HAD5772: Final Project Guidelines

Overview

This project gives you hands-on experience summarizing and analyzing data of your own interest using R. You should use publicly available healthcare data to address a research question of interest, as discussed in class. Feel free to consult with me or with other faculty if you need help finding a specific type of data.

This project prepares you to answer descriptive questions such as:

1. Measurement: On average, how big is variable X?
2. Heterogeneity: How widely does X vary?
3. Correlation (causality?): Are variables X and Y positively or negatively correlated? How strong is this relationship?
4. Regression: How can I use variable X to predict variable Y? How can I start to measure causal effects of X on Y?

Your project should be submitted as a writeup that presents and discusses the following information in a consistent and coherent way. Formatting of tables and figures is important, as discussed in class – make sure everything is readable and succinctly conveys the story you want to tell.

Part 1 – Research Question & Motivation

1. *Research question* – Identify a research question of interest and clearly state it. Note that this can (and likely should) be a descriptive question relating two or more variables, given the scope of this course.
2. *Motivation* – in your introduction, defend the reasoning behind selecting your question, what makes it relevant for current research, and what answering this question would mean for policy.
3. *Literature Review* – briefly discuss past literature relevant to your question. What is known already? What previous attempts have been made at answering your question? Can you isolate the contribution your analysis would make in this space? (Note that this project is not meant to be a groundbreaking project; your contribution can be replicating previous results in a new setting, for example).
4. *Identify your audience*

   Identify some audience that might find these data interesting: a policy maker, a business or industry leader, a consumer, etc. In Part B, you will report your findings to this audience. List any questions (at least two) that this audience might have, that you believe your data can shed (at least partial) light on.

Part 2 – Data Collection & Summary

1. *Basic data description*
    a. Collect your data of interest and report where you got it. You do *not* need to submit your data files.
    b. How were these data originally collected, and by whom?
    c. What is your unit of observation? How many observations do you have?
    d. List at least one variable that is *quantitative* (e.g. price, number of sales, age, GDP) and another that is *binary* (e.g. gender, race, industry, political party, sport position). If any of these variables are not obvious already, explain how they are determined or measured (e.g. what units) [1] or constructed.

Note that, while not required, it is often interesting to pull data from multiple sources, or to construct new variables from existing data. In the spreadsheet below, for example, government finance variables come from one source and a binary political variable comes from another. Per capita variables are then computed simply as ratios; growth variables are computed simply as differences (as a ratio of the original level); and additional binary variables are constructed either by reducing a quantitative variable into "high" and "low" categories (e.g. GDP growth above or below 1.5%) or by comparing two existing variables (e.g. Gov. growth > GDP growth?).

Unit	Original Variables				Constructed Variables					
	GDP	Population	Gov. Spending	Republican House?	Per capita GDP	Per capita GDP growth	GDP Growth > 1.5%?	Per capita Gov. spending	Per capita Gov. growth	Gov. growth > GDP growth?
	($ bil.)	(mil.)	($ bil.)		($ thous.)	(%)		($ thous.)	(%)	
Year										
2008	14,834	304	4,665	0	48.8	-	-	15.3	-	-
2009	14,418	307	5,179	0	47.0	-3.7%	0	16.9	10.1%	1
2010	14,779	309	5,057	0	47.8	1.7%	1	16.3	-3.2%	0
2011	15,052	312	5,116	1	48.3	1.1%	0	16.4	0.4%	0

[1] For example, a humanitarian agency might rate sovereign governments as "corrupt" or not, and designate individuals as "in poverty" or not, but how are these categories assigned? What exactly do they mean?

2012	15,471	314	5,042	1	49.3	2.0%	1	16.1	-2.2%	0
2013	15,759	316	4,955	1	49.8	1.1%	0	15.7	-2.5%	0
2014	16,077	319	4,957	1	50.4	1.3%	0	15.5	-0.7%	0

2. *Present summary statistics* -- In general, summary statistics should include the information relevant to your context. These could include: the minimum and maximum, mean, median, and/or standard deviation/error.
    a. Present summary statistics in 1-2 tables. These tables should include useful information about your data, and at a minimum should include:
        i. At least one binary variable
        ii. At least one continuous variable.
    b. One of these tables should report subgroup summary stats, where your sample is divided into subgroups based on a binary or limited variable (e.g., average wages by male and female workers). For this analysis, test if means differ significantly across these groups, and report this information using either confidence intervals, p-values, or both.
3. *Descriptive Figures*
    a. Represent at least one quantitative variable graphically, using a histogram or other well-designed figure.
4. *Correlation and causation:* For the two variables in your research question:
    a. Identify reasons why the variables might be positively or negatively correlated. Might one cause the other to increase or decrease? Is reverse causation possible? Are there outside factors that might cause both variables to move? If so, what additional data could be collected and examined, to control for these outside factors?
    b. Visualize the relationship between your variables in a well-designed figure.
    c. Compute, report, and interpret the correlation coefficient between your variables.
5. *Regression*

    Regress one variable Y on another variable X, and answer the following. You may include control variables as needed

    a. Report the slope and intercept parameters for your regression. Interpret your slope parameter in the context of your two variables.
    b. Give a confidence interval for β_1, at the level of your choice.

c. Perform a one- or two-sided test, at the level of your choice, of the hypothesis that β_1 equals a benchmark value of your choice. State the associated p-value and interpret the results.

d. Predict Y for some special value of X of your choice. Interpret this value.

Part 3 – Executive Summary

After your results section above, include a summary addressed to the audience identified above (this could be a discussion section for an academic article, or a memo to policymakers). Summarize and explain any patterns in your data that you find interesting or useful. In your report, you should do the following:

1. Clearly state the question or issue that this analysis addresses.
2. Make sure the nature of the data, including key variables, is clear to the executive.
3. Clearly explain key findings based on the results section.
4. Explain and emphasize the practical significance of any key results. Include any policy recommendations that your analysis favors.
5. This is key: Be clear and forthright about any caveats, assumptions, or limitations of your data, your analysis, or your policy recommendations, particularly questions of causation. Indicate what additional data, analysis, and assumptions would be necessary to answer the questions of interest more completely.
6. Organize your thoughts effectively.
7. Write ideas clearly, cleanly (e.g. without grammatical errors), directly, and succinctly (executive audiences don't have time for tangents and roundabout arguments). As if your audience has only a limited knowledge of statistics, avoid overly technical jargon. (For example, units of dollars are easier to understand than standard deviations or correlation coefficients.)