17.835: Problem Set 1

Professor: In Song Kim
TA: Jesse Clark, Sean Shiyao Liu, and Nicole Wilson

Due: September 16 before class
(Late submissions will not be accepted)

This problem set will help you get acquainted with working with data in `R`, the statistical software we will be using for this course. Please post all questions to the Piazza discussion forum, which you can access through the link on the class website or at:

https://piazza.com/mit/fall2020/17835

For this and future problem sets, you will need to submit two files onto the Canvas (under `Assignment`): 1) your write-up with your answers to all questions (including computational questions and any related graphics) as a PDF file, and 2) your code as an `R` file (so that we can verify it runs without errors). Both files should be identified with your last name (e.g., `wilson.R` and `wilson.pdf`). Please ensure that all of these are completed *before* class on the day of the due date – late submissions will be automatically flagged to us by Canvas. For problems that require calculations, please show all the steps of your work. For problems involving `R`, please ensure that your code is thoroughly commented.

Problem 1: Predicting COVID-19 Transmission by Population Flows

An accurate prediction of where an epidemic would likely occur could help save lives, because health authorities will be able to get prepared and take precautionary actions to avoid a "bank run" on medical resources. Some social scientists from both China and United States have recently published on *Nature*. They find population movement data can predict COVID 19 cases 14 days in advance. In this problem set you will be asked to reproduce some of their results, through which you will practice data manipulation and plot generation in R.

1. First, load the `distance_to_wuhan.csv` and `cities_info.csv` datasets found on the class website. To do this, set the working directory to whichever folder the file is in. Then load the file into R using the `read.csv` command. If you type `?read.csv` in your `R`, you can see more on how you can use the command, e.g. on how to set the header or separator.

   `distance_to_wuhan.csv` contains the geographical information of about 300 Chinese cities. The dataset contains the following variables:

   - **city_name**: city name
   - **latitude**: Latitude of the city
   - **longitude**: Longitude of the city
   - **distance_to_wuhan(km)**: distance of the city to Wuhan in unit of kilometer
   - **wuhan_outflow(Jan1 to Jan24)**: The total number of cell phone users who have travelled out of Wuhan, the epicenter of COVID 19 in China, between Jan 1 and Jan 24. Jan 24 is the date when Wuhan was put under city-wide quarantine.

cities_info.csv contains city-level information about roughly 300 Chinese cities. The dataset contains the following variables:

- **cumulative_confirmed_cases(Feb19)**: Number of cumulative COVID-19 confirmed cases in the city as of Feb 19.
- **population(10 thousand)**: Number of residents in the city in the format of 10 thousand
- **GDP**: Annual GDP of the city in the format of 10 thousand Chinese Yuan

2. Combining two datasets will make following questions much easier. Using `R`, merge the geographical info and socio-economic info of the cities into one dataframe by `city_name`. The new dataframe should contain all variables from previous datasets. Display the first several rows of the merged dataset to show your that you have completed your work.

3. As the distribution of our data is highly skewed – some of the cities have large numbers of confirmed cases, and other cities have large numbers of outflow from Wuhan, for a better interpretation of the data, we want to take logrithm for both the cumulative number of cases and outflow from Wuhan. Name these two variables `log.wuhan.outflow` and `log.cum.cases` respectively.

4. Next, we want to explore the relationship between the cumulative number of cases and size of outflow poulation from Wuhan. We're going to make a scatter plot, which are a type of plot that shows the relationship between two variables. Using the `plot` function, plot the logged population outflow from Wuhan on the X-axis, and the logged cumulative confirmed number of cases on the Y-axis. Use `pch` argument to set the shape of the points to 16, and the color of the points to red, and use `par` to set the font size to help with your readers. Make sure you have labelled each axis so that your readers understand your plot.

5. An alternative way to visualize the relationship is to plot the texts instead of points. Create a similar plot as above. But this time, create a blank plot first. Then, use `text` argument to add the city names and set `cex`, namely the size of label, to 0.45. You may use `jitter` argument to avoid text overlapping.

