Online Assignment

Anahi Rodriguez

due 9/16/2022

There will be two parts to this online assignment. In the first part of this assignment, you will practice applying what we've been learning about to a dataset of your choosing. In the second part of this assignment, you will begin performing simulations. We'll return to the results of the second part of this assignment in a future lecture.

Part 1 [30 points]

1. First, select a dataset on a topic that you are interested in. Be sure that the dataset has at least two quantitative variables.

Here are some sources for a dataset to consider:

- Kaggle
- Tidy Tuesday
- Data is Plural
- ICPSR
- UCI Machine Learning Repository
- FiveThirtyEight
- Google's Dataset Search

If your dataset is located online, provide a link to the dataset. If the dataset is from another source, provide a brief description of the dataset and indicate how you have access to the dataset.

Answer: https://gwis.jrc.ec.europa.eu/apps/country.profile/continent/NA

2. Load this dataset into R. To do so, download the dataset. Then, place the dataset in the same folder as your .Rmd document. If your data is a .csv file, load it in with code like the following: `mydf = read.csv('filename.csv')`. If your data is a .txt file, load it in with code like the following: `mydf = read.delim('filename.csv)`. If your data is in a different format, try searching the internet for sample code, asking in office hours, or posting to campuswire. Once your dataset is in R, print the first few rows of your dataset.

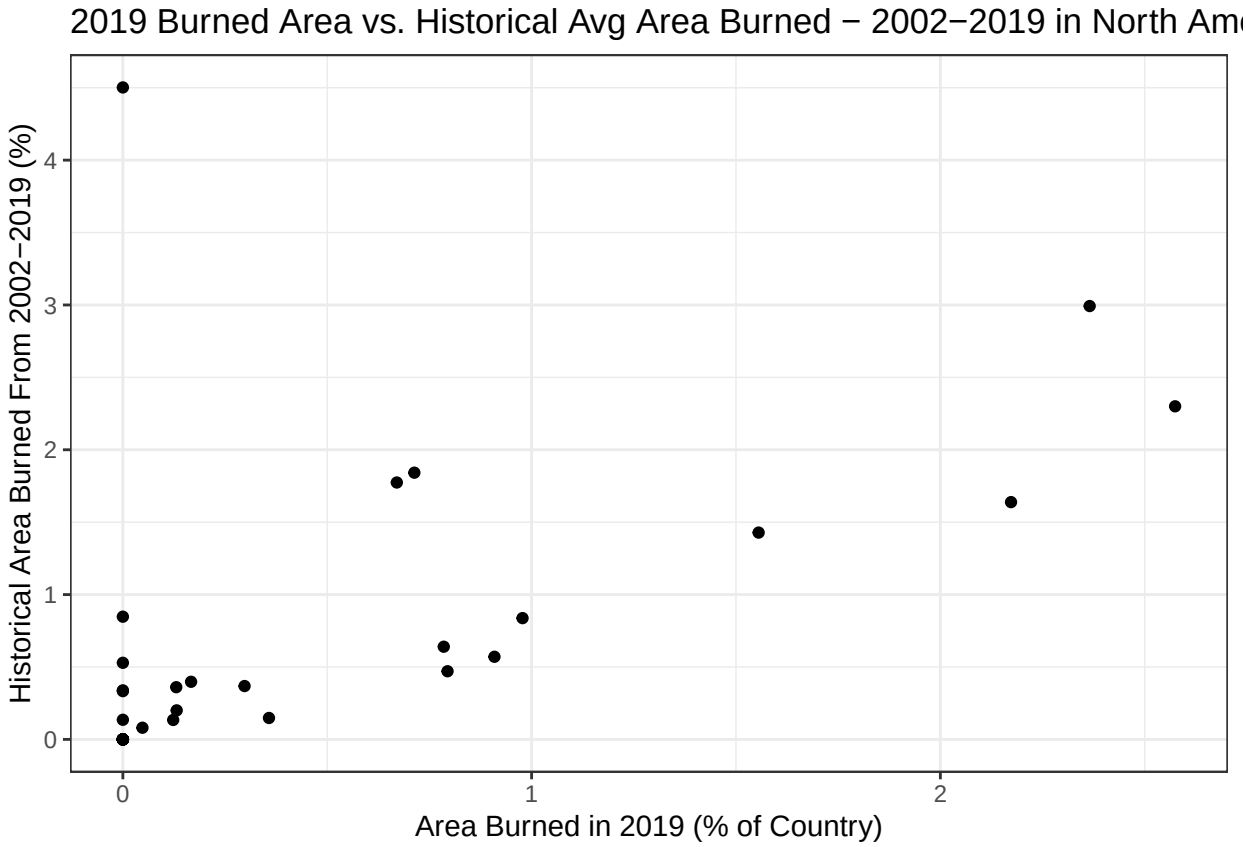
```
# Use this code chunk for your answer.
setwd("~/Desktop/data")
df = read.csv("MCD64.006.estimates-country-ba.2019_2002-2019.USA_.csv")
```

3. Provide one critique or area of concern regarding this dataset. Think carefully about the data, and what variables or observations might be missing about the data. How could someone misuse this data?

Answer: I think that the ratios here are very easy to misinterpret and the name of the variables currently makes this more likely. Additionally, it might have been useful to include the most common causes of the wildfires per country/region.

