Econ 204
Spring 2015

LAB PROJECT INSTALLMENT 3: DATA REPORT

The main goal of this installment is to prepare your data for analysis. You won't start the analysis until the next installment of this project; for this installment, you will get your data cleaned up and organized so that it is all ready to be analyzed. This will entail obtaining all the data you need, importing it into Stata, and processing it as necessary (cleaning and merging files, generating new variables, etc.). When you have completed this installment, you should have created one or more clean, well-organized data files that contain all the data you will need for your project, with anything extraneous removed. I will refer to the cleaned and organized data file (or files) that you will use for your analysis as your "final data file (or files)."

A critically important aspect of this installment is beginning the process of documenting the sources of your data and all the steps of data analysis and management that you do to transform the data in the original data files into the final data files you use in your analysis. Documenting your work will be an ongoing process from now until the completion of this project.

The work you turn in for this installment will consist of a (printed) written essay and a set of electronic files that document your work.

The written essay will include three major components: (i) a statement of the questions you will investigate; (ii) a verbal overview of your data set, including descriptions of the contents of the data file or files you are using for the project and the source or sources from which you obtained the data; and (iii) a data appendix that shows definitions, coding and basic descriptive statistics for each of the variables in the final data set you have constructed for your analysis.

The electronic documentation will include: (i) your original data files, accompanied by appropriate metadata, (ii) importable versions of the original data files (if they are necessary), (iii) do-files containing Stata commands that import, clean and organize the data to create the final, cleaned dataset you will use for your analysis, (iv) a do-file that produces all the summary statistics and figures you present in the data appendix, and (v) a read-me file.

THE WRITTEN REPORT

The three components of the written report should be incorporated into one integrated and consistently formatted essay. This essay should be written in complete and grammatically correct sentences, organized in logically structured paragraphs, with the paragraphs assembled in such a way that they constitute one coherent essay.

Remember to follow the guidelines in the Writing Style Guide and the handout on APA style, both of which are posted on Moodle. Failures to follow those guidelines on this installment will result in large grade reductions.

(i) Statement of research question

This part of the report should include a clearly articulated statement of the particular questions or issues you plan to investigate in your project. It may be similar to the description of your research topic that was presented in your project proposal. But if your thinking has developed in any significant ways on the basis of the feedback I gave you on your proposals and/or on the basis of the independent work of your group, the developments in your thinking should be reflected in the statement of the research question.

Your ultimate goal is to develop a question or set of questions that are compelling, well-defined, and rooted firmly in previous research on your topic. As we have discussed, formulating a research topic with these qualities usually takes time—much of which must be spent reviewing previous studies, reflecting on the questions they asked and the evidence they presented, and then thinking synthetically and creatively about further research that would build on or contribute to the existing research in a useful or interesting way. Since this process takes time, it is one that you should be working on continuously throughout the semester.

It is in the installment that will come after this one—the preliminary draft—that it will be really essential to have nailed down a topic with a good hook and to know how your project will relate to previous research. In this installment, the main focus is on getting your data ready, which means that the conceptual development of your topic will not be the main focus. That does not mean, however, that you should forget all about the conceptual development of your topic as you work on this installment. In particular, here are three really crucial things to keep in mind:

> --As described above, the description of your topic that you give in this installment should reflect any developments in your thinking since you turned in your proposal.
>
> --Also as described above, reviewing and synthesizing previous research and generating a good research topic is a process that takes a long time and that you should be working on continuously. So even if this is not the installment in which the conceptual definition of your topic is front and center, you should be making progress on that aspect of the project at the same time that you write your data report over the next couple of weeks.
>
> --Finally, to write a useful data report, you need to have defined your topic well enough that you know what data you will need. If you have just a vague idea of what your topic will be, but still don’t have much a focus, you could end up putting together, cleaning up and writing a report about a bunch of data that turn out not to be relevant to what you actually end up doing.

(ii) Overview of the data

Broadly speaking, in this section of your essay you should describe the contents and structure of your final data, as well as the sources from which you obtained your original data.

You should mention and give some information about every source from which you obtained any of the data that you plan to use for your project. How much you should say about each source of data will vary, depending on the nature of the source. If you are using a well-known source (like the Current Population Survey produced by the Bureau of Labor Statistics or the World Bank's World Development Indicators), just naming the source may be sufficient. If the data source is not well known, it may be useful to provide some additional information about it.

Exactly what additional information will be relevant will depend on the questions you are investigating and the nature of your data, so it is impossible to give specific instructions that will be applicable for all groups. One way to think about what to include in your data overview is to think about how you would describe your data to someone who has a basic idea of what working with statistical data is about, but who is not familiar with your project or the data you plan to use—maybe a friend who took this course last year. What would you have to tell that person to give her or him a concrete understanding of what your data consist of, where they came from, and how they are organized?

