14.310x: Data Analysis for Social Scientists
Introduction Unit Homework Assignment

Welcome to your first homework assignment! You will have about one week to work through the assignment. We encourage you to get an early start, particularly if this will be your first time using R. We have provided this PDF copy of the assignment so that you can print and work through the assignment offline. You can also go online directly to complete the assignment. If you choose to work on the assignment using this PDF, please go back to the online platform to submit your answers based on the output produced.

Good luck ☺!

Unit 1- Data is beautiful!

Dell (2010) studies the long-run impacts of the mita, an extensive forced mining labor system in effect in Peru and Bolivia between 1573 and 1812. The mita required over 200 indigenous communities to send one-seventh of their adult male population to work in silver and mercury mines. The mita took place within the boundary shown in Figure 1 (take a close look at the figure and be sure you understand it). It also graphs the altitude of the area with respect to the Earth's sea level (browner areas are at higher levels). Based on this map answer the following questions:

Figure 1

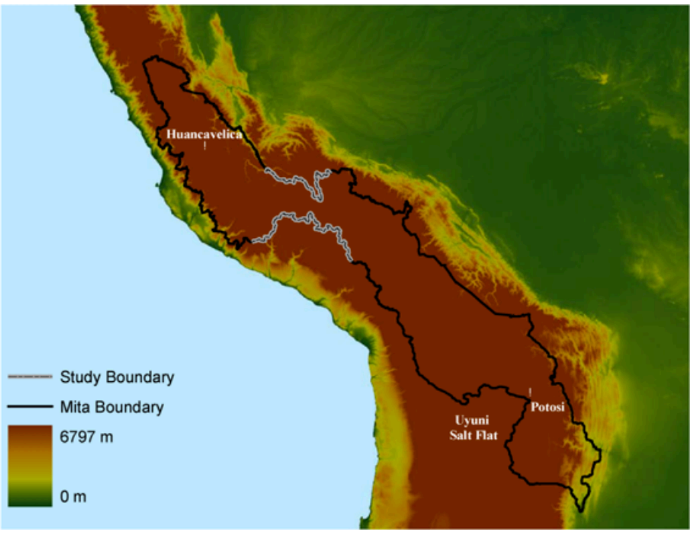


FIGURE 1.—The *mita* boundary is in black and the study boundary in light gray. Districts falling inside the contiguous area formed by the *mita* boundary contributed to the *mita*. Elevation is shown in the background.

1. Which of the following statements are true? (Select all that apply)
   a. The region where the mita took place has lower altitude levels than the region where it did not.
   b. The region outside the boundary has higher altitude levels than the region inside.
   c. The region inside the boundary has higher altitude levels than the region outside.

d. The region where the mita did not take place has lower altitude levels than the region where it did.
e. The region where the mita took place has lower altitude levels than the region where it did not.
f. The mita took place in Argentina and Chile.

2. Looking at the figure, and how the color of the area changes within and outside the boundary, what can you conclude?

   a. Across both, the black and grey boundaries, there is a sharp change in the altitude of the area.
   b. There is a sharp change in the altitude of the area across most of the black boundary, but not across the grey one.
   c. There is a sharp change in the altitude of the area across the grey boundary, but not across most of the black one.
   d. There is no sharp change in the altitude of neither the area across the grey or that of most of the black boundary.

