Literate programming

Using log2markup, basetable and matrixtools

Niels Henrik Bruun

Dept. Of Public Health, Aarhus University

AARHUS UNIVERSITY

AARHUS UNIVERSITY

Reprodiciple research

Peng, Dominici, and Zeger (2006), Peng (2009), Patil, Peng, and Leek (2016), Barba (2016)

- Data
  - A prepared dataset
    - a lot of programming behind
    - and data management
    - statistical conclusions are very sensitive to variable definition
- Analysis
  - all code used for the published article
  - analysis and reportting becoming more complex
- Accessibility
  - internet presentations and summaries?

A lot of programming (Learn from programming):

- a researcher today must be or work together with a statistical programmer
- need for documentation eg by using literate programming

AARHUS UNIVERSITY

Newer attemps on integrating Stata log output and text/comments

- weaving - Rising (2008)
- sar - Magno (2013)
- markdoc - Haghish (2016b), Haghish (2016a)
- texdoc - Jann (2016)
- webdoc - Jann (2017)
- markstat - Rodríguez (2017)
- asdoc - Shah (2018)
- dyndoc
- putdocx
- putpdf
- **log2markup**

AARHUS UNIVERSITY

Tables transform commands into eg. latex and html

- Stata commands: *statsby*, *parmby*, *matlist*
- estout - Jann (2007), Jann (2014)
- outreg - Gallup (2012a), **Gallup (2012b)**
- outreg2
- table1 - Clayton (2002)
- table1_mc - Chatfield (2017)
- tabout - Watson (2015)
- ietoolkit
- **basetable**
- **matprint**
- **sumat**
- **crossmat**
- **metadata**
- **regmat**, *to appear soon*

AARHUS UNIVERSITY

Requirements (my opinion) I

- Documents should have multiple layers/functions (Knuth!!)
  - keep thoughts, analysis and "presentation" in one document
  - not just code and commented code
- Integrate markup text blocks with Stata log outputs
  - text is the most
    - So a need for text blocks integrated with code
  - Allow do-files to be runable
    - Use Stata text marking as base for text
- Keep/hide blocks within text
  - to do (eg adding labels) and document (verification, analysis), but not show
- Make documents as flexible as possible (choose destination format lately)

AARHUS UNIVERSITY

Requirements (my opinion) II

- Show/hide commands
- Show/hide output
- Integrate table output in nice layout
- Keep external tools like pandoc out of Stata commands
  - requires needless parsing
  - there will always be shortcomings which are harder to handle

log2markup, basetable and matrixtools

An set of commands for integrating text and table output in html, tex and markdown

- **log2markup** for transforming text log output into markup files
- **basetable** for easy building of standard report table 1 in articles
- **matrixtools**, tools for reporting using Stata matrices
  - **metadata** a mix of *describe* and *sumarize* on current dataset, a dataset, or a directory (including subdirectories)
  - **matprint** a command for printing matrix content - a user friendly version of *matlist*
  - **sumat** an extension of *summarize* returning results in a Stata matrix
  - **crossmat** an extension of *tabulate* returning results in a Stata matrices
  - **regmat** - tabulation of regression results. *To appear soon*
  - and growing ...

Documented at Bruun (2017). All uses the same Mata backbone: **lmatrixtools.mata**

AARHUS UNIVERSITY

log2markup, Add markup text blocks

Command that select marked text, code and/or output blocks

- comments surrounded by `/***` and `***/` are kept for text processing
  - Only these comments are integrated with code
  - can be written in markdown, html, latex, a mix or something completely different
- Text is ignored if
  - surrounded by `/*` and `*/`
  - in lines starting with `*`
  - in lines starting with `//`

AARHUS UNIVERSITY

log2markup, code and output appearance

Prefix for integrating commands:

- /****/ Show only output from command without formatting - **SMART!!**
  - integrate table print from **basetable** and **matprint** into output
- For teaching purposes typically
  - /**/ Show only command
  - /***/ Show only output from command
  - No prefix: Show command and output (Just like Stata log)

AARHUS UNIVERSITY

log2markup, Internal blocks of code and text

Internal code blocks with comments

- Command and comment blocks surrounded by `//OFF` and `//ON` are ignored in transformed log file
  - Eg for analysis only worth showing in compressed or summarised form

Line with macro content:

- `/****/display "This is updated with ‘mymacro'"`
- It is possible to mark code and output differently

AARHUS UNIVERSITY

Make a word document out of a log2markup output using markdown

The do file **toWord.do** are transformed into the Word document **toWord.docx** by:

1. Generate a log file (/***/ ignores the command, but insert the output)

```
capture log close
log using toWord.log, text replace
/***/do toWord.do
log close
```

