THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact

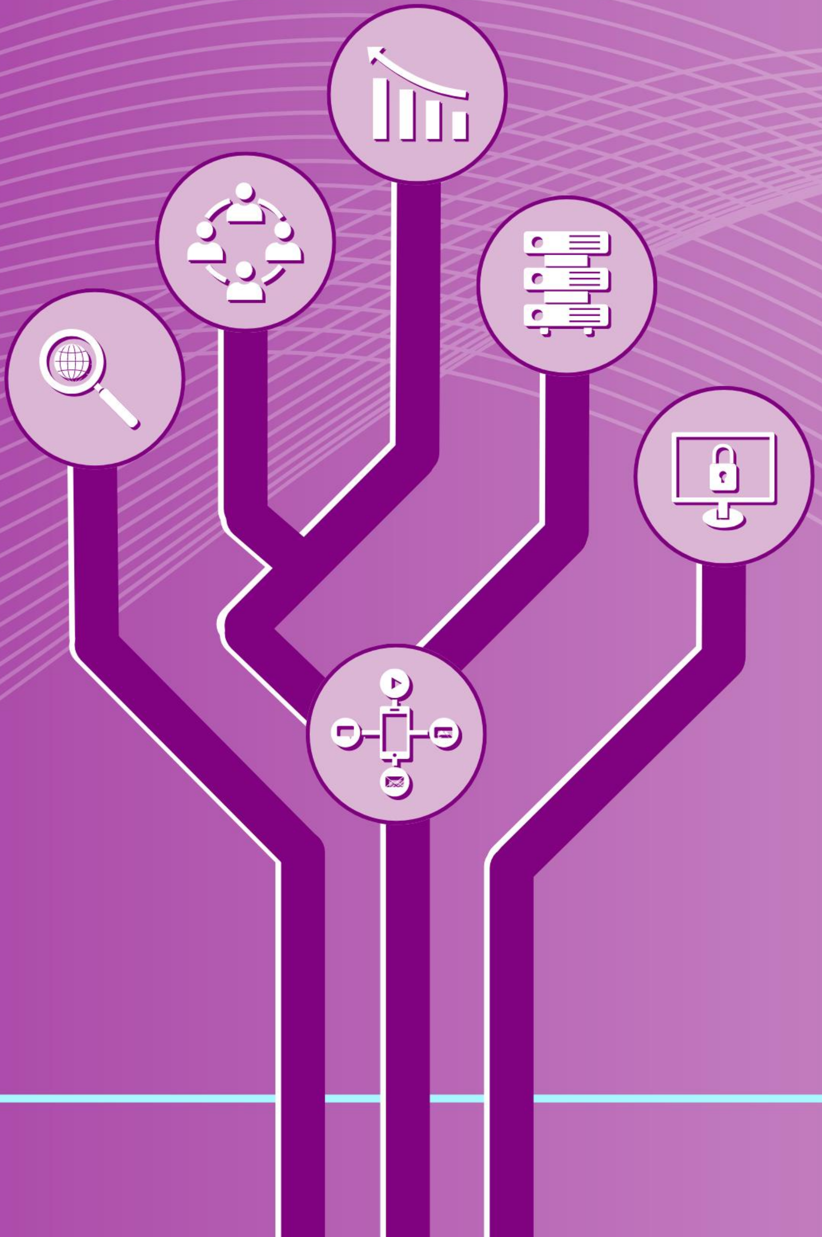
DIME
ANALYTICS

Ensuring Reproducibility in Stata

Insights from the World Bank's Reproducible Research Repository

Luis Eduardo San Martin

Junior Data Scientist

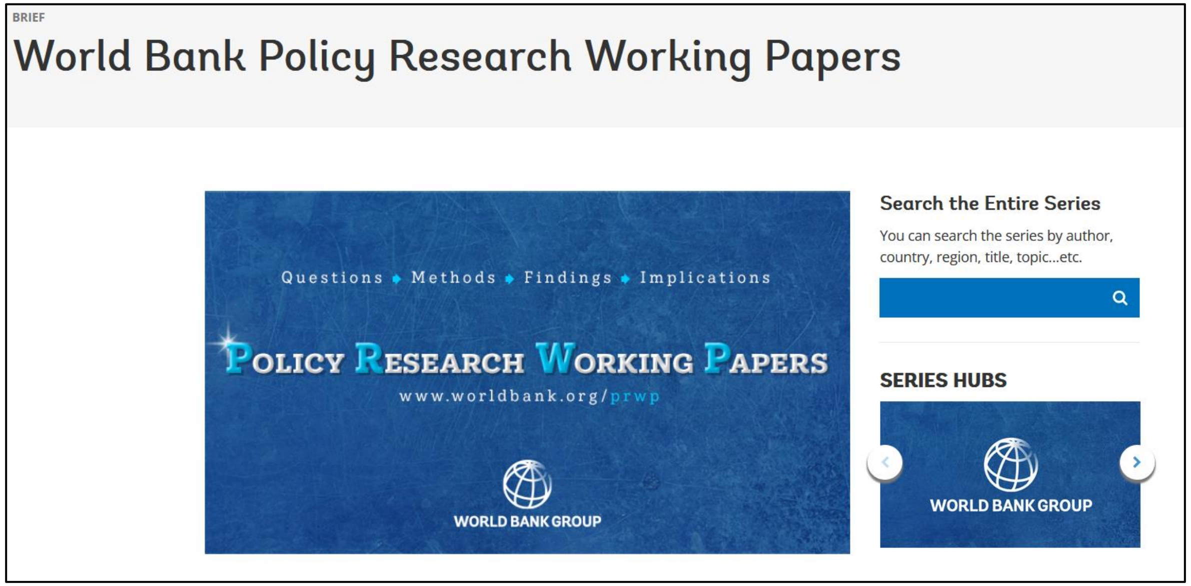


Introduction

About this session

About this session

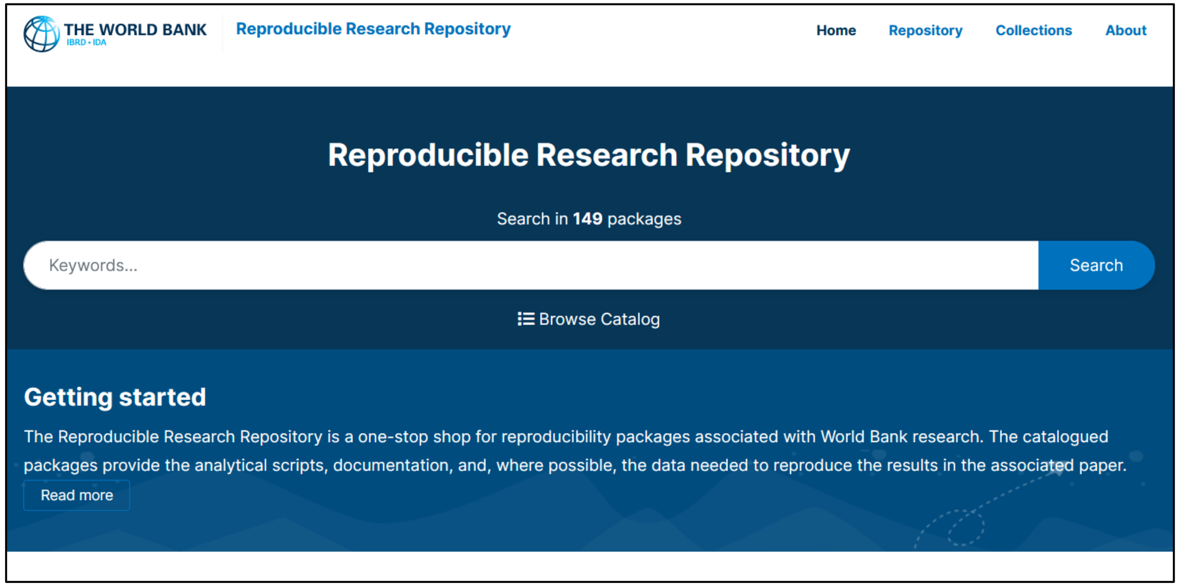
- This is based in our work reviewing and curating 140+ Word Bank reproducibility packages and working papers


BRIEF
Questions ◆ Methods ◆ Findings ◆ Implications
Policy Research Working Papers
www.worldbank.org/prwp
WORLD BANK GROUP
SERIES HUBS
WORLD BANK GROUP

World Bank Policy Research Working Papers

Search the Entire Series

You can search the series by author, country, region, title, topic...etc.