4. Select two quantitative variables from your dataset to further explore. To start, create a scatterplot of these two variables; make sure that your scatterplot is clear and easy to read. Describe the scatterplot.

```
# Use this code chunk for your answer.
library(ggplot2)
ggplot(data = df, aes(x = ba_ratio_2019.... , y = ba_ratio_2002_2019....)) +
  geom_point() +
  labs(title = '2019 Burned Area vs. Historical Avg Area Burned - 2002-2019 in North America',
       x = 'Area Burned in 2019 (% of Country)',
       y = 'Historical Area Burned From 2002-2019 (%)') +
  theme_bw()
```



Answer: We see a weak but positive linear relatinship between the amount of area burned by wildfires in 2019 and the average amount of area burned between 2002-2019. This tells us that countries that historically have higher (lower) percents of their total land burned have higher (lower) amounts of burned area in 2019. We also see that there exists countries that typically expect none of their land to be burned due to wildfires. Similarly, we can also see that although this is a problem that exists for many North American Countries, the total percent of a country that burns due to wildfires typically stays below 3% with few exceptions both before and during 2019.

5. Generate a linear model to summarize the relationship between your two variables. Write the fitted model and provide an interpretation of your estimated coefficients.

```
# Use this code chunk for your answer.
lm = lm(ba_ratio_2019.... ~ ba_ratio_2002_2019...., data = df)
summary(lm)
```

```
##
## Call:
```

```
## lm(formula = ba_ratio_2019.... ~ ba_ratio_2002_2019...., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03246 -0.13574 -0.13512 -0.01167  1.46980
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.13512    0.10856   1.245    0.221
## ba_ratio_2002_2019....  0.42143    0.09542   4.417 8.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5679 on 36 degrees of freedom
## Multiple R-squared:  0.3514, Adjusted R-squared:  0.3334
## F-statistic: 19.51 on 1 and 36 DF,  p-value: 8.789e-05
```

Answer: Estimated area burned in 2019 = 0.13512 + (0.42143 * average area historically burned between 2002-2019). The intercept coefficient tells us the average percent of area burned in 2019 when the average area burned between 2002-2019 is 0% is on average 0.13512%. The 2002-2019 average variable coefficient tells us that for every additional 1% increase in average area burned between 2002-2019, the amount of area burned within a country in 2019 will increase by 0.42143% land.

Part 2 [30 points]

We'll rely on R to help us create "fake" data and then practice understanding a linear model applied to this data.

We'll provide some initial information to R to set up our data, including our sample size, our x values, and other population characteristics:

```
sample_size = 21
x_vals = seq(from = 0, to = 10, length.out = sample_size)
sigma = 3
```

1. **Replace the following code with your birthdate in mmddyyyy form**. Currently, the birthday of June 13, 1876, which is William Gosset's (pen name Student's) birthday is below.

```
set.seed(02162001)
```

2. Next, we'll set some important characteristics for our data. We start by generating the randomness of our data.

```
epsilon = rnorm(n = sample_size, mean = 0, sd = sigma)
```

Now, generate the values of y based on the following relationship:

$$Y = 7 - 1.4x + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 = 9)$ (independently).

The values of ϵ have been generated in the above code chunk. Save the values of y as `y_vals` in R

```
# Use this code chunk for your answer.
y_vals = 7 - (1.4 * x_vals) + epsilon
```

3. Uncomment the following line of code to create a data frame that contains both x and y.

```
sim_data = data.frame(x_vals, y_vals)
```

Report the dimensions of this data frame. Print the first few rows of this data frame. Calculate the correlation between x & y.

```
# Use this code chunk for your answer.
dim(sim_data)
```

```
## [1] 21  2
```

```
head(sim_data)
```

```
##   x_vals    y_vals
## 1    0.0 10.792018
## 2    0.5  8.284465
## 3    1.0  7.798642
## 4    1.5 11.060227
## 5    2.0  3.846448
## 6    2.5 -2.172807
```

```
cor(sim_data$x_vals, sim_data$y_vals)
```

```
## [1] -0.8234135
```

Answer: This data has 21 observations (rows) and 2 variables (columns). The correlation between the x and y values is -0.8234: a strong negative correlation exists.

4. Create a linear model that predicts y from our x values. Print the summary values for this model. What are the fitted coefficients? Make sure to type the fitted coefficients below. Do they seem reasonable based on our true relationship printed above?

```
# Use this code chunk for your answer.
lm1 = lm(y_vals ~ x_vals)
summary(lm1)
```

```
##
## Call:
## lm(formula = y_vals ~ x_vals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7471 -2.0222  0.1536  1.7195  5.2466
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2378     1.3542   6.083 7.53e-06 ***
## x_vals       -1.4654     0.2317  -6.325 4.53e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.214 on 19 degrees of freedom
## Multiple R-squared:  0.678,  Adjusted R-squared:  0.6611
## F-statistic: 40.01 on 1 and 19 DF,  p-value: 4.533e-06
```

Answer: The fitted coefficients are 8.2378 for the intercept and -1.4654 for our x variable. These coefficients do seem reasonable because as our x values increase, our y values would fall which coincides with a negative correlation.

5. Write out the fitted model. Based on this fitted model, what is $\hat{y}$ when $x = 2$ and when $x = 12$. Which (if either) of these values do you trust more? Why?

```
# Use this code chunk for your answer.
x_2 = 8.2378 + (2 * -1.4654)
x_2
```

```
## [1] 5.307
```

```
x_12 = 8.2378 + (12 * -1.4654)
x_12
```

```
## [1] -9.347
```

```
summary(x_vals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     2.5     5.0     5.0     7.5    10.0
```

Answer: y_hat = 8.2378 + (x_val * -1.4654). I would trust the prediction when x = 2 more since our simulated x values range from 0-10, so the predicted values were calculated using these values of which 2 is included and 12 is not.