   Further, for better illustration, we want to add a line that describes the bi-variate relationship. In this problem, we give you the intercept = -1.4193 and the slope = 0.5662 for the regression line. Please use the `abline(a=intercept, b=slope)` to draw the line.

6. Now suppose that you were the Director of China CDC and had this dataset before the spread of the virus. You would like to evaluate *ex ante* the number of the cities that have the potential to be affected seriously by the disease if actions were not taken pre-emptively. To do this, you want to know the number of the cities that have a population over 10 million[1] and an outflow from Wuhan larger than 10 thousand.

   There are more efficient ways of doing this, but here we want you to practice for-loop and if-else in `R` for our instructional purpose.

   You perform your analysis with the following steps:

   (a) Create an object named `cities.with.risk` and set it as a vector of character with length zero

[1] Note in the dataset, population are recoreded in the unit of 10 thousand.

(b) Write a for-loop that goes through each row. For each iteration, if both the city's population is larger than 10 million and the outflow from Wuhan to this city is larger than 10 thousand, add the city's name to the vector `cities.with.risk`

(c) Print the vector to see the list of the city that have the potential to be seriously affected if actions were not taken pre-emptively.

Hint: `city.names` may be a vector of factors in your data.frame. You may need to convert it into a vector of characters with `levels` or `labels`, as you may decide, before you your loop.

Problem 2: "Shy" Conservatives and Post-Material Politics

Surveys are frequently used to measure political behavior, for example to learn about the demographic characteristics of voters. They are also used by analysts, the media and the public to try to predict election outcomes. But researchers are generally concerned that people might not be honest when they answer surveys. In fact, one notable phenomena that researchers have identified is the "Shy-Tory factor,"[2] by which conservative voters are more likely to conceal their vote. The failure of pollsters to predict Trump's election in the US or the *Brexit* referendum result are associated with this phenomena.

A related problem is the fact that progressive parties tend to be favoured among young people, who expresses support for them in surveys but then does not attend to the polls. In surveys, young people tends to say that they voted or will vote for a progressive party, but many of them "forget" to go to the polls on the election day. This is a typical political science problem because in surveys it is difficult to distinguish "attitudes" from actual behavior.[3]

Across the world, countries have different electoral systems to select their representatives. The two most common systems are single-member plurality and proportional representation (PR) systems. In single-member constituencies, each electoral district in the country - generally a small geographical area - is assigned one seat in the lower chamber, all the candidates in that district compete for that single seat, and whoever receives more votes gets the seat. In PR systems, districts are larger and there are many seats to be assigned. The competetion is among parties instead of individual candidates and the parties are assigned a proportion of seats that corresponds to their proportion of votes.

Germany is a special case because it has a hybrid electoral system that combines single-member constituencies (like the US) with proportional representation (PR) to elect the Bundestag (lower house). This means that each voter casts two votes at the same time: one for a candidate in their constituency and one for a party list.

To examine whether the "Shy-Tory factor" and the "lazy" progressives also exist in Germany, we are going to use data from Germany collected by the European Social Survey (ESS). The ESS is a survey that runs every two years and covers social and political issues in most European countries. Below, we display the names and descriptions of variables in the `ESS_Germany_2018.csv` data (available on Canvas).

- **vote**: Whether the respondent voted ("Yes","No") or is not elegible to vote ("Not elegible to vote")
- **prtvede1**: Party voted in the First vote (constituency) in the 2017 election for the Bundestag (lower house).
- **prtvede2**: Party voted in the Second vote (list) in the 2017 election for the Bundestag (lower house).

[2] The name comes from the fact that it was first identified in the 1992 UK general election. To know more about this factor you can read this paper written in the aftermath of the 1992 election by Jowell et. al (1993

[3] For a discussion of the young voters issue see Holbein and Hillygus (2015).

- **gndr**: Gender
- **agea**: Age

Since we want to compare the data from the survey with the actual results we have also included the national vote totals and percentages by party for each type of vote in the `results_germany_2017.csv` file.

1. Alternative for Germany (AfD) is a far-right populist party associated with Neo-Nazi positions. The Greens is a progressive environmentalist party. Both of these parties have experienced a relevant increase in their vote shares in the last elections.