Reproducible Research Repository

Search in **149** packages

Getting started

The Reproducible Research Repository is a one-stop shop for reproducibility packages associated with World Bank research. The catalogued packages provide the analytical scripts, documentation, and, where possible, the data needed to reproduce the results in the associated paper.

https://www.worldbank.org/en/research/brief/world-bank-policy-research-working-papers

https://reproducibility.worldbank.org

About this session

- Insights for:
  - Researchers looking to submit or publish a reproducibility package for a paper
  - Stata coders looking to make their code easier for collaboration with colleagues or future self-collaboration
  - Advocates for transparency and openness in science
  - Stata users who have ever noticed their results change using the same code and data and have no idea why

About this session

- But wait, didn’t you guys cover this yesterday? Aren’t all reproducibility problems in Stata detectable with your package `reprun`?

About this session

- Well, no ☺
- Reprun confirms stability. Reproducibility issues can still happen across different computers for stable code

Stable

Produces the same outputs every run

Consistent

Tables and figures produced match exactly those from the authors

Reproducibility Verifications

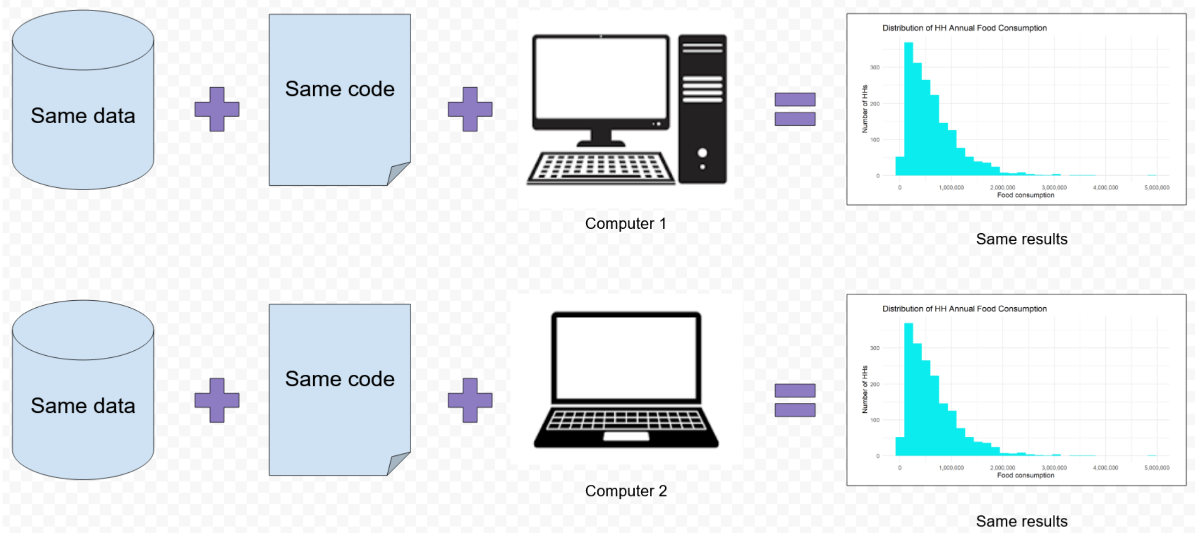
Reproducibility

"the ability [...] to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator"

– Bollen et al., Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science (2015)

Reproducibility

"the ability […] to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator"



In our team's work:

The ability to reproduce outputs using the same code and data inputs.

(computational reproducibility)

Reproducibility verifications

- Principle of **computational empathy** is strongly encouraged for teams sending us their works:
  - Don’t require users to do tedious things
  - Make code run as easy as possible
- 77% of works reviewed use Stata
- Only 20% are reproducible after the first try
- If not reproducible, our team reviews the code and collaborates with the authors to identify reproducibility issues

Reproducibility in Stata

Reproducibility in Stata

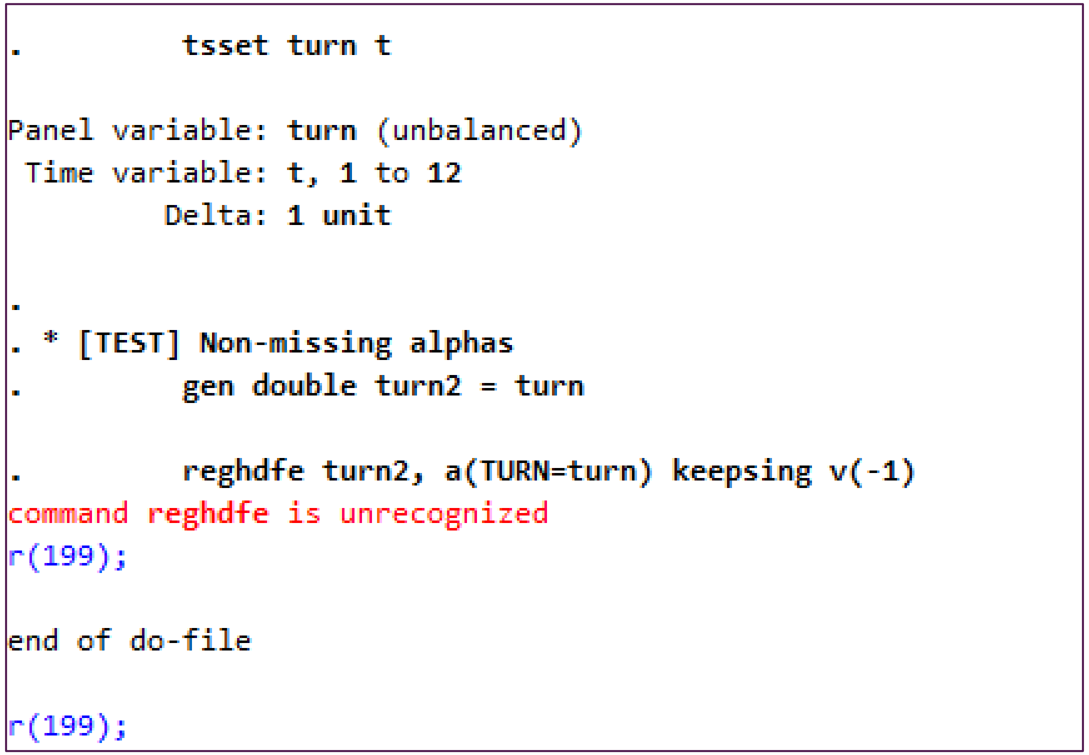
Four rules for reproducibility in our experience:

- Same dependencies
- Seed number
- Stata Version
- Unique sorting

But exceptions can still happen!

