability of getting a tails is 0.5. Empirically, we have 42 tails out of 100 tosses, which makes the empirical probability 0.42.

c) If you were to flip the coin 1000 times and record the proportion of times that you get heads, what empirical probability would you expect to observe? Why?

Using the Law of Large Numbers, we know that as the number of flips increases, the empirical probability should converge to the theoretical probability. Thus, we would expect the empirical probability to be very close to 0.5.

d) Give an example of a real-life situation where empirical probabilities would be useful.

Let's say you want to predict whether it will rain tomorrow or not. This is an event that does not have a well known theoretical probability. In order to make your predicition, you can look at the weather from the previous week or two and calculate the empirical probability. If it rained frequently in the past few days, then it is more likely that it will rain tomorrow as well.

Note: Pretty much any answer is fine as long as an appropriate explanation is given.

Exercise 5: Three experiments, each comprising a different number of trials, were conducted. The table below displays the outcomes of rolling a fair six-sided die in each of these

experiments. Answer the questions about empirical probabilities using the table. Compare the empirical probabilities to the theoretical probability, and explain what they show.

Outcome on Die	20 Trials	100 Trials	1000 Trials
1	3	20	169
2	4	20	166
3	4	14	167
4	2	20	166
5	4	13	166
6	3	13	166

a) What is the empirical probability of rolling a 4 for 20 trials?

There are 2 4s out of 20 rolls so the probability is $\frac{2}{20} = 0.1$

b) What is the empirical probability of rolling a 4 for 100 trials?

There are 20 4s out of 100 rolls so the probability is $\frac{20}{100} = 0.2$

c) What is the empirical probability of rolling a 4 for 1000 trials?

There are 166 4s out of 1000 rolls so the probability is $\frac{166}{1000} = 0.166$

d) What is the theoretical probability of rolling a 4 with a fair six-sided die?

On a fair six-sided die, each number appears once and is equally likely to be rolled as any other number. Thus, the theoretical probability of rolling a 4 is $\frac{1}{6} \approx 0.167$

e) Compare the empirical probabilities to the theoretical probability, and explain what they show.

Based on parts (a)-(c), we can see that the empirical probability of rolling a 4 grows closer and closer to the theoretical probability of rolling a 4 as the number of rolls increases. This demonstrates the Law of Large Numbers. By the time we reach 1000 rolls, the empirical probability is only 0.001 away from the theoretical probability (rounded).