2. Transform log file using **log2markup**

```
log2markup using toWord.log, replace extension(md)
```

3. Use **pandoc** to create a Word document for distribution

```
shell pandoc -s toWord.md -o toWord.docx & timeout 30
```

AARHUS UNIVERSITY

pandoc

- Pandoc, see MacFarlane (2006), transforms a document in one format into another
  - usefull (almost) no matter what the original document format
- Simple command examples are (from dos promt or after Stata shell command):
  - pandoc [-s] markdown.md -o output.[html|pdf|tex|docx]
    - option -s for single full document, otherwise a fragment
    - file suffix determines type of output
- Pandoc markdown

AARHUS UNIVERSITY

mkdocs

- MkDocs, see Christie (2014), generates static HTML sites
  - fast, simple
- Source files are written in Markdown (plain markdown, Gruber (2002))
- Configured with a single YAML configuration file
- Static HTML sites you can host anywhere
- Example: StataHacks

AARHUS UNIVERSITY

Data description (using metadata from matrixtools)

Name	Index	Label	Value Label Name	Format	Value Label Values	n	unique	missing
bwlt1500	1	birthweight < 1500g	no_yes	%10.0g	0 “No” 1 “Yes”	189	2	0
bwlt2500	2	birthweight < 2500g	no_yes	%8.0g	0 “No” 1 “Yes”	186	2	3
age	3	age of mother		%8.0g		185	24	4
lwt	4	weight at last menstruation (kg)		%8.0g		189	76	0
race	5	race	race	%8.0g	1 “white” 2 “black” 3 “other”	183	3	6
smoke	6	smoked during pregnancy	smoke	%8.0g	0 “No” 1 “Yes”	189	2	0
ftv	7	number of visits to physician during 1st trimester	ftv	%8.0g	0 “0 visits” 1 “1 visit” 2 “2 visits” 3 “3 visits” 4 “4 visits” 6 “6 visits”	189	6	0
bwt	8	birth weight (grams)		%8.0g		189	133	0

Summary of metadata

- in current dataset
- in a noncurrent dataset (specified by using)
- in a folder (specified by using)
- in a folder with subfolders (specified by using + option **searchsubdirs**)

AARHUS UNIVERSITY

Basetable, description

- Easy build of a summary table (used in almost every article)
- requires labels and value labels for used variables
- Format and report options
  - Continous variables: format + sd, 95% ci, iqi, iqr
  - Categorical variables: row/col percentages, single value (ci)
- Total reported
- Comparison test: chisquare or anova(means)/Kruskal-Wallis(medians)
- Titles for groups of variables
  - Sub conditioning
- Missing values with option **`missing`**
- Export to Excel with option **`toxl`**
- With option **`style`** it can be integrated in md/csv/latex/html documents
- Hide counts less than 5 default in reports with option **`hidesmall`** (Statistics Denmark)

AARHUS UNIVERSITY

Log output

```
basetable bwlt2500 [**Quartile interval for age**] age(%6.1f, iqi) ///
        [**Counts and % for categorical variables**] race(c) ///
        [**CI for single categorical value**] smoke(Yes, ci) ///
        , /*style(md)*/ caption(A basetable demo) /*missing*/ ///
        toxl(lbw tables.xls, Table 1, replace)
```

A basetable demo:

	No	Yes	Total	P-value
n (%)	127 (68.3)	59 (31.7)	186 (100.0)	
Quartile interval for age				
age of mother, median (iqi)	23.0 (19.0; 28.0)	22.0 (19.0; 25.0)	23.0 (19.0; 26.0)	0.24
Counts and % for categorical variables				
race, n (%)				
white, n (%)	71 (57.3)	23 (39.0)	94 (51.4)	
black, n (%)	14 (11.3)	11 (18.6)	25 (13.7)	
other, n (%)	39 (31.5)	25 (42.4)	64 (35.0)	0.06
CI for single categorical value				
smoked during pregnancy (Yes), % (95% ci)	32.3 (24.2; 40.4)	50.8 (38.1; 63.6)	38.2 (31.2; 45.2)	0.02

```
Table send to Excel succesfully...
```

Integrated output using prefix /****/ and style(md)

Table 2: A basetable demo Table send to Excel succesfully...