Same dependencies

- Missing dependencies will stop code execution



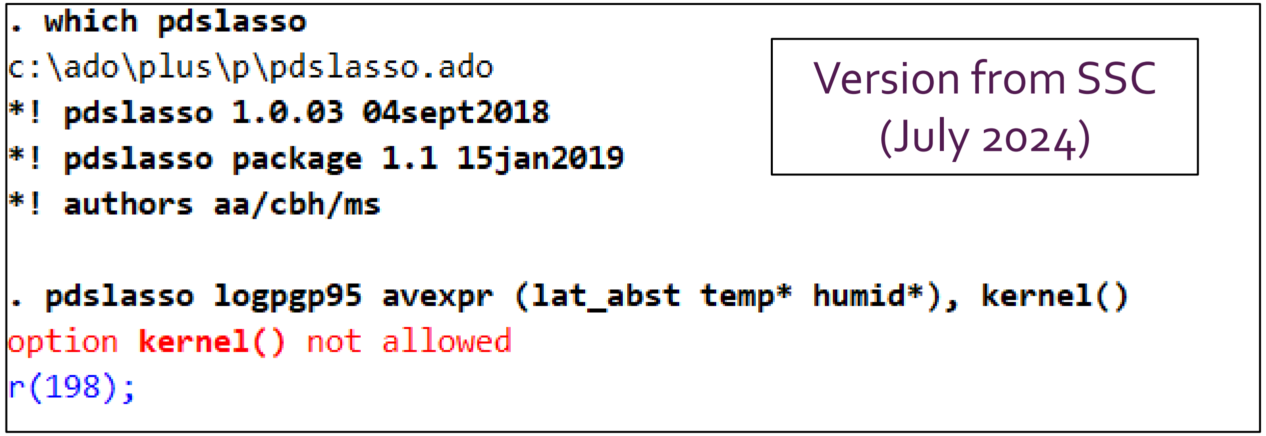
```
.         tsset turn t

Panel variable: turn (unbalanced)
 Time variable: t, 1 to 12
         Delta: 1 unit


.
. * [TEST] Non-missing alphas
.         gen double turn2 = turn

.         reghdfe turn2, a(TURN=turn) keepsing v(-1)
command reghdfe is unrecognized
r(199);

end of do-file

r(199);
```

Same dependencies

- Missing dependencies will stop code execution

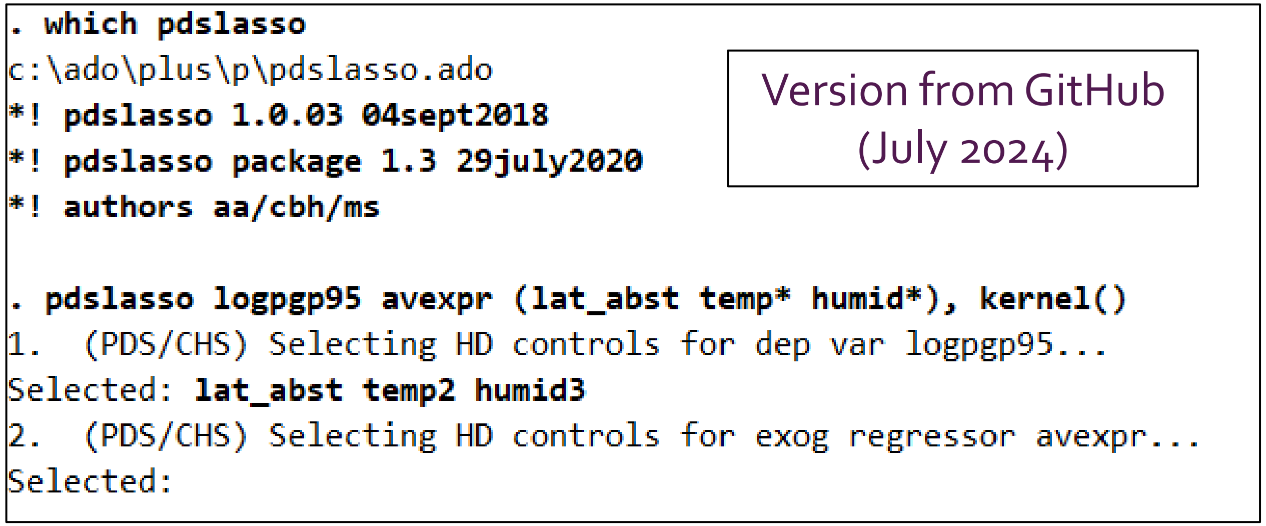
- Different versions of dependencies can use different options or produce different results

Recommended: *Save your environment: The (often) overlooked problem of research reproducibility in economics*



```
. which pdslasso
c:\ado\plus\p\pdslasso.ado
*! pdslasso 1.0.03 04sept2018
*! pdslasso package 1.1 15jan2019
*! authors aa/cbh/ms

. pdslasso logpgp95 avexpr (lat_abst temp* humid*), kernel()
option kernel() not allowed
r(198);
```

Version from SSC (July 2024)



```
. which pdslasso
c:\ado\plus\p\pdslasso.ado
*! pdslasso 1.0.03 04sept2018
*! pdslasso package 1.3 29july2020
*! authors aa/cbh/ms

. pdslasso logpgp95 avexpr (lat_abst temp* humid*), kernel()
1.  (PDS/CHS) Selecting HD controls for dep var logpgp95...
Selected: lat_abst temp2 humid3
2.  (PDS/CHS) Selecting HD controls for exog regressor avexpr...
Selected:
```

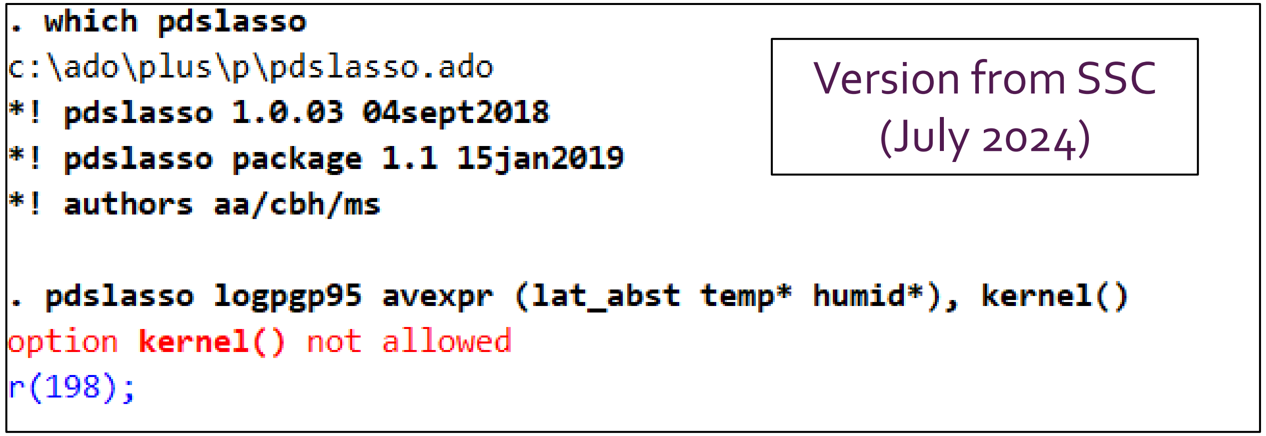
Version from GitHub (July 2024)