   Load the `results_germany_2017.csv` dataset which has the results (the **"truth"**) for the 2017 German election. What was the percentage of votes of the AfD and The Greens in the First and Second votes? (*Hint:* To correctly display this dataset you will need to set the encoding to `UTF-8` and the separator to `";"`).

2. Since the ESS is a large representative sample, the vote percentages by party obtained by asking respondents in the survey who they voted for should be pretty close to the actual results from the national election.

   Calculate the vote percentages of each party in the First and Second votes using the ESS data (*Hint:* `prop.table` may be useful here. Also note that your denominator will be those who reported that they voted *and* said which party they voted for. You will notice that there are some respondents who reported voting but did not disclose a particular party, but we will not worry about these respondents for now). Then, create a table adding a column with the estimated (ESS data) and the actual data, as well as a column with the difference between the two. Your table should have seven colunms: party name, estimated First vote, actual First vote, difference between estimated and actual First vote, estimated Second vote, actual Second vote, difference between estimated and actual Second vote

   Make sure that the table you create have clear descriptive names in its columns. Order your table by First true vote percentages, in descending order.

   Using this table, create a barplot showing the survey percentage and the actual percentage for the First vote. Use only the top five parties by First vote in your plot. The x-axis should have party names and the y-axis the percentage. The bars for each party should be next to each other. Use a different color to distinguish the survey percentage from the actual percentage, and add a legend. Your plot should look similar to the final bar plot covered in our recitations. (*Hint*: You might want to check the `barplot` function and its documentation, paying attention to what types of objects you feed into it).

   What do you conclude? Which party benefits the most from people "lying"? Which one do voters seem to feel more "ashamed" of saying that they voted for? Are German progressive voters "lazy" and conservative voters "shy"? By "benefit", we mean a voteshare lower in the actual result but higher in the estimated survey result. It is a benefit because it makes the party look popular in surveys, but we have a quotation mark to indicate that this might mislead the party's campaign and thus is not necessarily a benefit.

3. **Extra credit:** Until a couple of decades ago, most political scientists thought that the main determinant of partisan identity was the social class. That is, workers and low-income individuals

identify with left-wing parties, while capitalists and upper-middle classes identify with right-wing conservative parties.

In the mid-90's, some authors started arguing that there was an undergoing change in political preferences. They observed that among new generations issues like the environment, gender equality, diversity and tolerance were more important than classical left-right issues like taxes or welfare. Ronald Inglehart and Pippa Norris called these new preferences "post-materialist values" (as opposed to the "materialist" preferences of the past) and they argued that age and gender (instead of class) were the new key determinants of partisan identities (with women and young people identifying with parties promoting progressive values and old people and males identifying with culturally conservative parties).[4]

Let's then explore the relationship between some demographic characteristics and voting behavior to see if there are differences in age and gender between The Greens and AfD supporters. First, create a table showing the percentages of AfD and The Greens supporters (i.e., claim to have voted for that party) who are male and female. (*Hint:* Note that you will probably need to subset the data to these two parties). Second, create a density plot with the age of the voters of the AfD and The Greens. For simplicity, in this question use the First vote.

Both densities should be on the same plot. Use red for the AfD and green for The Greens. The main title of the plot should be "Distribution of Voters' Age". There should also be a legend referencing the colors.

What do you notice in the table and in the plot? Is there evidence of a "post-materialist" partisan divide? What could be a problem with this interpretation?

Problem 3: Income inequality and redistribution

Income inequality and government redistribution (i.e., social policies that reduce income inequality) often play a major role in political debates. For example, the idea of establishing a wealth tax is frequently debated in the current Democratic primary as a potential policy solution to the USA's high levels of inequality. In this problem, we will use time-series cross-sectional data (that is, data that covers multiple countries over multiple years) to explore the relationship between inequality and redistribution among high-income democracies.[5]

There are two main insights to this problem. First, inequality can be measured in many different ways, and the way we measure makes a difference for how unequal a particular society appears, and for the kinds of policy solutions one would propose. Second, political scientists have long argued that there is a relationship between inequality and redistribution, but the theoretical connection between the two (i.e., our model about how they are related) and the empirical evidence to support that connection are still being debated. These academic debates feed directly into intense political and policy debates about inequality and redistribution.

