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 this 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 that was 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 the figure below (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).

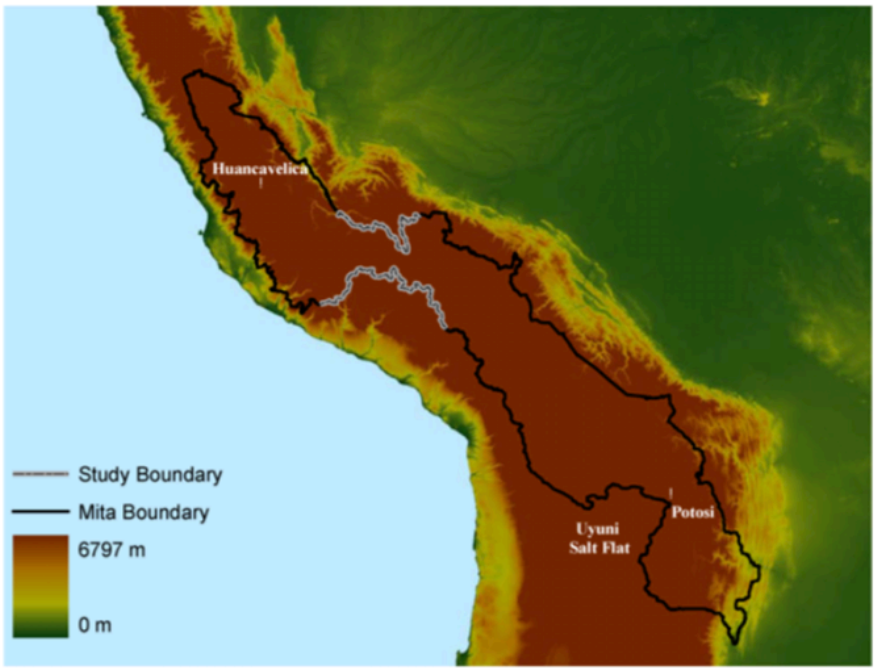

Figure 1 The mita boundary is in black and the study boundary in light gray. Districts falling inside the contiguous area formed by the mita boundary contributed to the mita. Elevation is shown in the background.

Based on this map answer the following questions:

Question 1
Which of the following statements are true? (Select all that apply)

- ☐ The region where the mita took place is in an exclusively low altitude area.
- ☐ The region outside the grey and black boundaries is exclusively high altitude.
- ☐ The region inside the grey and black boundaries has mostly high altitude levels.
- ☐ The region where the mita did not take place is exclusively low altitude.
- ☐ The mita took place in Argentina and Chile.

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

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

Question 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 it be better 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.

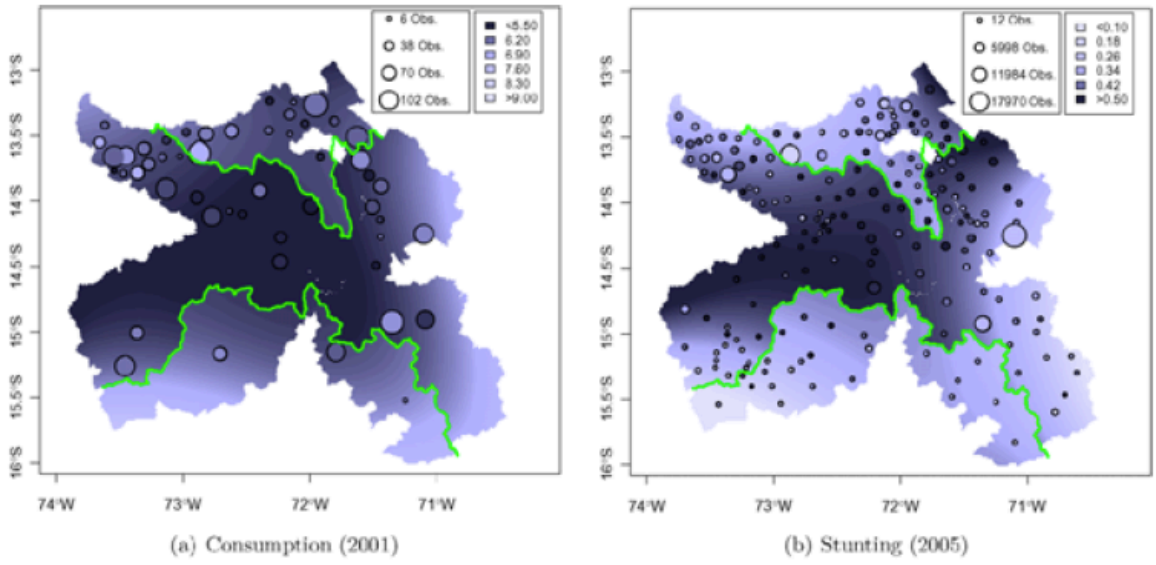

Figure 2

The colors on the map correspond to consumption levels and stunting rates, respectively. From 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:

Question 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.

Question 5
What can you conclude from the maps? (Select all that apply)

- a. While the consumption level in 2001 is higher in regions where the mita took place, the stunting rate is actually lower in these places. Thus, it is not possible to conclude whether the mita has 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.
- 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.

Question 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

Question 7
Now consider if we plot the consumption level (Panel A) in 2001 in the y-axis. Fill in the blanks for the following statements:

The negative side of the x-axis will show a _____ relation between consumption levels and its position (degree to the north).

