Statistical Software in the Census Research Data Center

Texas Census Research Data Center*

Updated: July 19, 2012

Overview

The objective of this document is to offer researchers a starting point from which to learn about the software products available for computing in a Census Research Data Center (CRDC). There are several reasons for this to be of use to (prospective) CRDC researchers:

- A researcher's preferred software is not available in the CRDC, so that an alternate software package is required to conduct statistical analysis of CRDC data.
- A researcher's preferred software is *available* in the CRDC, but is sub-optimal due to the volume of data included in the analysis.[1]
- A researcher would like to learn more about the available software and their capabilities.

This is not an exhaustive list of software available in the CRDC, and it is far from being an exhaustive list of resources available to learn about these software suites. Mainly, this document serves as a compilation of links to online tutorials, references, and advice regarding the software. Selected printed texts are also suggested for each of the included statistical packages.

Following the discussion of specific software available in the CRDC, I include some more general programming advice and suggestions/resources for working with large data sets.

*For questions on this document, contact Jeremy West at j.west@tamu.edu

[1]For instance, Stata first loads all data into the computer's memory (RAM) before a researcher can analyze them. Because of the size of many CRDC data sets, this could be problematic for some analyses.

1 Stata

Stata is arguably the most "user friendly" statistical package available in the CRDC, but this does not mean Stata is lacking in power. Stata may be used through either user-issued commands entered into a command-line or via "point-and-click" graphical user interface (GUI) contextual menus. Additionally, scripts (.do files) may be composed and executed.

Official Resources

- Official Website: http://www.stata.com
- Support Website: http://www.stata.com/support/
- FAQ (how-to for many common analyses): http://www.stata.com/support/faqs/
- The Stata Journal: http://www.stata-journal.com/
- Stata Listserve: http://www.stata.com/statalist/
- Listserve Archive: http://www.stata.com/statalist/archive/

"Unofficial" Resources

- German Rodríguez provides a concise and straightforward introduction and tour of Stata for the novice user here: http://data.princeton.edu/stata/
- Some additional Stata resources at Princeton are listed here: http://www.princeton.edu/wwac/academic-review/stata/
- One of the most comprehensive online resources for learning Stata is provided by UCLA: http://www.ats.ucla.edu/stat/stata/
- In particular, novice Stata users may find the "Stata Starter Kit" beneficial: http://www.ats.ucla.edu/stat/stata/sk/
- A "function-oriented" tutorial and list of many commands is provided by UNC: http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial
- Andrew Grogan-Kaylor's "Two-page Stata" provides a concise way to hit the ground running: http://www-personal.umich.edu/~agrogan/stata/TwoPageStata.pdf
- At a slightly more advanced level, Gabriel Rossman has a nice guide to basic programming in Stata: http://gabrielr.bol.ucla.edu/stataprogramming.pdf
- A number of websites offer video tutorials, but many videos for learning Stata can be found on YouTube: http://www.youtube.com/results?search_query=stata+tutorial

Some Additional Resources

[1] Alan C. Acock. *A Gentle Introduction to Stata*. Stata Press, 3rd edition, 2012.

[2] Ulrich Kohler. *Data Analysis Using Stata*. Stata Press, 2nd edition, 2008.

[3] A. Colin Cameron and Pravin K. Trivedi. *Microeconometrics Using Stata*. Stata Press, 2nd edition, 2010.

[4] Christopher F. Baum. *An Introduction to Stata Programming*. Stata Press, 1st edition, 2009.

[5] Kyle C. Longest. *Using Stata for Quantitative Analysis*. Sage Publications, 1st edition, 2011.

[6] J. Scott Long. *The Workflow of Data Analysis Using Stata*. Stata Press, 1st edition, 2008.

[7] Michael N. Mitchell. *Data Management Using Stata: A Practical Handbook*. Stata Press, 1st edition, 2010.

[8] Michael N. Mitchell. *A Visual Guide to Stata Graphics*. Stata Press, 3rd edition, 2012.

2 R

R is an open-source language and environment for statistical computing that originated from the S language and environment developed at Bell Laboratories. A strength of R is the availability of numerous user-created packages and statisical routines available in the Comprehensive R Archive Network (CRAN) repositories. Although users may issue command-line prompts directly, R is designed to operate primarily by running programming routines (R-scripts).