	No	Yes	Total	P-value	Missings / N (Pct)
n (%)	127 (68.3)	59 (31.7)	186 (100.0)		3 / 189 (1.59)
Quartile interval for age					
age of mother, median (iqi)	23.0 (19.0; 28.0)	22.0 (19.0; 25.0)	23.0 (19.0; 26.0)	0.24	4 / 189 (2.12)
Counts and % for categorical variables					
race, n (%)					
white, n (%)	71 (57.3)	23 (39.0)	94 (51.4)		
black, n (%)	14 (11.3)	11 (18.6)	25 (13.7)		
other, n (%)	39 (31.5)	25 (42.4)	64 (35.0)	0.06	6 / 189 (3.17)
CI for single categorical value					
smoked during pregnancy (Yes), % (95% ci)	32.3 (24.2; 40.4)	50.8 (38.1; 63.6)	38.2 (31.2; 45.2)	0 / 189 (0.00)	0 / 189 (0.00)

AARHUS UNIVERSITY

And in Excel:

A1

	A	B	C	D	E	F
1		Normal	Low	Total	P-value	Missings / N (Pct)
2	n (%)	127 (68.3)	59 (31.7)	186 (100.0)		3 / 189 (1.59)
3	**Quartile interval for age**					
4	age of mother, median (iqi)	23.0 (19.0; 28.0)	22.0 (19.0; 25.0)	23.0 (19.0; 26.0)	0.24	4 / 189 (2.12)
5	**Counts and % for categorical variables**					
6	race, n (%)					
7	white, n (%)	71 (57.3)	23 (39.0)	94 (51.4)		
8	black, n (%)	14 (11.3)	11 (18.6)	25 (13.7)		
9	other, n (%)	39 (31.5)	25 (42.4)	64 (35.0)	0.06	6 / 189 (3.17)
10	**CI for single categorical value**					
11	smoked during pregnancy (Yes), % (95% ci)	32.3 (24.2; 40.4)	50.8 (38.1; 63.6)	38.2 (31.2; 45.2)	0 / 189 (0.00)	0 / 189 (0.00)
12						

Figure 1: The Excel file “lbw tables.xls”

AARHUS UNIVERSITY

sumat/matprint, description

- Stata matrices is the best data container we have in Stata
  - Overlooked
- `sumat` is an extension of summarize
  - Summarize summarised in matrix
  - Handles string variables when possible
  - More functionality, eg ci, iqi, unique, missing etc
    - Option `rowby`
    - Result are returned in `r(sumat)`
- `matprint` is a simple to use command for visualising Stata matrices

AARHUS UNIVERSITY

sumat/matprint, example

```
sumat age lwt bwt, statistics(n missing unique mean ci) style(md) rowby(smoke) decimals((0,0,0,2))
```

		n	missing	unique	mean	ci95% lb	ci95% ub
smoke(No)	age of mother	115	0	23	23.43	22.43	24.43
	weight at last menstruation (kg)	115	0	59	130.90	125.71	136.10
	birth weight (grams)	115	0	87	3054.96	2917.44	3192.47
smoke(Yes)	age of mother	70	4	20	22.93	21.78	24.08
	weight at last menstruation (kg)	74	0	45	128.14	120.44	135.83
	birth weight (grams)	74	0	61	2772.30	2621.97	2922.63

AARHUS UNIVERSITY

crossmat - if it is worth showing, it is worth reusing

A wrapper for `tabulate` returning everything in Stata matrices, eg:

```
crossmat race bwlt2500
matprint 100 * r(pct), decimals(0) style("md")
```

		No	Yes	Total
race	white	39	13	51
	black	8	6	14
	other	21	14	35
	Total	68	32	100

```
return list

matrices:
             r(lrchi2) :  4 x 3
               r(chi2) :  4 x 3
               r(cpct) :  4 x 3
               r(rpct) :  4 x 3
             r(greeks) :  3 x 2
              r(tests) :  2 x 3
           r(expected) :  4 x 3
                r(pct) :  4 x 3
```

regmat - Regression matrix (Table 2) (**In next update**)

Tabulation of regression results

```
regmat, outcome(bwlt2500 bwlt1500) exposure(i.smoke age) adjustments("" "ftv i.race") drop(se) labels: ///
logit, vce(robust) or
```

		Adjustment 1				Adjustment 2			
		b	Lower 95% CI	Upper 95% CI	P value	b	Lower 95% CI	Upper 95% CI	P value
birthweight < 2500g	smoked during pregnancy (Yes)	2.17	1.15	4.09	0.02	4.21	1.94	9.13	0.00
	age of mother	0.95	0.90	1.01	0.08	0.97	0.91	1.03	0.31
birthweight < 1500g	smoked during pregnancy (Yes)	1.04	0.17	6.39	0.97	2.10	0.20	21.60	0.53
	age of mother	1.16	1.05	1.28	0.00	1.28	1.08	1.51	0.00

```
return list

macros:
       r(Adjustment_2) : "ftv i.race"
       r(Adjustment_1) : "Crude"

matrices:
             r(regmat) :  4 x 8
```

AARHUS UNIVERSITY

Final remarks and questions

- Diversity is good
- Many different tools!
  - Choose the ones that fits You the most!
- Questions?

