Stata Tutorial for DEAL

Ricardo E. Miranda [1] and Xiaoqi (Jade) Peng[1]

Duke University[1]

DEAL, 2022 Fall
Aug 29-30, 2022

Outline

1. Replication
2. Research Folder Structure
3. Version Control
4. Data Analysis Workflow
   - Do File - quick overview
   - Data Cleaning
   - Data Cleaning

Replication

Why we need replication?

- Transparency
- Improves communication
- Helps other researchers (within and outside the research team/project)

Exact documentation of all the resources and steps necessary to produce an identical research result.

Ideally it should be easy, not just feasible.

Research Folder Structure

An Example

- Raw data: **raw data should never be touched**
- Working data: active manipulation of data
- Tables: collection of useful data points derived from data analysis
- Figures: graph representation to demonstrate critical results

Benefits of following this structure:

1. Error Tracing - localize errors
2. Clarity - all collaborators know where to look for graphs, tables, or data
3. Workflow - data processing typically follow the procedure of data analysis to table to figures

Version Control

Introduction

Importance of version control

- Smooth collaboration with team members
- Tracking changes; reversing to older versions if necessary

An easy tool: Github Desktop (free for download)

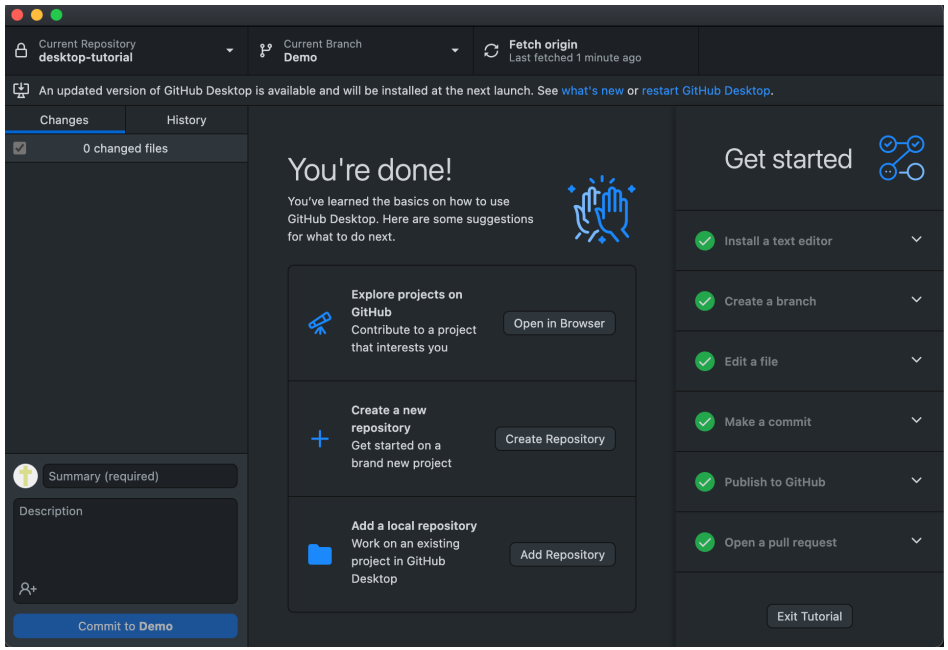

Figure: Screenshot for starting page

Data Analysis Workflow

Do File - Brief Summary

Logistics

- Author, date, description
- Temporary and auxiliary files

Global variables and working directories
Global Variables, once defined, are available anywhere in Stata.
In contrast, **Local variables** exist solely within the program or do-file in which they are defined. When a program or do-file ends, its local variables are permanently deleted.

This property makes global variables particularly suitable for defining working directories when the same set of files are shared between collaborators. See next slide for an example.

global variable - an example

```
*Initial commands
clear
set more off
set maxvar 5000
pause off

*Set working directory for a folder that will be shared
global User "Jade"
if "$User"=="Ricardo"{
    *Approach 1:
    cd "D:\DEAL\Lecture1"
    *Approach 2
    global Raw "D:\DEAL\Lecture1\Raw"
    global Working "D:\DEAL\Lecture1\Working"
    global Tables "D:\DEAL\Lecture1\Tables"
    global Figures "D:\DEAL\Lecture1\Figures"
    global Dos "D:\DEAL\Lecture1\Dos"

    do "$Dos\01_CleanConcentradoHogares.do"
    do "$Dos\02_CleanLivingPlaceCharacteristics.do"
    do "$Dos\03_VariableConstruction.do"
}

if "$User"=="Jade"{
    cd "/Users/jadepeng/GoogleDrive/22Fall/DEAL_Lec_1"

    do "Dos/01_CleanConcentradoHogares.do"
    do "Dos/02_CleanLivingPlaceCharacteristics.do"
    do "Dos/03_VariableConstruction.do"
}
```

Figure: Master do file specifies different paths for Ricardo versus Jade

Data Cleaning - Overview

Run stata_example.do for a quick walk-through on the pre-installed auto.dta.

- Reading datasets
  - cd desired_location
  - use ... , clear
  - import excel/delimited; use filename.extension/file_link
- the help command
- Viewing and Cleaning
  To view: describe, summarize, list, tabulate, browse, bysort
  To modify: label, drop, destring/tostring, generate, replace
  Important distinction: **description (return) vs estimation (ereturn)**
- Dealing with missing data (imputation)
  drop altogether; assume a value; assume similarity with nearby entries or adopt other extrapolation strategies
- Saving data

Data Cleaning - Multiple datasets

Look at a real-world example - Mexican Household data

- convert data into desired formats
- take care of missing/duplicate values
- relabel to make data more readable
- split or combine datasets when necessary

Different Ways to Combine Datasets

- append (vertically); merge (horizontally)
- joinby using data2: form pairwise combinations of observations from data1.dta in memory with those from data2.dta using all common variables and drop unmatched observations

Thank you!

E-mail contact: jade.peng@duke.edu