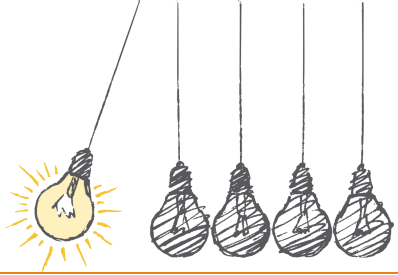
Programming 101



DIME Analytics RA Onboarding

February 4, 2020

Development Impact Evaluation (DIME)
The World Bank

WORLD BANK GROUP

i2i DIME
TRANSFORM DEVELOPMENT

Introduction

Can I use Excel?

The main reason why we code

- In Excel you make **changes directly to the data** and save **new versions of the data set**

- In Stata (or R) you make **changes to the instructions** on how to get from the raw data to the final analysis and save **new versions of the instructions**

Your code is an output

Researchers often treat code as a **mean to an end**

The main point of this presentation is that your **code is** equally as much **an end in itself** as the paper or the report you are writing!

Objective

- To become a great coder you both need to code like a coder and think like a coder.

- Both takes a lot of practice to master, but the objective of this session is to give you a framework to think like a coder by answering questions like:
  - Why do we code?
  - How do I code so it is the most helpful for other people in my team?

Objective

- In academia:
  - Being correct is what matters

Objective

- In academia:
  - Being correct is what matters
- In the workplace:
  - Correct is equally important as in academia
  - Past, current and future team members will contribute to the same code, and therefore we need to standardize how we code, and focus on skills for coding as a team

Critical thinking about data

- Do I believe this number?
- What can go wrong in my code?
- How will missing values be treated in this command?
- What would happen if more observations would be added to the data set?
- What would happen if some observations would be removed from the data set?

Data Management

Explore a raw data set

- What is the first thing you want to look for every single time you open a new data set for the first time?

Explore a raw data set

- What is the first thing you want to look for every single time you open a new data set for the first time?

    1. Unit of observation
    2. Uniquely and fully identifying ID variable

Explore a raw data set

Household_data.csv

HHID	Village	District	HH Number	HH Head	HHH Age
22501	25	2	1	Andrew	52
22502	25	2	2	Patrick	48
23207	32	2	7	Charles	29
23205	32	2	5	Jeffrey	37
12501	25	1	1	Walter	48
11103	11	1	3	Anne	26
11205	12	1	5	Lawrence	61
24502	45	2	2	Dennis	45
24501	45	2	1	Nancy	41

Explore a raw data set

Clinic_data.csv

Clinic ID	Clinic Number	District	Patient	Age
2452	542	2	Andrew	52
2543	543	2	Patrick	48
2156	156	2	Charles	29
1152	152	1	Jeffrey	37
1152	152	1	Walter	49
1238	238	1	Anne	26
1122	122	1	Lawrence	61
2122	122	2	Dennis	45
2122	122	2	Nancy	41

ID variables

- Only work with data set that has an ID variable. If the data set that you have received does not have one, then creating it is your first task
- Test that the ID variable is uniquely and fully identifying
- Use only one variable as ID variable

Role division in data work

- Research Assistants:
  - No one will look at the data as much as the RA
  - Irregularities in the data that the RA does not identify will often never be discovered

- Economists:
  - In charge of deciding which irregularities will be corrected and how
  - Economists completely depend on RAs to identify irregularities and get the information to make the best call

Coding styles

Is this slide easy to read?

White Space. Stata does not distinguish between one empty space and many empty spaces, or one line break or many line breaks. It makes a big difference to the human eye and we would never share a Word document, an Excel sheet or a PowerPoint presentation without thinking about white space - although we call it formatting.

White Space

- Stata does not distinguish between one empty space and many empty spaces, or one line break or many line breaks

- It makes a big difference to the human eye and we would never share a Word document, an Excel sheet or a PowerPoint presentation without thinking about white space – although we call it formatting

Vertical lines

```
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_spend c2_inp_total_spending
rename c_IAAP_harv_value c1_IAAP_harv_value
rename c_plots_total_area  c1_total_plotsize
rename  c1_cropPlotShare_??? c1_cropPlotShare_all_???

tempfile BL_append
save  `BL_append'
```

```
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_spend           c2_inp_total_spending
rename  c_IAAP_harv_value       c1_IAAP_harv_value
rename  c_plots_total_area      c1_total_plotsize

rename  c1_cropPlotShare_???    c1_cropPlotShare_all_???

tempfile BL_append
save    `BL_append'
```