- Flat
- Positive
- Negative

The plot will show _____ at x=0.

- No change

- A positive jump
- A negative jump

The plot will show a _____ between consumption and its position (degree to the north) on the positive side of the x-axis.

- Flat relation
- Positive slope
- Negative slope

Question 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

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.

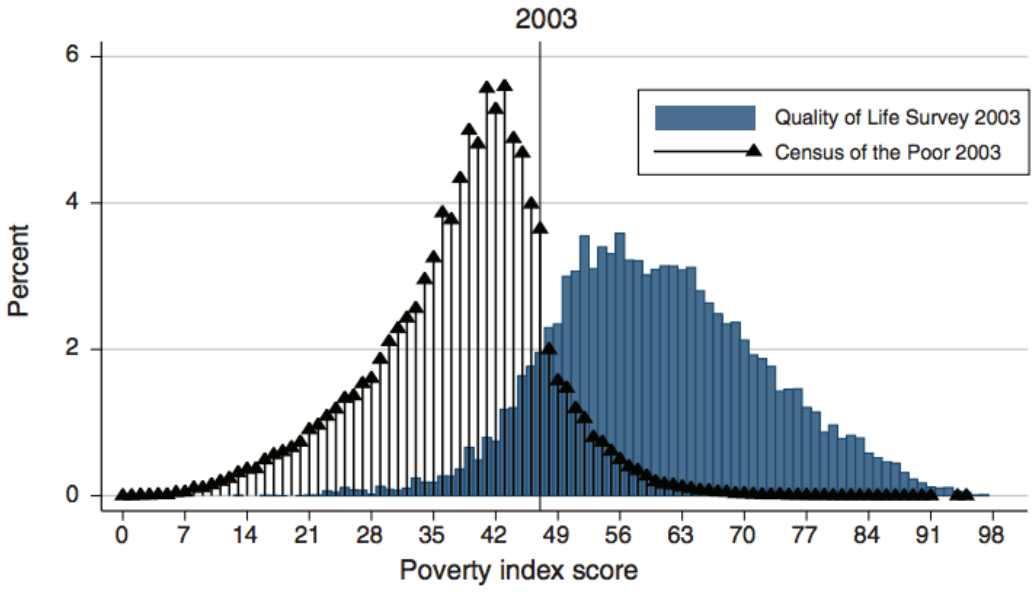

Figure 3

Question 9
What can you conclude from the graph? (Select all that apply)

- The two datasets (the one used by the government to target social welfare programs and the alternative data sources) suggest two different levels of poverty for the same population considered.
- Due to the source of the data, it would be expected to see these differences between the histogram represented by the black arrows and the one shown by the blue bars.
- The two data sources are measuring different outcomes so it's not surprising that the histograms show different patterns.
- The data from the Quality of Life Survey corroborates the census data collected by the government, as evidenced by the blue and black histograms.
- The black arrows show a discontinuity in the mass of the population exactly at a poverty score of 48 (the eligibility score used by the government). Since this is not shown with the blue bars, this suggest some sort of manipulation of social welfare targeting.

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

Question 10
Based on this visualization of the data (see Figure 4, above), 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

Question 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 school and income at the individual level. She also argued that some of these externalities could stem 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! Fromt his statistical evidence, could she conclude that there are human capital externalities?

- a. Yes
- b. No

Unit 5 – Introduction to R

Question 12
Suppose that you want R to display "Hello world!" Drag and drop to create the correct R input. Note that there may be multiple ways to write this in R but we want you to use this command specifically.

(	[	"Hello world!"	display	print	)

Question 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.

Question 14
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.

Question 15
If you have a missing value and you try to add it to 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

Question 16
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)]`

Question 17
Download the data "CitesforSara.csv" into RStudio. This dataset includes paper-level citations from 1969 to 1998. First, read the CSV file into R using these commands:

```
library(tidyverse)

papers <- as_tibble(read_csv("[YOURFILEPATH]/CitesforSara.csv"))
```

Great! Let's create a simplified dataset which only keeps the following variables contained in the papers dataset in this order: journal, year, cites, title, and au1. Use the method `select()` to accomplish this. Set this output to the variable `papers_select`. Drag and drop to create the code. Note that there may be more than one way to do this but there is only one correct answer from the following drag-and-drop options.

`[`	`select`	`journal,`	`(`	`papers_select`
`is.na`	`<-`	`papers,`	`year,`	`)`
`"`	`au1`	`"`	`cites,`	`title,`

Question 19
Let's take a look at some of the most popular papers. Using the `filter()` method, how many records exist when there are greater than or equal to 100 citations?

- a. 22
- b. 100
- c. 205
- d. 2251

Question 20
Use the `group_by()` function to group papers by journal. How many total citations exist for the journal "Econometrica"?

- a. 2251
- b. 75789
- c. 4182
- d. 3738

Question 21
How many distinct primary authors (au1) exist in this dataset?

- a. 2332
- b. 4132
- c. 205
- d. 1242

Question 22
Use the *dplr* `contains()` method to create a new dataset *papers_female* which contains only the columns from papers containing the string "female". Drag and drop to create the code. Note that there may be more than one way to do this but there is only one correct answer from the following drag-and-drop options.

"female"	(	<-	)	all_papers	)	contains
is.na	papers_female	select	papers,	(	print	