Same dependencies

- Missing dependencies will stop code execution

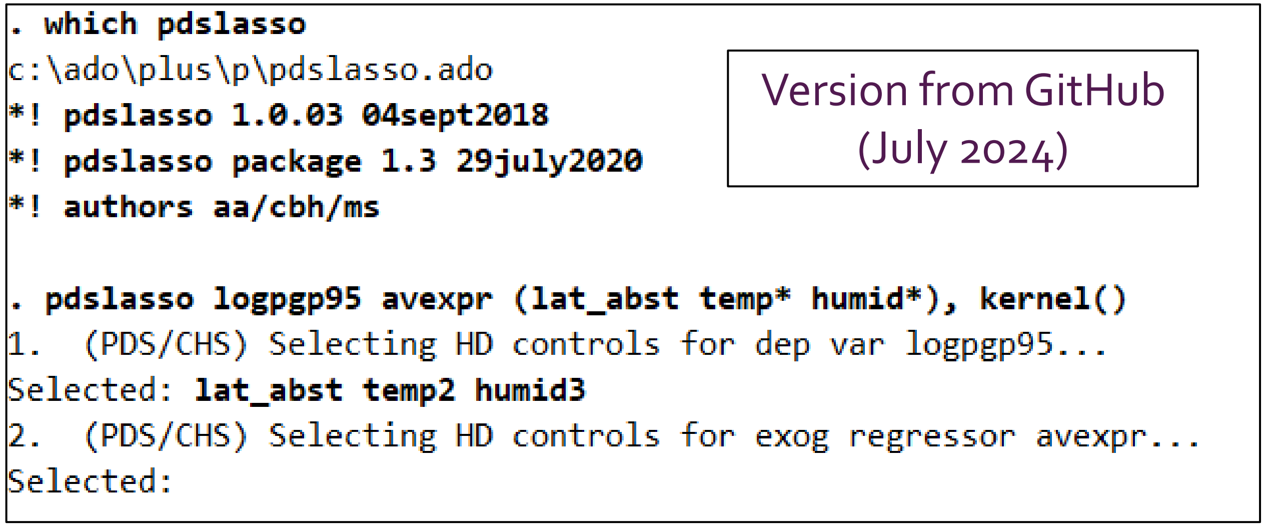
- Different versions of dependencies can use different options or produce different results

- Solution:
  - use an ado folder for each paper or project
  - Use repado



```
. which pdslasso
c:\ado\plus\p\pdslasso.ado
*! pdslasso 1.0.03 04sept2018
*! pdslasso package 1.1 15jan2019
*! authors aa/cbh/ms

. pdslasso logpgp95 avexpr (lat_abst temp* humid*), kernel()
option kernel() not allowed
r(198);
```

Version from SSC (July 2024)



```
. which pdslasso
c:\ado\plus\p\pdslasso.ado
*! pdslasso 1.0.03 04sept2018
*! pdslasso package 1.3 29july2020
*! authors aa/cbh/ms

. pdslasso logpgp95 avexpr (lat_abst temp* humid*), kernel()
1.  (PDS/CHS) Selecting HD controls for dep var logpgp95...
Selected: lat_abst temp2 humid3
2.  (PDS/CHS) Selecting HD controls for exog regressor avexpr...
Selected:
```

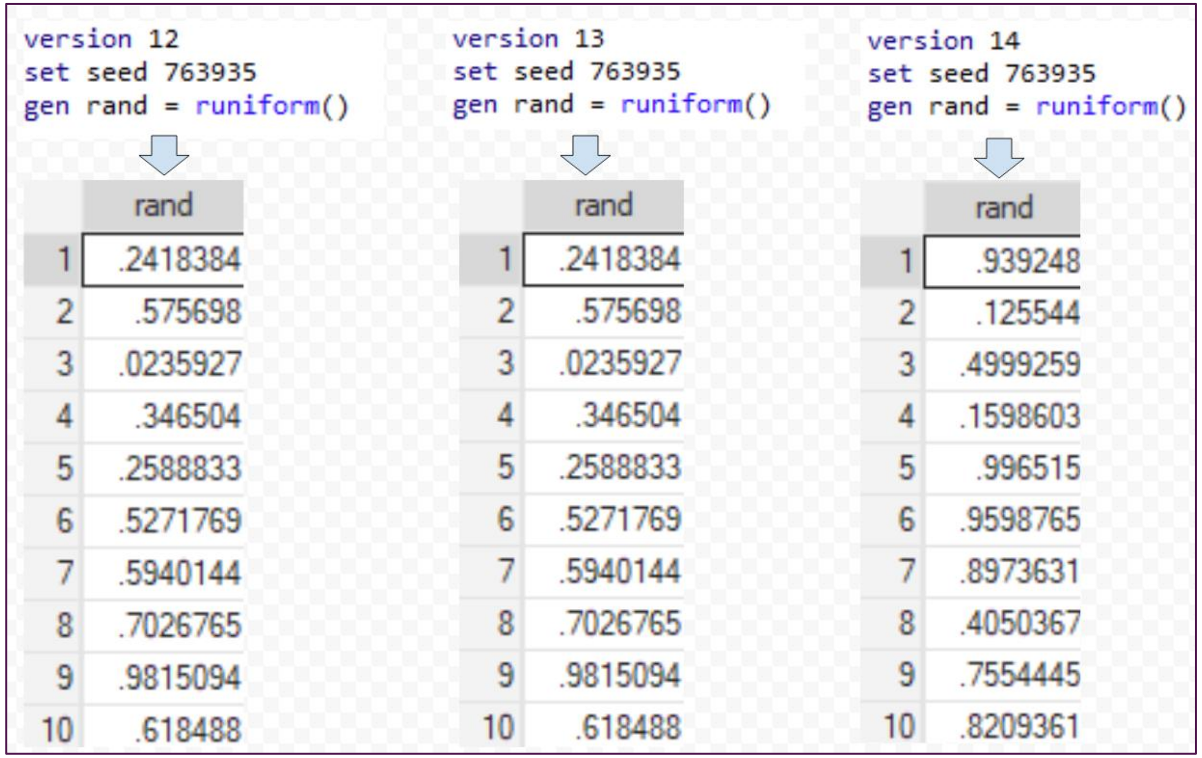
Version from GitHub (July 2024)

Random Seed Number and Versioning

- The generation of random numbers occasionally changes between versions

- Setting the Stata version using `version` ensures that the same random numbers are generated

- Importantly, **this will ensure your code remains reproducible** even if the generation of random numbers changes in a future version of Stata



```
version 12
set seed 763935
gen rand = runiform()
```

	rand
1	.2418384
2	.575698
3	.0235927
4	.346504
5	.2588833
6	.5271769
7	.5940144
8	.7026765
9	.9815094
10	.618488

```
version 13
set seed 763935
gen rand = runiform()
```

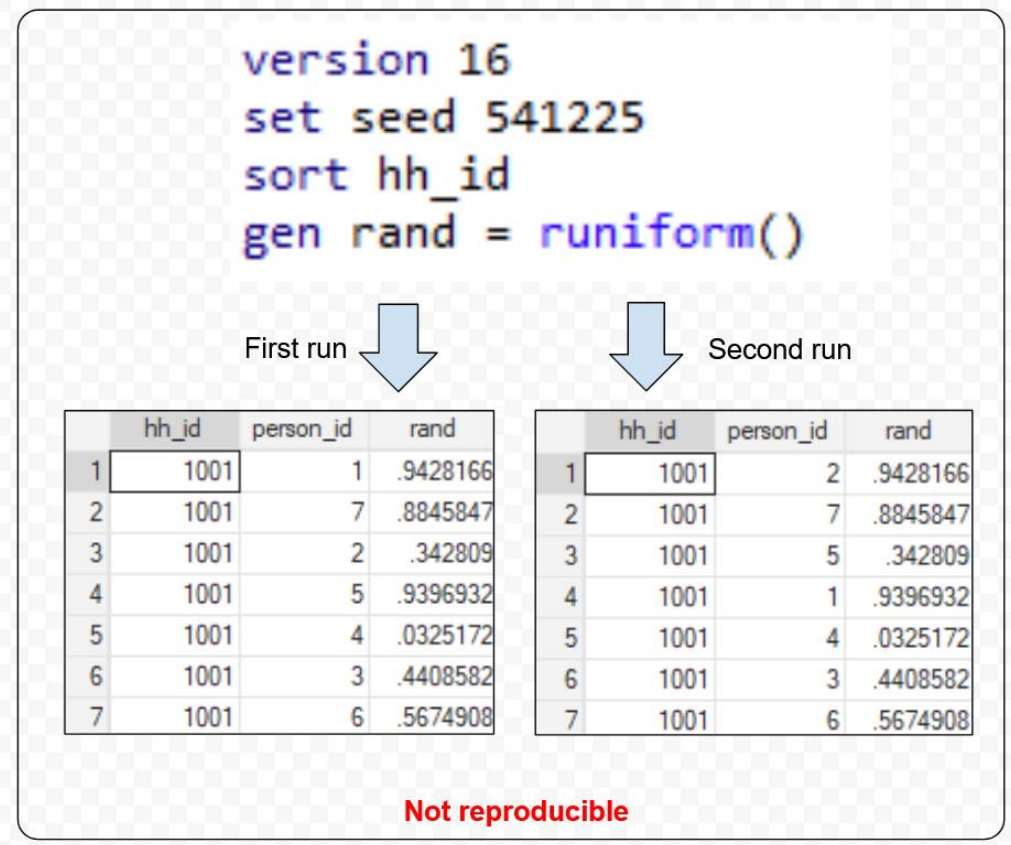
	rand
1	.2418384
2	.575698
3	.0235927
4	.346504
5	.2588833
6	.5271769
7	.5940144
8	.7026765
9	.9815094
10	.618488

```
version 14
set seed 763935
gen rand = runiform()
```