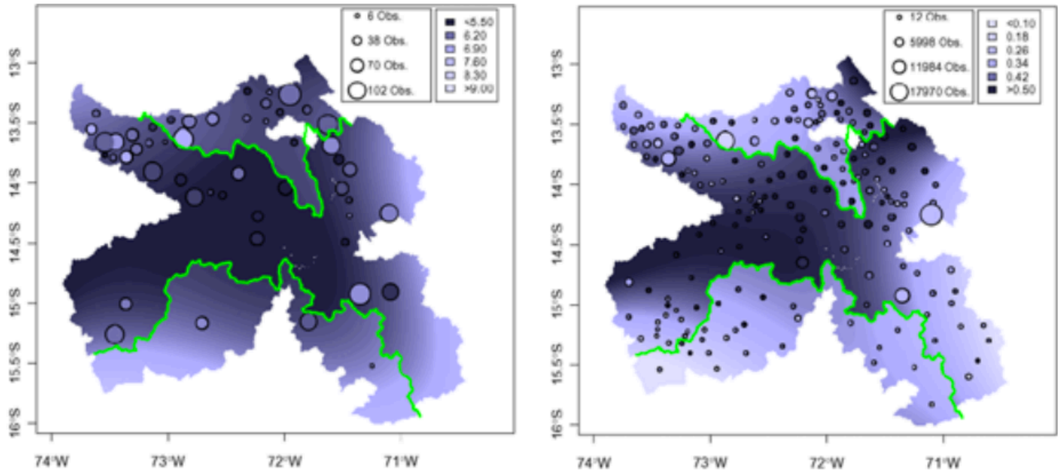
3. In the lecture we discuss the differences between causation and correlation, and the potential risks of confounding the two. If you were interested in studying the **causal** effect of the mita on long-run development, would you prefer to compare regions within and outside the grey or the black boundary?

   a. Grey
   b. Black

Unit 2 – Data is insightful

Continuing with Dell's research, she looks at the way in which more recent welfare variables look like in areas where the mita took place versus areas where it did not. Figure 2 shows a map zooming across the grey boundary: Panel A presents consumption levels in 2001, and Panel B the stunting rate in 2005. Take a look and some time to understand the maps and compare them to the one shown in Figure 1.

Figure 2



(a) Consumption (2001)

(b) Stunting (2005)

The colors on the map correspond to consumption levels and stunting rates, respectively. From the map, you can see that the **darker** areas show *lower* levels of consumption in Panel A, and a *higher* stunting rate in Panel B. Taking this information into account, now answer the following questions:

4. What does the green line in the maps represent?
   a. It corresponds to the black boundary in figure 1.
   b. It shows the grey boundary in figure 1.
   c. It shows the frontier between Peru and Bolivia.
   d. It shows the frontier between the region where Lima is located, and the rest of Peru.

5. What can you conclude from the maps? (Select all that apply)
   a. While the consumption level in 2001 is higher in regions were the mita took place, the stunting rate is actually lower in these places. Thus, it is not possible to conclude whether the mita had a positive or negative effect.
   b. The map shows that both consumption levels in 2001 and the stunting rate in 2005 are higher outside the boundary, showing a negative causal effect of the mita.
   c. Inside the boundary, the consumption level in 2001 is lower and the stunting rate in 2005 is higher, implying a negative effect of the mita in the long run.
   d. From the maps, it is not possible to conclude whether the mita had a positive, negative, or ambiguous impact. It is necessary to collect more data.

In the lecture, Professor Duflo presented Michael Greenstone and coauthors' research, where the relationship between pollution and the distance to the Huai river had two different visualizations: (1) a map similar to the ones in Figure 2, (2) a two-dimensional plane of the data. The latter showed the degree to the north in the x-axis and the level of pollution in the y-axis. Suppose that we were trying to do a similar visualization here. To simplify the plot, we only take the boundary in the south. Assume that the x-axis corresponds to the degree in the north, and that we normalize the boundary to zero. It might be helpful to make some drawings for a better visualization of the plot.

6. From this visual representation, are the regions that had mita presence in the negative or positive side of the x-axis?
   a. Negative
   b. Positive

7. Now think that we plot the consumption level in 2001 in the y-axis. Fill in the blanks for the following statement:

   The plot will show: (i) In the negative side of the x-axis, a _______ relation between consumption levels and the degree of the north, (ii) a _______ jump in zero, (iii) and a ________ slope between consumption and the degree of the north in the positive side of the x-axis.

