Software Overview

Paul E. Johnson[1] [2]

[1] Department of Political Science

[2] Center for Research Methods and Data Analysis, University of Kansas

2014

Outline

1 Overview

2 Common Chores

3 Replication

4 Naming Conventions

Outline

1 Overview

2 Common Chores

3 Replication

4 Naming Conventions

What is this Presentation

- Opinions about research
- Statistical packages
- Miscellaneous bits of advice I've been eager to share to new students (and prospective GRAs in the Center for Research Methods and Data Analysis).

Software Frameworks

- There is no “perfect” framework for doing statistical analysis.
- I’ve used: SPSS, SAS, SST, LimDep, SyStat, WinRATS, S+, Stata, R, GRETL, WinRATS, Matlab, Octave, Minitab. If you count specialty programs, add Mathematica, Swarm, OpenBUGS, JAGS, MPlus.
- These differ in
  - Cost
  - Convenience
  - Sophistication (fancy models!)
  - Openness to user-contributed components

Your Career Goal May Affect Your Choices

- I tell everybody to learn R now. I can get huge-scale projects done with R, doubt any alternative exists.
- But
  - If you are going into industry, chances are good they want you to use SAS or SPSS
- Stata is a low priority as far as industry is concerned, but it is very widely used in some social sciences.
- If you could collect up all of the articles that introduce new statistical methods, I predict that the software they use would be

  R (30%), Matlab (25%), Stata (10%), BUGS (Open/WinBUGS, JAGS) (10%), SAS (10%), SPSS (0.000001%), and the rest

The Biggest Difference? The "One Dataset" philosophy

- SPSS and Stata insist the user can access only one rectangular data set at a time. Every user interaction with the program is framed by this limitation.
- SAS and S/R have always allowed users to have many datasets in the same session.
  - SAS insists that sets must be merged into one block before use in most statistical procedures.
  - R similar, not quite so strict.

Another Big Difference: Interrogatability of Results

At one end of the spectrum, we cannot interact with results at all

- SPSS “plops” out some estimates. There’s a big splat of printout. And the analysis is finished

Some programs are not quite so limiting

- SAS procedures provide an option to save elements from results into new datasets, which can then be investigated

Some are at the opposite end altogether

- S/R procedures generally provide no printed output. They save structured “objects” which the user must interrogate the result. It is impossible to get anywhere without investigating it.

Open Source: A priority?

- Some people just want a program that "works". They don't mind paying. They don't care why it works (It "just works")
- That's a natural, understandable human tendency (lazy).
- R is an open source "free software" program. GRETL and PSPP are open. If you wonder "how do they calculate a regression model?" go look at their source code.
- John Chambers (*Software for Data Analysis*, 2008) Prime Directive: researchers must be accountable to for all calculations, beginning from initial data to final report.

Outline

1 Overview

2 Common Chores

3 Replication

4 Naming Conventions

Things We Would Like From Every Program

- Bring in data, from a diverse set of input formats
- Reorganize, “Recode” variables
- Statistical procedures
- Nice Graphs

What are we looking for?

It appears the things we can't live without in CRMDA are

1. Scriptability
2. Cutting edge statistical models

Scriptability

We don't want a program that won't allow us to write scripts.

- Need to run a program 1000s of times. Must Automate
- Scripting is part of replication: script the creation of tables and plots

Numeric versus Categorical Variables

- All of these programs can handle numeric data
- Programs vary enormously in the way they handle categorical information
- Do you think of a variable $\{Male, Female, Female, Male, Male\}$ as a 1) character, or 2) label
- SPSS was first. They decided
  - assign numeric values, any integers the user desires
  - assign a separate set of "value labels" that can be used in presentations about the data
  - Problem: statistical procedures intended for numeric information would treat the categorical data as if it were a number.

Categorical Variable Revolution

- The S language & interpreter was introduced in the 1980s. One of its primary distinguishing characteristics was the new way of treating categorical variables.
  - Called them “factors”, users are encouraged to think of the substantive label as the actual value of the variable (discouraged of thinking of their underlying numeric values)
  - Statistical and Graphing procedures procedures were customized in the handling of factor variables.
- SAS, SPSS, Stata later followed, to various degrees.

The Program Editor/Environment Problem

- R doesn't really try to provide a user interface (except on Mac); they expect you will use an editor like Emacs, RStudio, Notepad++, WinEdt, Eclipse, TINN-R, or whatnot.
- SAS and Stata do provide editors and a somewhat richer environment, but still not very convenient.
- By far the best "provided with" user environment is in the Matlab program.
- I urge everybody to learn to use the editor Emacs.
  - multi-program, multi-platform
  - has features for preparing stat scripts that no other editor can match
  - long blather about it:`http://pj.freefaculty.org/guides/Rcourse/emacs-ess/emacs-ess.pdf`

Outline

1 Overview

2 Common Chores

3 Replication

4 Naming Conventions

Replication is Priority #1

- The ability to repeat an analysis from start to finish is the single highest priority.
- We need to replicate
  - across teammates
  - across time
  - across computers
- There must never be any
  - unaccounted for "point and click" revision of a data set
  - "copy and paste" translation of information from one program to another
- If everything works correctly, I should be able to take your work folder, open it, and re-do every step

Scripts!