	rand
1	.939248
2	.125544
3	.4999259
4	.1598603
5	.996515
6	.9598765
7	.8973631
8	.4050367
9	.7554445
10	.8209361

Non-unique sorts

- By default, Stata handles ties in sorts randomly
- If not handled correctly, **this can be a problem for reproducibility**



```
version 16
set seed 541225
sort hh_id
gen rand = runiform()
```

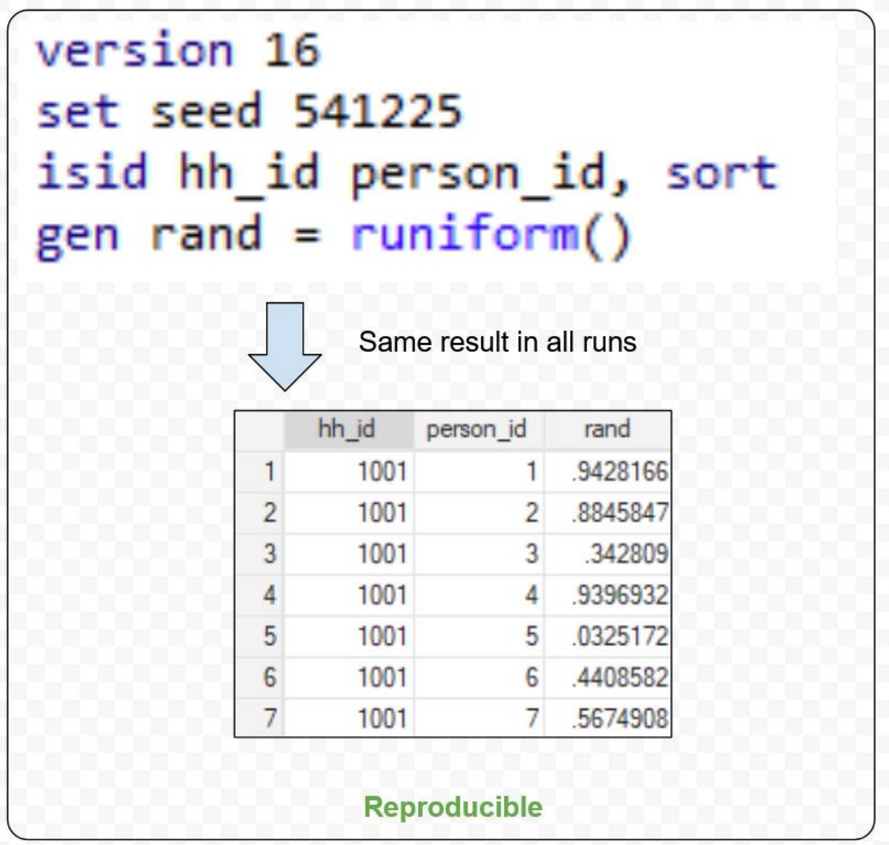
First run

	hh_id	person_id	rand
1	1001	1	.9428166
2	1001	7	.8845847
3	1001	2	.342809
4	1001	5	.9396932
5	1001	4	.0325172
6	1001	3	.4408582
7	1001	6	.5674908

Second run

	hh_id	person_id	rand
1	1001	2	.9428166
2	1001	7	.8845847
3	1001	5	.342809
4	1001	1	.9396932
5	1001	4	.0325172
6	1001	3	.4408582
7	1001	6	.5674908

Not reproducible



```
version 16
set seed 541225
isid hh_id person_id, sort
gen rand = runiform()
```

Same result in all runs

	hh_id	person_id	rand
1	1001	1	.9428166
2	1001	2	.8845847
3	1001	3	.342809
4	1001	4	.9396932
5	1001	5	.0325172
6	1001	6	.4408582
7	1001	7	.5674908

Reproducible

Non-unique sorts

- How to handle this?

1. Avoid using `sort` but use instead `isid [varlist], sort`

2. Be aware of "implicit sorts" Stata applies when using other commands

Some implicit sorts

Command	Issue	Solution
`merge 1:m …`	Stata will sort observations by the key variable, but randomly within it	Add a unique sort after the merge

`bysort [varlist]: gen …`
`bysort [varlist]: egen …`
	If `[varlist]` doesn't produce a unique sorting, results of `gen …` might not be reproducible if they depend on the observations' positions. For example:
	`bysort hh_id: gen sample = 1 if _n == 1`
		- Do not sort with `bysort`
		- Sort first and then use `by` separately:
		`isid [var1 var2 …], sort`
		`by var1: gen …`

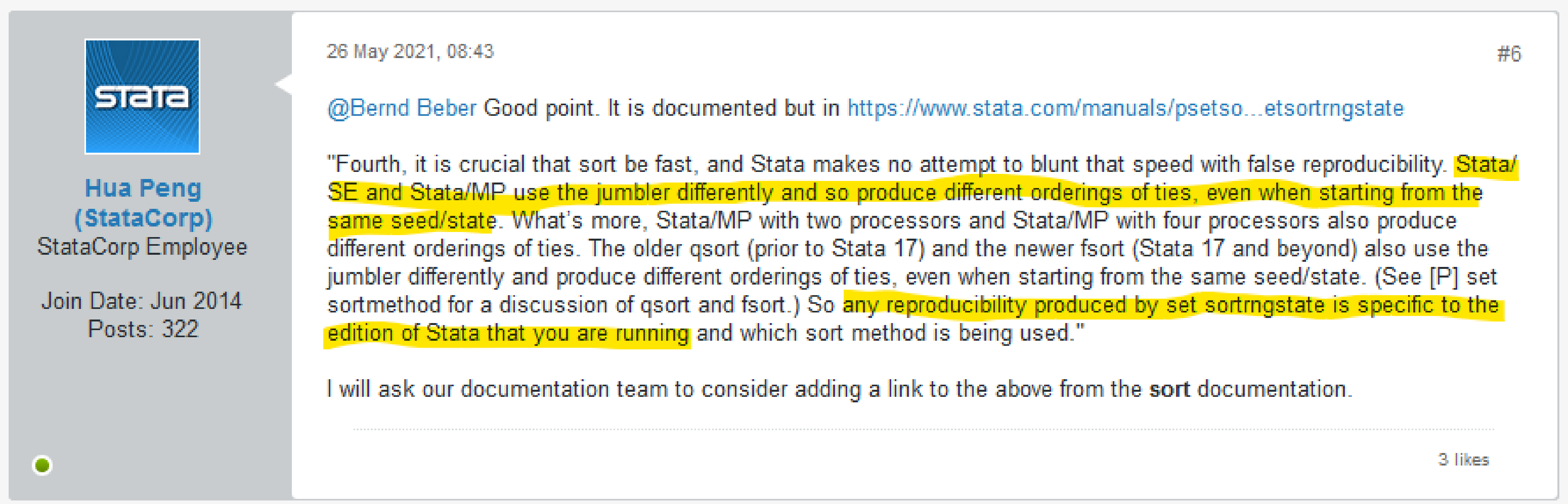
`duplicates drop [varlist], force`	Stata will select randomly which observations to drop after sorting by [varlist]. If it doesn't produce a unique sorting, it might not be reproducible	Think of why you have duplicates in `[varlist]` and choose a criteria for dropping them based on your assessment. Some examples: - Which obs was collected first - Which obs has less missings across all variables