Thinking about questions such as the following might be help you decide what information would be useful to include in the data overview. Note, however, that these questions are ***examples*** of the kind of questions that might be useful to think about as you decide what the most important aspects of your data are, and how to present this information in your essay. They are ***not meant to constitute an outline or a checklist*** of what you should include in the data overview. You should use your judgment to decide what questions are relevant for your project and how to organize and present the relevant information.

> --What is the "unit of analysis" of your data? Usually this just means indicating what each ***row*** of the data represents (e.g., a person who responded to an interview, a country, a country/year pair, or one incident of an event such as a crime or a recession). How many observations are there in your dataset?
>
> --Is your dataset a cross-section? A time-series? A panel?
>
> --What is the total number of observations in your data set?
>
> --If your data can be broken down into sub-samples by one or more categorical variables (e.g., country and year), what are the categories (e.g., what countries? what years?), and how many observations do you have for each category?
>
> --What is the scope of your sample; i.e., how is your sample defined or delimited? For instance, does your sample consist of all the people who responded to a certain survey? Or does your sample consist of observations for a certain group of countries for a certain year (in which case you should indicate which countries are included—possibly

in a table—and what year the data represent)? Or does it consist of annual observations for a certain group of countries over a certain number of years (in which case you should indicate which countries and years you have data for)? Or monthly observations?

--What population was your sample drawn from? All adults in the US? All adults in PA? All counties in the US? All individuals convicted of felonies in the US during a certain time period?

Finally, the data overview section of your essay should include a concise verbal description of every variable in your data set (other than variables that serve purely as labels, like country names or ID numbers of survey respondents). If your data come from more than one source, indicate which source each variable was taken from. Explain what the variable represents or how it is defined, and how it will be useful in your study. Also discuss any other aspects of the variable—like units of measurement, scale and coding—that would be useful for the reader to understand. You will provide additional technical details in the data appendix, described below; in the data overview, the objective is to give the reader enough information to understand concretely what all the variables represent.

(iii) Data appendix

The data appendix should contain an entry for every variable in your final data file or files.

For each variable you should give the name that you are using for the variable. State the source from which the variable was taken. (This should be one of the sources described in the data overview. If all your data came from one source, you may indicate this fact just once at the beginning of the data appendix, and then not repeat it in the entry for every variable.) Indicate the number of observations of the variable that are missing (out of the total number of observations in your data set), and give a brief description of what the variable measures. Give the name or label for the variable that was used in the source from which you obtained it.

For every categorical variable, indicate the coding scheme for the variable. Present a frequency distribution for all the non-missing observations of the variable. If there is an ordering to the categories of the variables, be sure that they are listed in that order in the frequency distribution, and include the cumulative percent frequencies (as well as the frequencies and percent frequencies). If there is no ordering to the categories, do not include the cumulative percent frequencies. Finally, present a bar graph showing the number of observations in each of the categories (unless for some reason that is infeasible—e,.g., if the number of categories of a variable is very large).

For every quantitative variable, indicate how the variable is defined and what the units of measurement are. For example, if the variable measures income, is it the income of an individual, or the total income earned by everyone in the individual's household? Is it measured in dollars, pesos or yen? Corrected for inflation (and if so, what is the base year)? Pre-tax or post-tax? Including government transfers like social security or any form of welfare payment? Report the mean, median, standard deviation, minimum and maximum of the variable, and finally present a histogram.

Formatting the written report

The written report for this installment should be organized and presented as follows:

Title page. The title page should give a title for the assignment, the names of everyone who contributed to the assignment, and the date you turn it in.

The title page should not be numbered, but in principle it counts as page 1 of your report.

Essay. This should be a coherent essay that includes a statement of your research question and a verbal description of your data (as explained in points (i) and (ii) above).

A page number should appear on every page of the essay. The first page of the essay should be numbered as page 2.

Data appendix. The data appendix (described in item (iii) above) should begin at the top of a new page, immediately following the essay. The heading "Data Appendix" should appear centered at the top of the first page of the data appendix.

A page number should appear on every page of the data appendix. The numbering of the pages of the data appendix should pick up where the numbering of the essay left off. (That is, if the last page of the essay is page *n*, then the first page of the data appendix should be numbered *n*+1.)

Reference list. The reference list should begin at the top of a new page, immediately following the data appendix. The heading "Reference List" should appear centered at the top of the first page of the reference list. Your reference list should contain citations of the sources of your statistical data, as well as of books, articles, etc. you have consulted.

A page number should appear on every page of the reference list. The numbering of the pages of the reference list should pick up where the numbering of the data appendix left off.