Vertical lines

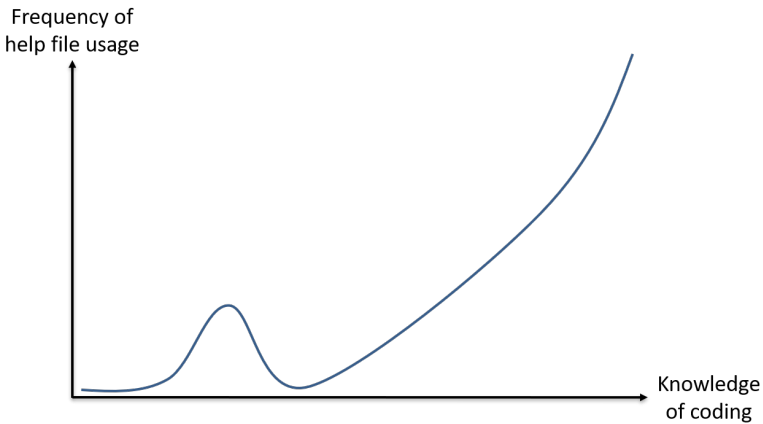
```
*-create dummy for employed
gen employed = 1
replace employed = 0 if _merge == 2
label var employed "Person exists in employment data"
label define yesno 1 "yes" 0 "no"
label val employed yesno
```

```
*-create dummy for being employed
gen          employed = 1
replace      employed = 0 if _merge == 2
label var    employed "Person exists in employment data"
label def             yesno 1 "yes" 0 "no"
label val    employed yesno
```

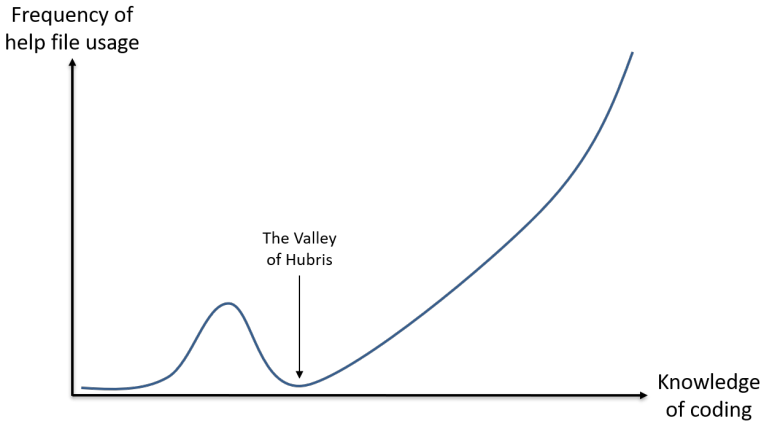
Help files and documentation

Help file usage and coding knowledge

Help file usage and coding knowledge



Help file usage and coding knowledge



Help files

- Type in Stata: `help command_name`
- Get in the habit of using the help file as often as possible!
  - Even with familiar commands, always more to learn
- Help files are only summaries of the reference manual
  - coding practices, common mistakes, alternative approaches
- Access the reference manual by clicking:

[R] regress

How can I get better?

Where are the regressions?

- Nothing I have said so far relates to analysis
- **In coding**, analysis is the easy part as long as the data set is properly set up for analysis
- How to use analysis commands are much easier to google or ask someone how to do, than data cleaning, data management and data quality assurance

When your code works you are only half done

- Ancient proverb

Where are the regressions?

- Compare your code and discuss differences
  - Ask what is easiest to understand if you think of your do-file as an instruction? What is difficult?
  - Apply the question of critical thinking of data work to each others code. (If you do not know what will happen if you have missing data, test it)

- If no one lets you see their code, ask people to look at your code. Have you ever asked someone to help you proofread you your Word document?

Read other peoples code

- Look for code on GitHub
  - `https://github.com/trending/stata` (Stata)
  - `https://github.com/vikjam/mostly-harmless-replication` (Stata and other languages)

- Google code, but before using, ask yourself critical questions about the code you found
  - Why did this person code this way?
  - Does this apply to my context?

Data Base Management Literature

- Your project folder is an informal data base, and very smart people working with data bases have been thinking a lot about this
- I don’t have a specific book to recommend as I don’t know of a book written for our context, so this method is not for the faint hearted

Summary

- Your code is a tool that you should develop as if other people will be using it
- Ask for help from your peers to review your code
- When submitting code, format them as carefully as you would format your resume or your cover letter

Thank you!

Questions?

For more information or further questions please contact DIME Analytics at
`dimeanalytics@worldbank.org`

The End