Official Resources

- Official Website: http://www.r-project.org/
- CRAN: http://cran.r-project.org
- FAQ: http://cran.r-project.org/doc/FAQ/R-FAQ.html
- The R Journal: http://journal.r-project.org/

"Unofficial" Resources

- R Wiki: http://rwiki.sciviews.org/doku.php
- Emmanuel Paradis' *R for Beginners*: http://scicomp.evergreen.edu/docs/workshops/RforBeginners.pdf
- Tutorials and examples at UCLA: http://www.ats.ucla.edu/stat/r/
- Mahmood Arai has an econometrics-focused introduction to R: http://people.su.se/~ma/R_intro/R_intro.pdf
- Patrick Burns offers a—dare I say—humorous (and thorough) guide: http://www.burns-stat.com/pages/Tutor/R_inferno.pdf
- However, Burns' "Hints for the R Beginner" is perhaps a better place to start: http://www.burns-stat.com/pages/Tutor/hints_R_begin.html
- In keeping with the open-source nature of R, a (GNU) free book is provided online: *Introduction to Probability and Statistics Using R*, by G. Jay Kerns: http://ipsur.org/index.html
- For those who prefer video tutorials, Jeromy Anglim offers a nice list: http://jeromyanglim.blogspot.com/2010/05/videos-on-data-analysis-with-r.html

Some Additional Resources

[1] Alain F. Zuur. *A Beginner's Guide to R*. Use R! Springer, 1st edition, 2009.

[2] Norman S. Matloff. *The Art of R Programming: A Tour of Statistical Software Design*. No Starch Press, 1st edition, 2011.

[3] Paul Teetor. *R Cookbook*. O'Reilly Cookbooks. O'Reilly Media, 1st edition, 2011.

[4] Peter Dalgaard. *Introductory Statistics with R*. Springer, 2nd edition, 2008.

[5] Michael J. Crawley. *The R Book*. Wiley, 1st edition, 2007.

[6] John Verzani. *Using R for Introductory Statistics*. Chapman and Hall, 1st edition, 2004.

[7] Phil Spector. *Data Manipulation with R*. Use R! Springer, 1st edition, 2008.

3 SAS

SAS software is an integrated system of software products, as opposed to being a single standalone product. As such, SAS software may be used for database management and similar needs. In the CRDC, SAS software is available for data analytics, and this document focuses on this application for SAS software. Processing and analysis of data using SAS is conducted by writing programs, which are then executed ("run").

Official Resources

- Official Website: http://www.sas.com/
- SAS Resource Center: http://www.sas.com/resources/
- SAS Support: http://support.sas.com

"Unofficial" Resources

- SAS-L community and listserve: http://www.sascommunity.org/wiki/SAS-L
- Texas A&M's Statistics department hosts a nice list of tutorials: http://dist.stat.tamu.edu/flash/SAS/
- Dwight Galster has a solid list of guides as well: http://galsterhome.com/stats/Tutorial/sas_tutorial_contents.htm
- As with Stata and R, UCLA provides a comprehensive set of resources: http://www.ats.ucla.edu/stat/sas/
- The UCLA "SAS Starter Kit" is a great starting place for novice SAS software users: http://www.ats.ucla.edu/stat/sas/sk/modules_sk.htm
- Karl L. Wuensch's compilation of lessons and programs for SAS Base: http://core.ecu.edu/psyc/wuenschk/SAS.htm
- For those seeking video tutorials, the StudySAS Blog has compiled links to many online video guides: http://www.studysas.blogspot.com/2008/08/sas-tutorials-video-free.html

Some Additional Resources

[1] Ron P. Cody. *Applied Statistics and the SAS Programming Language*. Prentice Hall, 5th edition, 2005.

[2] Lora Delwiche and Susan Slaughter. *The Little SAS Book: A Primer*. SAS Institute, 4th edition, 2008.

[3] Andy Field. *Discovering Statistics Using SAS*. Sage Publications Ltd., 1st edition, 2010.

[4] Alan C. Elliott and Wayne A. Woodward. *SAS Essentials: A Guide to Mastering SAS for Research*. Research Methods for the Social Sciences. Jossey-Bass, 1st edition, 2009.

[5] SAS Institute. *SAS Certification Prep Guide: Base Programming for SAS 9*. SAS Institute, 3rd edition, 2011.

[6] Michele Burlew. *SAS Macro Programming Made Easy, Second Edition*. SAS Publishing, 2nd edition, 2007.

