Documentation of Social Science Data with STATA

Haile Selassie Okuku
Data & Open Access Consultant

29th September 2018

Background

- OA & OD policies make more and more social science data available for secondary analysis
- In secondary data analysis, documentation plays a critical role in transferring knowledge about data from data producers to secondary users
- Documentation (aka metadata) usually includes codebooks & data dictionaries, related bibliographies and data collection instruments
- Metadata serves 3 main purposes: resource discovery; preservation; & administration
- Documentation for social science data is mainly used for resource discovery (searching and judging the relevancy of the data) and secondary analysis

Metadata

- 4 types of metadata are needed for using and archiving social science data [1]
  - Study-level metadata: abstracts, study descriptions - purpose of the study, the major conceptual categories studied, the characteristics of the sample, measures, etc.
  - File-level metadata: describe the properties of individual files in a data collection
  - Variable-level metadata
  - Administrative and structural metadata : These are critical to ongoing maintenance and preservation of the electronic data collections
- Documentation using stata refers to the 2nd and 3rd categories of metadata

[1]Gutmann, M., K Schürer, D Donakowski and Hilary Beedham. the selection, appraisal, and retention of digital social science data. Data Science Journal, Volume 3, 30 December 2004

Stata do-file

- STATA Interface: There are 3 ways to communicate:
  - Interactive mode (Real Time)
  - non-interactive mode (batch mode)
  - Point-and-click (Menu based)
- Stata documentation is best done using “a do-file”: text editor that saves commands and comments
- This session is a sample work session, introducing a few of the basic tasks that can be done in Stata

Best Practices in Variable Naming

- Use characters (a- z and A-Z), numbers (0-9), or underscore (_) only. Do not use special characters such as -, space, ~, !, @, #, $, %, ^, &, *, (,), {, }, [,], <, >, ?, and /.
- Begin with a letter. It is because underscore is often used in system variables such as _N, _n, _pi, _b, _coef, and _cons
- The shorter, the better. Do not exceed 10 characters unless necessary, though STATA allows up to 32 characters as a variable name
- Avoid reserved words or keywords (e.g., command and function)
- Use meaningful names associated with contents of the variable
- Make it consistent and systematic. You can benefit from using array and wild card as in score1 - score10, score??, vote*
- Use lower cases unless necessary or required
- Use underscore instead of space

Examples of Naming Rules

Good Example	Bad Example	Description
gnp2002 gnp-2002	gnp#2002	Avoid special characters
real_int	real interest rate	Use underscore
score1; gnp2003	1st_score; 2003gnp	Begin with a character
reg_out; glm1	REG; glm; ttest	Avoid reserved words
invest; interest	xxx; yyy; zmdje	Use meaningful names
score1; score2;…	math; math_1; math02	Consistent and systematic
citizen	Are_you_a_US_citizen?	The shorter, the better
income; intUS03	INCOME; Int_us2003	Use lower cases

Length of Names and Labels

Keyword	Maximum Length	Notes
Variable Name	32 characters	.gen; .egen
String Variable	244 characters	
Dataset Label	80 characters	.label data "…"
Variable Label	80 characters	.label variable var_name "…"
Value Label Name	32 characters	.label define lbl_name # "…";
Value Label	32,000 characters*	.label values var_name lbl_name

* The intercooled allows only 80 characters.

Stata Wildcards

Wildcards	Descriptions	Examples
?	Any character	d? (da db dc… d1 d2 d3)
*	Any characters	re* (retain return)
~	Zero or more characters	in~t (invent interest)
-	Specifying range of variables	gender-rank

General structure of do-file

- Header information: Description of the program and what is does including names, creator; dates of creation; project name
  - All put under comments (/* */, *)
- Setting environment: system parameters e.g.
  - erase everything in the current workspace memory - clear
  - Memory – set memvar, set min_memory, set max_memory
  - Default system locations – cd, tempfile
  - Tell Stata to pause or not pause for --more—messages
- Opening / importing data
  - Use – for Stata datasets
  - File → Import
  - User written – usespss, usesas
- Listing observations
  - Starting a log file named to echo the session – log using
  - Closing opened log file – log close