Problem Set #4: Using R and other tools

Economics 435: Quantitative Methods

Fall 2011

Some academics and politicians have argued that public sector unions can be a powerful political force in increasing the size of government. Public sector employees have a clear financial interest in a larger government, and unions may enable them to organize politically to pursue this interest (along with other more altruistic goals, of course).

You and I will be writing a paper together that evaluates this argument. The current draft of the paper is available on the course website at `http://www.sfu.ca/~bkrauth/econ435/restricted/psu0.docx`. I have set up the basic skeleton of the paper, but I will need you to fill in the rest over the course of the term. The idea here is to get used to some of the mechanics of writing a research paper in a controlled setting before you write your own term paper.

PLEASE READ THE INSTRUCTIONS HERE CAREFULLY. IN PARTICULAR, PLEASE DON'T TRY TO DO THE ENTIRE PAPER RIGHT NOW. JUST DO THE PARTS I AM SPECIFICALLY ASKING YOU TO DO.

Since the product of your work will be computer files, you will need to turn in your assignment using WebCT. I will want you to give me the Word document with your revisions, along with all computer files needed to generate the results (i.e., the CSV files and R script).

1 Public sector unionization and size of government: Part I

We will start by gathering some data and calculating summary statistics.

Before we get started, a mention of the principles of good data and project management. The most important principle is to always remember that you will make mistakes, and that you will not remember what you have done. When you run into a problem, the key to fixing it will be the ability to retrace your steps.

- Set aside a directory specifically for each project. Keep everything you do together in that directory (don't just save it to the desktop, for example).
- Anytime you get raw data from an external source, keep an unaltered copy of the original file, and document where it came from. If you are going to edit the file, make a copy under a new name, and edit the copy.
- When creating new files, choose informative names, but avoid spaces in names.
- When manipulating and setting up data, write script files in R instead of doing things by hand (or by menus).
- Sometimes you cannot avoid doing something by hand. If you are doing anything by hand, document it. You can do this by creating a little text file or Word file, and typing in a few notes.

Now, the data we need:

- First we will get some data on public sector unionization. Barry Hirsch (Trinity University) and David Macpherson (Florida State University) are kind enough to provide data they have gathered on this subject at `http://www.unionstats.com/`. Get the Excel file containing 2003 state-level data on union membership, coverage, density, and employment.
- Next, we will get some data on state government payrolls from the US Census Bureau. The data set is at `http://www.census.gov/govs/apes/index.html`. Get the Excel file containing the 2003 summary of estimates by function at state level.
- Next, since the payroll data is in absolute terms rather than per capita terms, we need to get some population figures for each state from the Census Bureau. An Excel file containing this information is available at `http://www.census.gov/popest/states/NST-ann-est.html`.
- Finally, we will read these three data sources into R, and merge them into a single data set that contains 2003 state-level cross-sectional data on all of the variables we want. The data set will have the following features:
  - Each row will correspond to a state.
  - Each column will correspond to a variable.
  - The variables will include full state name, total full time employees, full time equivalent employment, and total March payroll (all from the payroll data set), as well as total public sector employment, total and percent public sector union membership, and total and percent public sector union coverage (from the unionization data set) and (July 1, 2003) population (from the Census Bureau's estimates).
  - You should also add a variable for each state's 2-letter postal abbreviation.
  - Some of the variables will be observed for the District of Columbia. D.C. is not a state and does not have the same level of autonomy from the federal government as a state does, so it is probably inappropriate for our purposes. Exclude D.C. from your data set.

During class I will help you to get one of these data sets into R. The process is:

1. The Excel data files are formatted for reading by humans, not computers. We will need to change the formatting to be read by R. This includes:
   (a) R does not directly read Excel files, so we will need to use Excel to convert each file into a comma-separated values (CSV) file. This is just a text file in which each cell is separated from the next cell by a comma.
   (b) R is going to interpret the top row of the CSV file as variable names and the remaining rows as observations. So we need to make our Excel file look like that.
2. Once we have the CSV file, we can use the R command `read.table` or `read.csv` to read it in.
3. We will then have to check the resulting data set in R to make sure it looks like we think it should (`names` is a useful command here, as is `summary`), and then clean it up:
   (a) Get rid of D.C.
   (b) Merge the state postal abbreviations into each of the 3 data sets. The postal abbreviations are already available in R, in a variable called `state.abb`. You might find `merge` to be a useful command here.
   (c) Merge the 3 data sets together into a single data set. `merge` is definitely a useful command here.

(d) Add any calculated variables.

Once you put together your data set, we will calculate some basic summary statistics.

a) Use your data to fill in Table 1 in the paper. In creating Table 1, useful R commands include `summary`, `mean`, `median`, `var`, `sqrt`, `min`, `max`, and especially `apply`.

b) Use your data to create Figure 1 and Figure 2 in the paper. Adjust the captions as needed. In creating Figure 1, you will find the `hist` command useful (try `density` if you're feeling ambitious). In creating Figure 2, you will find the `plot` command useful. If you have time, you can use the `text` command to plot each point with the 2-letter state abbreviation instead of just as a generic point.

c) Complete Sections 2.2 and 4.1 in the paper.

2 Public sector unionization and size of government: Part II

Find three articles that we will want to cite in Section 1.1 ("Related literature"). Try to find the most important articles on this subject, and try to find a mix of published articles and working papers.

Cite those articles to the Related Literature section of the paper (Section 1.1), and put bibliographic references in the References section of the paper. You do not need to read these articles for this assignment, and your citations in Section 1.1 do not need to be any more elaborate than "Related articles include [insert list of papers here]."