References I

Barba, Lorena A. 2016. "The Hard Road to Reproducibility." *Science* 354 (6308). American Association for the Advancement of Science:142–42. https://doi.org/10.1126/science.354.6308.142.

Bruun, Niels Henrik. 2017. "Hacks for Stata Users." http://www.bruunisejs.dk/StataHacks/.

Chatfield, Mark. 2017. "Using and Interpreting Restricted Cubic Splines." 2017. https://www.statalist.org/forums/forum/general-stata-discussion/general/1420930-announcing-improved-table1_mc-stata-module-to-create-table-1-docx-of-baseline-characteristics-for-a-manuscript.

Christie, Tom. 2014. "MkDocs. Project Documentation with Markdown." http://www.mkdocs.org/.

Clayton, Phil. 2002. "TABLE1: Stata Module to Create Table 1 of Baseline Characteristics for a Manuscript." https://ideas.repec.org/c/boc/bocode/s457730.html.

Gallup, J. L. 2012a. "A New System for Formatting Estimation Tables." *Stata Journal* 12 (1). College Station, TX: Stata Press:3–28(26). http://www.stata-journal.com/article.html?article=sg97_4.

———. 2012b. "A Programmer's Command to Build Formatted Statistical Tables." *Stata Journal* 12 (4). College Station, TX: Stata Press:655–673(19). http://www.stata-journal.com/article.html?article=sg97_5.

Gruber, John. 2002. "Markdown." https://daringfireball.net/projects/markdown/.

AARHUS UNIVERSITY

References II

Haghish, E. F. 2016a. "Markdoc: Literate Programming in Stata." *Stata Journal* 16 (4). College Station, TX: Stata Press:964–988(25). http://www.stata-journal.com/article.html?article=pr0064.

———. 2016b. "Rethinking Literate Programming in Statistics." *Stata Journal* 16 (4). College Station, TX: Stata Press:938–963(26). http://www.stata-journal.com/article.html?article=pr0063.

Jann, B. 2007. "Making Regression Tables Simplified." *Stata Journal* 7 (2). College Station, TX: Stata Press:227–244(18). http://www.stata-journal.com/article.html?article=st0085_1.

———. 2014. "Plotting Regression Coefficients and Other Estimates." *Stata Journal* 14 (4). College Station, TX: Stata Press:708–737(30). http://www.stata-journal.com/article.html?article=gr0059.

———. 2016. "Creating Latex Documents from Within Stata Using Texdoc." *Stata Journal* 16 (2). College Station, TX: Stata Press:245–263(19). http://www.stata-journal.com/article.html?article=pr0062.

———. 2017. "Creating Html or Markdown Documents from Within Stata Using Webdoc." *Stata Journal* 17 (1). College Station, TX: Stata Press:3–38(36). http://www.stata-journal.com/article.html?article=pr0065.

MacFarlane, John. 2006. "Pandoc, a Universal Document Converter." http://pandoc.org/.

Magno, G. L. Lo. 2013. "Sar: Automatic Generation of Statistical Reports Using Stata and Microsoft Word for Windows." *Stata Journal* 13 (1). College Station, TX: Stata Press:39–64(26). http://www.stata-journal.com/article.html?article=pr0055.

References III

Patil, Prasad, Roger D. Peng, and Jeffrey Leek. 2016. "A Statistical Definition for Reproducibility and Replicability." *bioRxiv*. Cold Spring Harbor Laboratory. https://doi.org/10.1101/066803.

Peng, Roger D. 2009. "Reproducible Research and Biostatistics." *Biostatistics* 10 (3):405–8. https://doi.org/10.1093/biostatistics/kxp014.

Peng, Roger D., Francesca Dominici, and Scott L. Zeger. 2006. "Reproducible Epidemiologic Research." *American Journal of Epidemiology* 163 (9):783–89. https://doi.org/10.1093/aje/kwj093.

Rising, Bill. 2008. "Reproducible Research: Weaving with Stata." https://www.stata.com/meeting/italy08/rising_2008.pdf.

Rodríguez, G. 2017. "Literate Data Analysis with Stata and Markdown." *Stata Journal* 17 (3). College Station, TX: Stata Press:600–618(19). http://www.stata-journal.com/article.html?article=pr0067.

Shah, Attaullah. 2018. "Asdoc: An Easy Way of Creating Publication Quality Tables from Stata Commands." 2018. https://www.statalist.org/forums/forum/general-stata-discussion/general/1435798-asdoc-an-easy-way-of-creating-publication-quality-tables-from-stata-commands.

Watson, Ian. 2015. "Publication Quality Tables in Stata: A Tutorial for the Tabout Program." http://www.ianwatson.com.au/stata/tabout_tutorial.pdf.

AARHUS UNIVERSITY