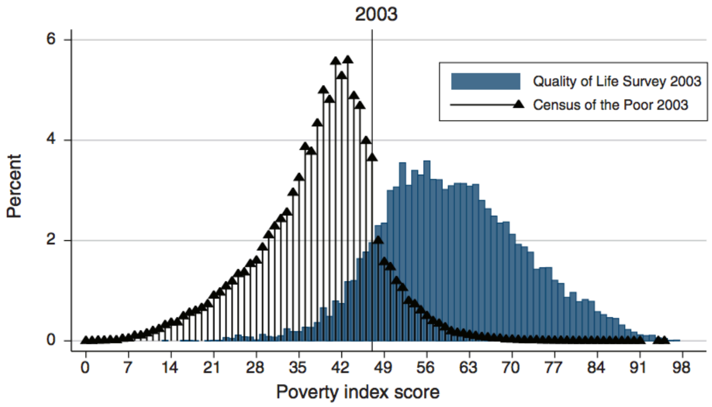
   a. negative; negative; flat
   b. positive; negative; positive
   c. null; positive; flat
   d. negative; null; negative
   e. null; negative; negative

8. Imagine a similar plot for the stunting rate in 2005 in the y-axis. Would you expect to find a jump in the zero of the x-axis?
   a. Yes, a negative jump.
   b. Yes, a positive jump.
   c. No, there would be no jump.
   d. We can't tell with the information provided.

Unit 3 – Data is Powerful

9. Camacho & Conover (2011) document manipulation of a targeting system for social welfare programs in Colombia. Take a look at the following figure, which shows two histograms: the black arrows present the histogram for a poverty score (lower numbers mean being poorer) that was calculated using the same data the Government collected to target social welfare programs – where only individuals with a poverty score below 48 were eligible to receive most of these programs. The blue bars correspond to the histogram reconstructing this poverty score using other data sources that were not used by the Government for this purpose.

Figure 3

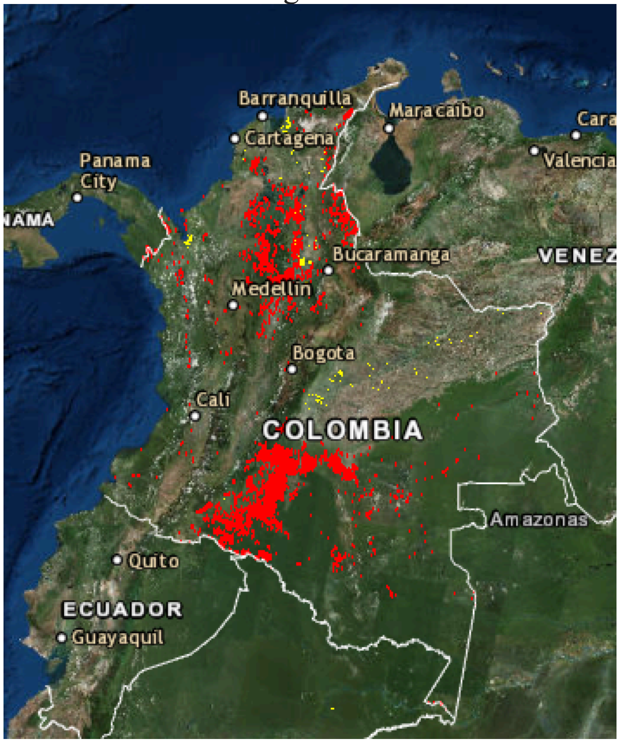


What can you conclude from the graph? (Select all that apply)

a. People were targeted as poorer in the data set used by the Government to determine social welfare eligibility than in the alternative data sources.
b. Due to differences in the way the data is collected it would be expected to see these differences between the histogram represented by the black arrows, and the one shown by the blue bars.
c. Since a lot of people apply and look for the benefits of social programs, it is not surprising to see the bunching in the poverty score at 48. It is likely that rational individuals answer surveys in order to have poverty scores that make them barely eligible for social welfare.
d. Since a histogram shows how does the poverty score vary across different values, it is indeed surprising the bunching at 48. It should be a continuous variable in both, the black arrows and the blue bars.
e. The black arrows show a discontinuity in the mass of the population exactly in a poverty score of 48. Since this is not shown with the blue bars, this suggest some sort of manipulation of social welfare targeting.