4 Matlab

MATLAB is both a programming language and numerical computing environment developed and maintained by MathWorks. As the software's name indicates, MATLAB is focused on matrix operations and manipulation, but MATLAB also allows for generating graphics and object-oriented programming. Although MATLAB includes a command-line and selected GUI contextual commands, most programming in MATLAB is conducted by composing and executing scripts.

Official Resources

- Official Website: http://www.mathworks.com/products/matlab/
- Official Examples: http://www.mathworks.com/products/matlab/examples.html
- Technical Documentation and Demos: http://www.mathworks.com/help/techdoc/index.html
- MATLAB Digest: http://www.mathworks.com/company/digest/current/
- Getting Started: http://www.mathworks.com/help/techdoc/learn_matlab/bqr_2pl.html
- Video tutorials: http://blogs.mathworks.com/videos/

"Unofficial" Resources

MATLAB's official resources are well-organized and approachable, but some additional resources are available at:

- Unofficial wiki FAQ: http://matlab.wikia.com/wiki/FAQ

- Mark S. Gockenbach's "Practical Introduction to MATLAB": http://www.math.mtu.edu/~msgocken/intro/intro.html
- Duke's "Partial List of On-Line MATLAB Tutorials and MATLAB Books": http://www.duke.edu/~hpgavin/matlab.html
- Somewhat more commercially-oriented, but nonetheless useful: http://www.matlabtutorials.com/

Some Additional Resources

[1] Stormy Attaway. *MATLAB: A Practical Introduction to Programming and Problem Solving*. Butterworth-Heinemann, 2nd edition, 2011.

[2] Rudra Pratap. *Getting Started with MATLAB: A Quick Introduction for Scientists and Engineers*. Oxford University Press, 1st edition, 2009.

[3] Amos Gilat. *MATLAB: An Introduction with Applications*. Wiley, 4th edition, 2010.

[4] Peter Kattan. *MATLAB for Beginners: A Gentle Approach*. Peter I. Kattan, Revised edition, 2009.

[5] David McMahon. *MATLAB Demystified*. McGraw-Hill, 1st edition, 2007.

[6] Timothy A. Davis. *MATLAB Primer*. CRC Press, 8th edition, 2010.

[7] David A. Rosenbaum. *MATLAB for Behavioral Scientists*. Psychology Press, 1st edition, 2007.

5 Programming Advice

Because of the extensive scope and prolonged duration of CRDC projects, as well as the volume of data included, programming etiquette is likely of higher importance in the CRDC than for many data analyses. In this section, I provide links to some programming advice and suggested "good behavior."

Writing Code

- Writing computer code, by Paul Murrell. Although this document is focused on HTML, the principles apply generally: http://www.stat.auckland.ac.nz/~paul/ItDT/HTML/node16.html

- Matthew Gentzkow and Jesse M. Shapiro's advice to their RAs on Writing Code (Economics-oriented): http://faculty.chicagobooth.edu/matthew.gentzkow/research/ra_manual_coding.pdf

Working with Large Data Sets

R

- CRAN suggestions: http://cran.r-project.org/web/views/HighPerformanceComputing.html
- "Taking R to the limit": large data sets in R: http://www.bytemining.com/2010/08/taking-r-to-the-limit-part-ii-large-datasets-in-r/

Stata

- Stata: http://www.stata.com/support/faqs/data-management/large-datasets/
- Some of my suggestions for Stata with large data sets:
  - As you generate them, define new variables as byte, integer, etc. if you know they will only take on small values (e.g. indicator terms)
  - Make liberal use of the *compress* command throughout your code
  - Consider breaking the data set into parts for some operations (not regression and analysis, of course); then, use loops or other means to process the pieces separately before rejoining the completed pieces. Pending availability of computing resources, you might consider running multiple instances of Stata simultaneously.

MATLAB

- Mathworks' suggestions for efficient use of memory: http://www.mathworks.com/help/techdoc/matlab_prog/brh72ex-25.html
- Iain Murray provides some useful advice on pre-allocation of matrices and other "tricks" for handling larger data sets more efficiently: http://homepages.inf.ed.ac.uk/imurray2/compnotes/matlab_octave_efficiency.html

6 Future Extensions to this Document

Some further additions that may be included in this document:

- Links to guides for moving data *between* these and other software formats
- Other software in the CRDC (e.g. GIS software)
- Additional resources, either ones that I stumble across or those suggested to me