Important: Avoid using `set sortseed`

- Experienced Stata programmers might have heard of `set sortseed`
- It allows to set a seed number for to set the random state for ties in sorts
- This is similar to how `set seed` sets a random state for random numbers generation
- However, we recommend never using it, as it is only a partial solution: `set sortseed` **only gives reproducible results within the same Stata edition (SE, MP).**

Important: Avoid using `set sortseed`

- See discussion in Setting version, seed, and sortseed not sufficient for reproducibility?



Hua Peng (StataCorp)
StataCorp Employee

Join Date: Jun 2014
Posts: 322

26 May 2021, 08:43

#6

@Bernd Beber Good point. It is documented but in https://www.stata.com/manuals/psetso...etsortrngstate

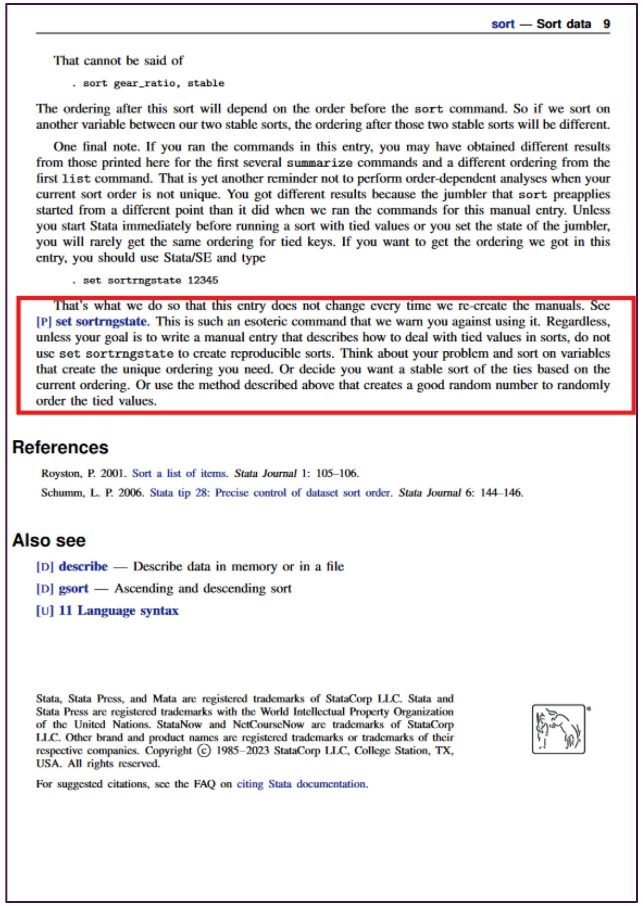
"Fourth, it is crucial that sort be fast, and Stata makes no attempt to blunt that speed with false reproducibility. Stata/SE and Stata/MP use the jumbler differently and so produce different orderings of ties, even when starting from the same seed/state. What's more, Stata/MP with two processors and Stata/MP with four processors also produce different orderings of ties. The older qsort (prior to Stata 17) and the newer fsort (Stata 17 and beyond) also use the jumbler differently and produce different orderings of ties, even when starting from the same seed/state. (See [P] set sortmethod for a discussion of qsort and fsort.) So any reproducibility produced by set sortrngstate is specific to the edition of Stata that you are running and which sort method is being used."

I will ask our documentation team to consider adding a link to the above from the **sort** documentation.

3 likes

Important: Avoid using `set sortseed`

- Also check Sorting with Ties in Stata's longer documentation for `sort`



sort — Sort data 9

That cannot be said of

```
. sort gear_ratio, stable
```

The ordering after this sort will depend on the order before the `sort` command. So if we sort on another variable between our two stable sorts, the ordering after those two stable sorts will be different.

One final note. If you ran the commands in this entry, you may have obtained different results from those printed here for the first several `summarize` commands and a different ordering from the first `list` command. That is yet another reminder not to perform order-dependent analyses when your current sort order is not unique. You got different results because the jumbler that `sort` preapplies started from a different point than it did when we ran the commands for this manual entry. Unless you start Stata immediately before running a sort with tied values or you set the state of the jumbler, you will rarely get the same ordering for tied keys. If you want to get the ordering we got in this entry, you should use Stata/SE and type

```
. set sortrngstate 12345
```

That's what we do so that this entry does not change every time we re-create the manuals. See [P] **set sortrngstate**. This is such an esoteric command that we warn you against using it. Regardless, unless your goal is to write a manual entry that describes how to deal with tied values in sorts, do not use `set sortrngstate` to create reproducible sorts. Think about your problem and sort on variables that create the unique ordering you need. Or decide you want a stable sort of the ties based on the current ordering. Or use the method described above that creates a good random number to randomly order the tied values.

References

Royston, P. 2001. Sort a list of items. *Stata Journal* 1: 105–106.

Schumm, L. P. 2006. Stata tip 28: Precise control of dataset sort order. *Stata Journal* 6: 144–146.

Also see

[D] **describe** — Describe data in memory or in a file

[D] **gsort** — Ascending and descending sort

[U] **11 Language syntax**

For suggested citations, see the FAQ on citing Stata documentation.

"[`set sortrngstate`] is such an esoteric command that ***we warn you against using it****. Regardless, unless your goal is to write a manual entry that describes how to deal with tied values in sorts, do not use `set sortrngstate` to create reproducible sorts.* ***Think about your problem and sort on variables that create the unique ordering you need.****"*

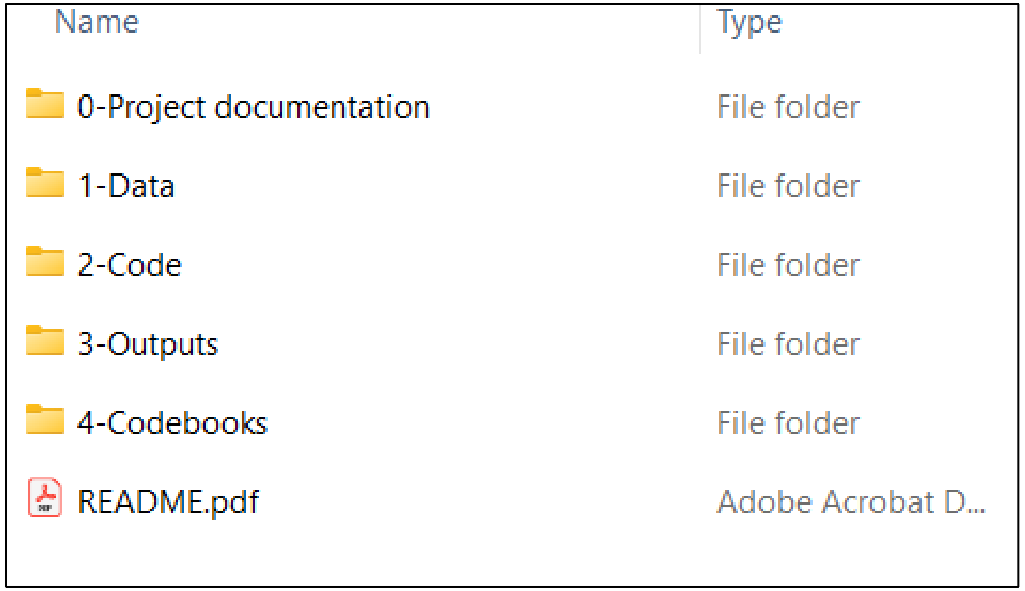
(emphasis added for this presentation)

Beyond Code Execution

Make it easy for replicators!