- Discourage
  - spreadsheets (ex Microsoft Excel)
  - any menu-driven statistical packages that make scripting seem obscure & uncomfortable
- Encourage
  - Write scripts that import data, re-compute values, draw plots, and make tables.
- If you interact with a program, do it to cultivate your script of commands.

Write Portable Scripts

- In a perfect world, your project folder can be translated to another computer and it can run AS IS!
- Implication: Avoid writing things like this in your scripts:

  ```
  setwd(”C:\users\pauljohn\myproject\version1”)
  ```

  because that command will only work for somebody who is named pauljohn

Workflow: Tables and Plots

- Bad scenario: A result “pops out” on the screen! Researcher
  - copies the results onto a napkin (er, types a table in a paper)
  - takes a screenshot of a plot
- Good scenario:
  - The script is smart enough to write tables and plots, which go into paper without modification!
  - Revisions in the data are easily incorporated in updated reports (no tedious re-typing)
- Possible in varying degrees with R, Stata, SAS

Batch mode: The Ground Floor

- The best stats programs are often delivered with the worst editors
- SPSS, SAS, R and MPlus originally provided no editors. One would
  - Write the script file
  - Run the script file non-interactively, in “BATCH” mode
- This is like “writing letters to yourself” by saving them on disk

Prog	suffix	example “run” command	Usual output files
SAS	.sas	sas whatever.sas	whatever.lst, whatever.log
R	.R	R CMD BATCH whatever.R	whatever.Rout, other files
Stata	.do	xstata -b do whatever.do	log and output files
Mplus	.inp	mplus whatever.inp	whatever.log, whatever.out

Outline

1 Overview

2 Common Chores

3 Replication

4 Naming Conventions

Cultivate a standard naming convention

- Project Directory. Everything from a project.
- Make subdirectories, use standard names
- Try to cultivate a consistent style for naming files

I’ve got 2 sorts of projects going

- In a simple case with a not-massive dataset (which is not “secret” and can be backed up)

Simple Subfolders

data	from client, read-only files. not editable
workingdata	we create, intermediate analysis
output	tables and plots
writeup	reports
doc	about the project
lit	literature (pdfs, etc)

Folders for Scripts

R

SAS

Stata

Mplus

Example Folders

- For our training program, we provide 2 working example folders

Example-els	Educational Longitudinal Study
Example-mpg	Miles per gallon analysis of automobiles

- Note how, inside the R, Stata, and SAS folders, we avoid non-portable directory names.
- From inside one folder, one can "go up" by referring to the folder "../", thus from R, Stata, or SAS, the data is available in "../data". No need to write "C:\users\pauljohn\project\" or such.

Projects with Secured Data

- The "work" part of the project–the part we back up–must be kept separate from the data
- Two large subfolders, "secure_data" and "work"

Secure folder

data	from client, read-only files. not editable
workingdata	we create, intermediate analysis
confidential	information
output	maybe here, maybe in other

Not Secure Team Work Folder

R	
SAS	
Stata	
Mplus	
output	
writeup	
doc	about the project
lit	literature (pdfs, whatnot)

Coherent File Names

- We suggest program file names like this
    - download: if data is retrieved from websites
    - import: brings in data, recodes variables, creates files in workingdata
    - analysis:

Naming Versions

- Here's a problem. People re-name versions chaotically.
- But they should be more disciplined: Change file names on the end to indicate version

- If the original text version of the dataset is called

ELS-national-integrated-2012.txt

- call your imported version in Stata

ELS-national-integrated-2012-1.dta

- call your next revision

ELS-national-integrated-2012-2.dta

- Do NOT do this

"New Version of ELS.dta"
"Jimmy's approved version of ELS.dta"
"Final data version.dta"

Write dates as YYYYMMDD

- It is acceptable to use the date as a version number, as in ELS-national-integrated-2012-20140811.dta
- Note the important element: YYYYMMDD. The sorted list of file names will come out "in the right order".
- The date = version naming scheme has some appeal because it is always crystal clear when the file was initiated
- Sometimes I need Version and Release Numbers
- Expect to make several versions on one day?
  ELS-national-integrated-2012-20140811-01.dta
  ELS-national-integrated-2012-20140811-02.dta

Change Variable Names on the End

- If the original is called “v44t51” and we
  - log it, call the new one “v44t51log”
  - take the square root, call the new one v44t51sqrt
  - re-organize categories, call the new one v44t51recode

What is the Take Away Supposed to Be?

- Don't be distracted programs that offer menus and buttons.
    - you need to build a script! The buttons are not helpful otherwise
- Develop (or conform to) a consistent naming scheme that puts each of your projects in a separate folder