To address these questions, we will work with the dataset of an academic paper: Noam Lupu and Jonas Pontusson. "The structure of inequality and the politics of redistribution." *American Political*

[4]The debate between authors who believe in the "materialist" argument to explain partisanship versus those who believe in the "cultural" explanation remains unsolved. Though Inglehart and Norris have written extensively on this issue, this paper is a good starting point for the discussion link (it also uses the Chapel Hill Expert Survey and the European Social Survey, two good datasets that you could use if you are interested in the topic.

[5]To get a sense of how your household income or the income you think you will make after you graduate compares to the income distribution in the US, and other cool facts about inequality, check out the OECD's page https://www.compareyourincome.org/.

Science Review 105(2): 316-336.[6] Here are the names and descriptions of variables in the dataset that we will use:

- **country**: country name
- **year**: year
- **redist**: percentage change in Gini coefficient (a measure of income inequality) when moving from gross market income (i.e., household income *before* taxes and transfers) to disposable income (i.e., income *after* taxes and transfers)
- **ratio9050**: earnings of a worker in the 90th percentile of the earnings distribution as a share of the earnings of the worker with a median income
- **ratio5010**: earnings of the worker with a median income as a share of the earnings of a worker in the 10th percentile of the earnings distribution

1. First, download the authors' dataset from their website (http://www.noamlupu.com/LupPon_APSR.dta) and put the file in the same directory as your R file. The dataset is in Stata format (`.dta`), so to read it into R you will need to install the package `foreign`, load the package, and then load the file with `read.dta`. How many countries are included in the dataset? How many years (and which years) do we have data for? For those, how many country-year observations have data on the level of redistribution, the 90-50 ratio, and the 50-10 ratio? For how many country-year observations do we have data on all three variables?

2. Next, let's look at which country-year observations have the 5 lowest and the 5 highest levels of inequality. There are many different ways of measuring inequality, but a common one is the 90-50 ratio, i.e. the earnings of a worker in the 90th percentile of the income distribution as a share of the earnings of a worker in the 50th distribution. This gives us a sense of how wealthy the wealthy are relative to the middle class. To see which country-years have highest and lowest levels of inequality in this metric, `order` the dataset by the variable `ratio9050`, and then look at the top 5 rows and the bottom 5 rows, remembering that we have missing data for many of the observations in the dataset. By doing so, we are indeed considering country-year as the unit of comparision. That is, we are checking which country and in which period has the largest/smallest level of inequality over the years and countries we have data about.

3. Does your answer change when you consider the 50-10 ratio, as an alternative measure of inequality? What does this measure of inequality focus on?

4. Now let's create a third measure of inequality focused on how large the distance between the rich and the middle is, relative to the distance between the middle and the poor. To do so, divide the ratio between the 90th and the 50th percentile by the ratio between the 50th and the 10th percentile. Let's call this variable `skew`, like the authors do.

   To visualize the distribution of the `skew` of the income distribution across countries, create a plot with 19 boxplots side by side (one per country), and overlay a horizontal line with the overall median skew in the dataset. To make sure that you have enough space for country names in the

[6] Available on http://www.noamlupu.com/structure_inequality.pdf. Note that most academic papers recently published in top journals make their datasets freely available online. If you see a paper using a dataset you would like to use for your projects, it is likely you can download the data from the publisher's or the authors' website.

plot, make two adjustments: (i) use `par(las = 2)` before you plot, so that axis labels are plotted horizontally, and use `par(mar = c(X, X, X, X))` to adjust the margins of the plot (you'll have to experiment with different values for X, which correspond to the margins on the 4 sides of the plot). The reason we are using a boxplot here is that we want to make the unit of comparison countries rather than country-year here. We would like to know in general how the levels of `skew` are different across different countries.

Looking at the plot, which countries have the smallest and the largest skew?

What would a skew below 1 mean in substantive terms? How many country-year observations are there with skew below 1? And above 1?