- Remember computational empathy?

- This means:
  - **Self-contained reproducibility package**
  - Main do-file with all settings
  - One-button reproducibility
  - Clear code
  - Documentation

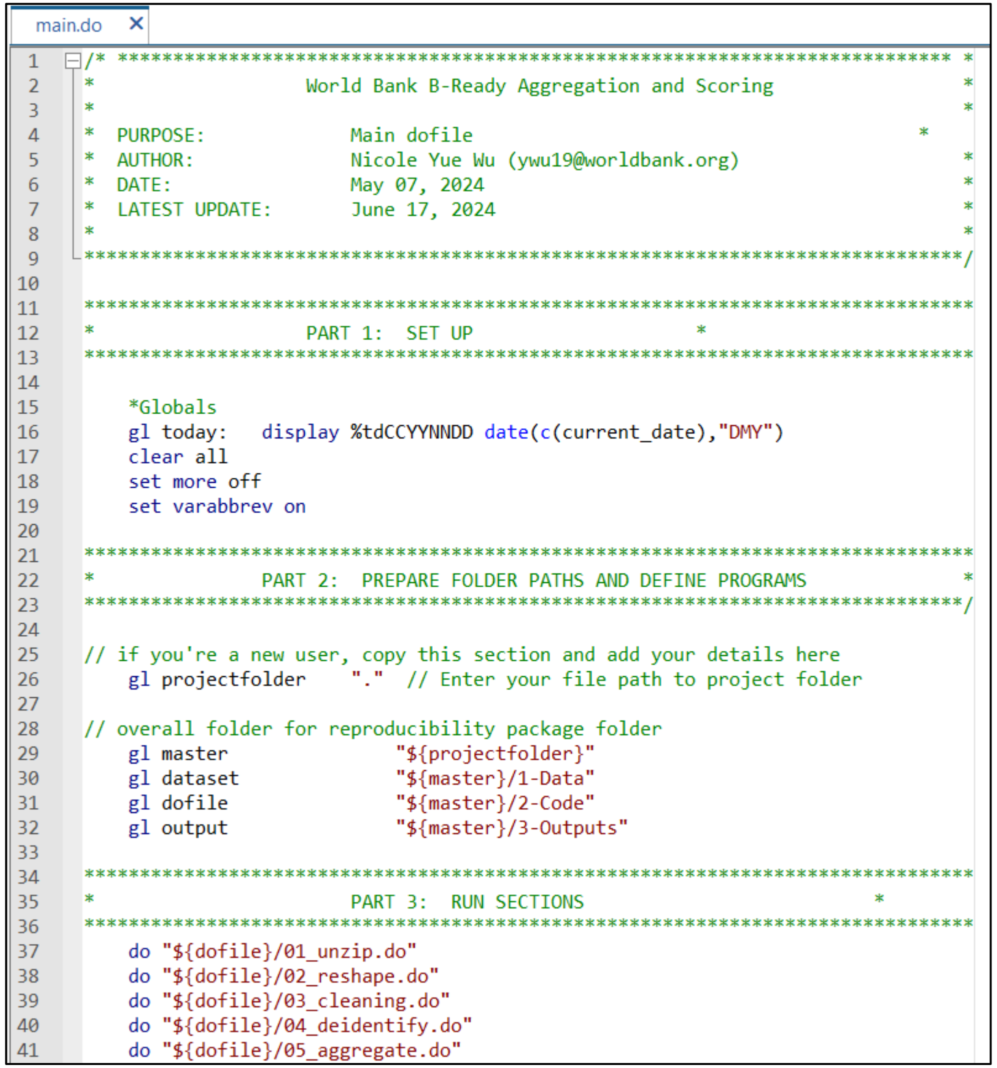


Name	Type
0-Project documentation	File folder
1-Data	File folder
2-Code	File folder
3-Outputs	File folder
4-Codebooks	File folder
README.pdf	Adobe Acrobat D...

Make it easy for replicators!

- Remember computational empathy?

- This means:
  - Self-contained reproducibility package
  - **Main do-file with all settings**
  - **One-button reproducibility**
  - Clear code
  - Documentation



```
/* *************************************************************************** *
*                   World Bank B-Ready Aggregation and Scoring                 *
*                                                                              *
*  PURPOSE:             Main dofile                                         *
*  AUTHOR:              Nicole Yue Wu (ywu19@worldbank.org)                    *
*  DATE:                May 07, 2024                                           *
*  LATEST UPDATE:       June 17, 2024                                          *
*                                                                              *
*******************************************************************************/

********************************************************************************
*                   PART 1:  SET UP                                 *
********************************************************************************


    *Globals
    gl today:   display %tdCCYYNNDD date(c(current_date),"DMY")
    clear all
    set more off
    set varabbrev on

********************************************************************************
*               PART 2:  PREPARE FOLDER PATHS AND DEFINE PROGRAMS              *
*******************************************************************************/

// if you're a new user, copy this section and add your details here
    gl projectfolder    "."  // Enter your file path to project folder

// overall folder for reproducibility package folder
    gl master               "${projectfolder}"
    gl dataset              "${master}/1-Data"
    gl dofile               "${master}/2-Code"
    gl output               "${master}/3-Outputs"

********************************************************************************
*                       PART 3:  RUN SECTIONS                           *
********************************************************************************
    do "${dofile}/01_unzip.do"
    do "${dofile}/02_reshape.do"
    do "${dofile}/03_cleaning.do"
    do "${dofile}/04_deidentify.do"
    do "${dofile}/05_aggregate.do"
```

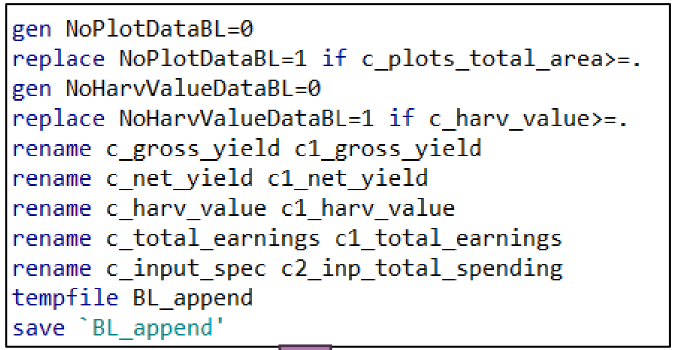
Code readability

“Programs must be written for people to read, and only incidentally for machines to execute.”