10. Continuing with Colombia, *www.laramaciudadana.com* is a blog that publishes quantitative information about different topics of national interest. Their objective is to inform public policy debate by collecting data on these controversial topics and displaying it to a general audience. Their most recent project uses satellite photos to map deforestation and evaluate industrial reforestation efforts in the country. The map is presented in Figure 4: the red dots show the locations where satellites detected deforestation activities, and the yellow dots give an overview of the industrial reforestation efforts made by the Government in recent years. Take a close look at the map.

Figure 4



Based on this visualization of the data, would you conclude that the efforts made by the Government are located in the areas where deforestation has taken place?

a. Yes
b. No
c. From the map this is impossible to conclude

Unit 4 – Data can be deceitful and correlation versus causality

11. During the introductory lecture, Professor Duflo discussed that human capital externalities are one potential explanation for the fact that the relationship between schooling and output at the country level is larger than the relationship between an additional year of schooling and income at the individual level. She also argued that some of these externalities could come from teaching or exchanging ideas within a city. A researcher decides to test this idea formally and she correlates the average schooling level in the city with the individual wage of a sample of individuals. She finds a strong positive correlation! From this statistical evidence could she conclude that there are human capital externalities?

a. Yes
b. No

Unit 5 – Introduction to R

12. Suppose that you want R to print the following statement “Hello world!”, write your code in the following box:

13. If you run the following code in R, what does the object my_sqrt contain?

```
z <- c(pi, 205, 149, -2)
y <- c(z, 555, z)
y <- 2 * y + 760
my_sqrt <- sqrt(y - 1)
```

a. A single number (i.e a vector of length 1).
b. A vector of length 0 (i.e. an empty vector).
c. A vector of length 1.
d. A vector of length 3.
e. A vector of length 9.

14. Are the two following codes in R equivalent?

Code 1

```
z <- c(1:4)
z * 200 + 3
```

Code 2

```
z <- c(1:4)
z * c(200, 200, 200, 200) +
c(3, 3, 3, 3)
```

a. Yes
b. No

15. What kind of matrix would *my_matrix* be if the following is run in R?

```
my_matrix <- matrix(1:6, nrow = 3, ncol = 2, byrow = TRUE)
```

a.

```
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

b.

```
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

c.

```
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

d.
```
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
```

.

16. Now assume that each row in the matrix represents a different person. We had created the following vector:

    names_of_students <- c("lisa", "juan", "diana")

    Write down the code that allows you to assign these names to the rows of the matrix *my_matrix*.

    ______________ <- names_of_students

17. Assume that you tell R to divide zero by zero, what would you get?
    a. NA which corresponds to not being a number.
    b. NaN which corresponds to a missing value.
    c. NA which corresponds to a missing values.
    d. NaN which corresponds to not being a number.
    e. Both, NA and NaN since for R they are the same object.

18. If you have a missing value and you try to add it a number, what result would you get?
    a. NA
    b. The number you are trying to add
    c. An error, since R is not able to perform operations with missing values

19. We have asked the age of a group of 12 students. While 10 of them provided us with this information, 2 of them did not. We have constructed the vector *age* that captures this information.

    ```
    age <- c(12, 28, 35, 27, NA, 25, 32, 45, 31, 23, NA, 34)
    ```

    If we were interested in getting the vector without the missing values, which of the following lines of code would be useful to achieve this purpose? (Select all that apply)

    a. age[c(5, 11)]
    b. age[-c(5, 11)]
    c. age[c(-5, -11)]
    d. age[1:10]
    e. age[c(1, 2, 3, 4, 6, 7, 8, 9, 10, 12)]
    f. age[is.na(age)]

g. age[!is.na(age)]

20. Which of the following statements are true? (Select all that apply)

a. Matrices in R only allow objects are numeric, if you try to create matrices with strings you will find an error.
b. If a vector contains numeric and string characters, and R transforms it into a matrix then the numeric objects would be treated as strings.
c. Matrices in R allow for collections of objects of different types, while lists do not.
d. Matrices in R allow for different types of objects, but all of them have to be of the same kind.
e. Contrary to matrices, lists allow for a collection of objects of different types.
f. If a vector contains numeric and string characters, and R transforms it into a matrix then it will keep the original type of the objects.
g. Matrix and lists are the same in R.