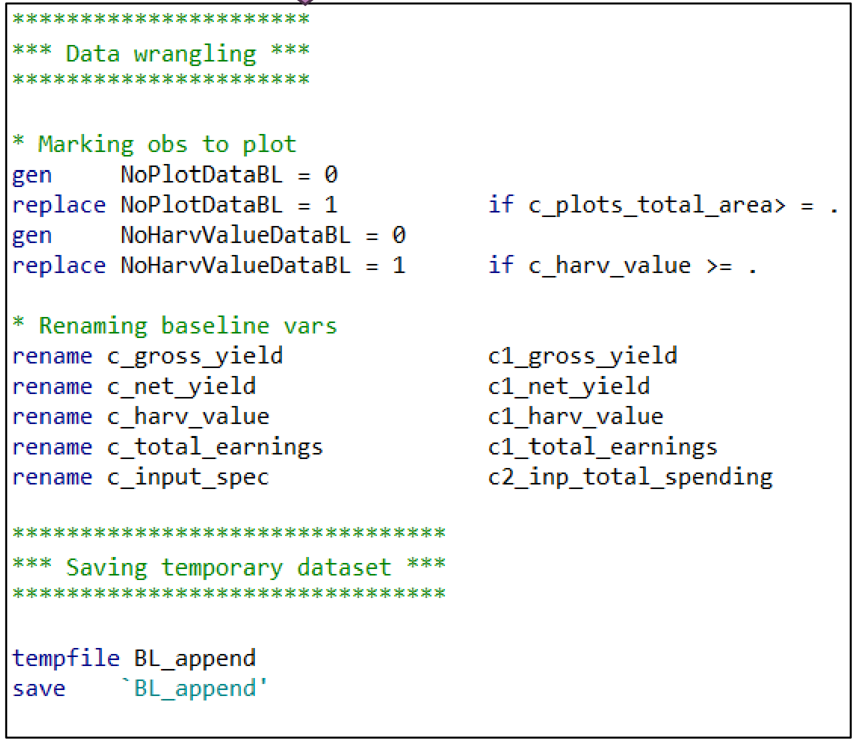
—Abelson, Susman and Susman, Structure and Interpretation of Computer Programs (1985)

Code readability

- Code linked to a paper should allow readers to understand the paper's logic, assumptions, and check its correctness
- You can do a lot with simple measures:
  - Horizontal and vertical spaces
  - Code comments
  - Section headers
- Use the Stata linter to improve your code



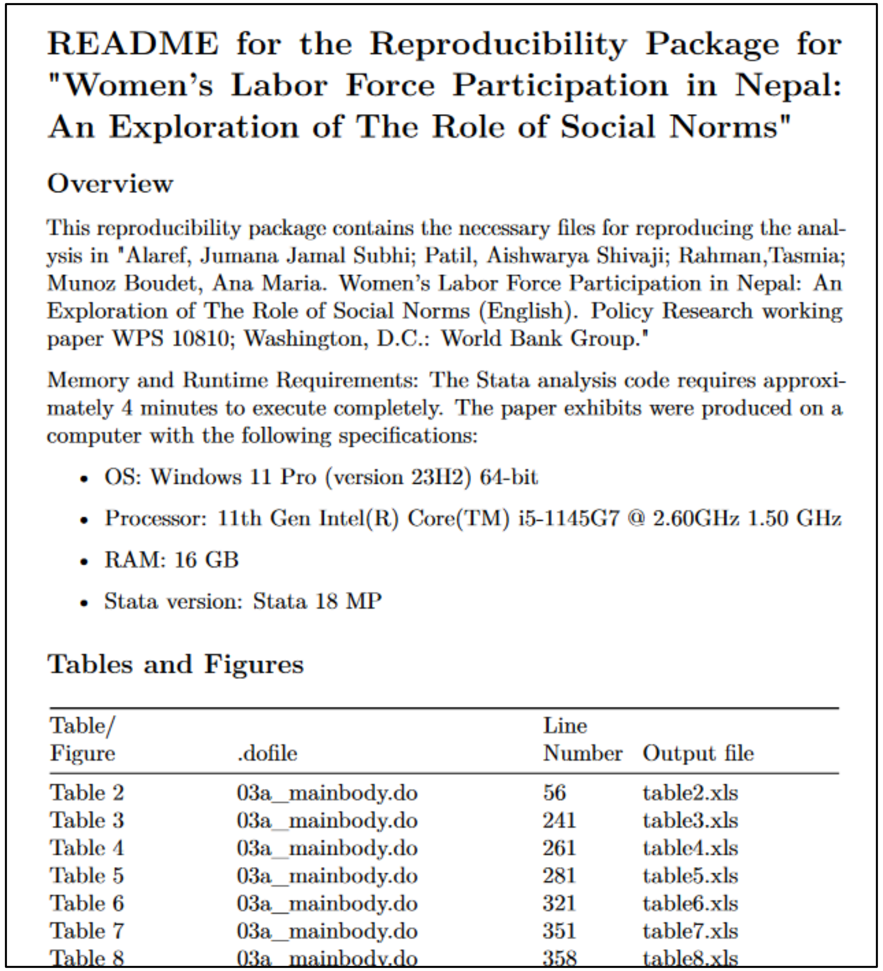
```
gen NoPlotDataBL=0
replace NoPlotDataBL=1 if c_plots_total_area>=.
gen NoHarvValueDataBL=0
replace NoHarvValueDataBL=1 if c_harv_value>=.
rename c_gross_yield c1_gross_yield
rename c_net_yield c1_net_yield
rename c_harv_value c1_harv_value
rename c_total_earnings c1_total_earnings
rename c_input_spec c2_inp_total_spending
tempfile BL_append
save `BL_append'
```



```
**********************
*** Data wrangling ***
**********************

* Marking obs to plot
gen     NoPlotDataBL = 0
replace NoPlotDataBL = 1           if c_plots_total_area> = .
gen     NoHarvValueDataBL = 0
replace NoHarvValueDataBL = 1      if c_harv_value >= .

* Renaming baseline vars
rename c_gross_yield               c1_gross_yield
rename c_net_yield                 c1_net_yield
rename c_harv_value                c1_harv_value
rename c_total_earnings            c1_total_earnings
rename c_input_spec                c2_inp_total_spending

********************************
*** Saving temporary dataset ***
********************************


tempfile BL_append
save     `BL_append'
```

Documentation

- Include a README file with the following:
  - Data provenance information
  - Code outputs - paper exhibits linkage (exm: scatterplot.png → Figure 3 in the paper)
  - System information for the generation of exhibits of the paper (OS, processor, RAM, Stata version and edition)



README for the Reproducibility Package for "Women's Labor Force Participation in Nepal: An Exploration of The Role of Social Norms"

Overview

This reproducibility package contains the necessary files for reproducing the analysis in 'Alaref, Jumana Jamal Subhi; Patil, Aishwarya Shivaji; Rahman,Tasmia; Munoz Boudet, Ana Maria. Women's Labor Force Participation in Nepal: An Exploration of The Role of Social Norms (English). Policy Research working paper WPS 10810; Washington, D.C.: World Bank Group.'

Memory and Runtime Requirements: The Stata analysis code requires approximately 4 minutes to execute completely. The paper exhibits were produced on a computer with the following specifications:

- OS: Windows 11 Pro (version 23H2) 64-bit
- Processor: 11th Gen Intel(R) Core(TM) i5-1145G7 @ 2.60GHz 1.50 GHz
- RAM: 16 GB
- Stata version: Stata 18 MP

Tables and Figures

Table/ Figure	.dofile	Line Number	Output file
Table 2	03a_mainbody.do	56	table2.xls
Table 3	03a_mainbody.do	241	table3.xls
Table 4	03a_mainbody.do	261	table4.xls
Table 5	03a_mainbody.do	281	table5.xls
Table 6	03a_mainbody.do	321	table6.xls
Table 7	03a_mainbody.do	351	table7.xls
Table 8	03a_mainbody.do	358	table8.xls

Thank you!

lsanmartin@worldbank.org

DIME
ANALYTICS

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP
Development Economics • Impact