Development Data Boot Camp for Young Economists

Kellogg Institute for International Studies

Summer 2023

Format

Two weeks over Zoom/remote class. Material will be delivered by Economics graduate student Ge Sun, supervised by Professor Taryn Dinkelman from the Department of Economics.

Dates: May 15-May 19 and May 22-May 26, 2023

Course time and arrangements:

- One 45-minute morning group Zoom lesson at 9:00 - 9:45 EST.
- One 60-minute afternoon group Zoom meeting at 15:00 -16:00 EST.

- There will be a topic-specific problem set to be completed after each lesson. Students will have the chance to review their solutions with the instructor in due course to troubleshoot.

Zoom link: https://notredame.zoom.us/j/91720300613

Meeting code: 917 2030 0613

Overview of the course

The course covers the core concepts necessary for quantitative research assistants. Completion of an Introduction to Statistics class or an equivalent course is a prerequisite for the bootcamp. The course stresses the importance of attention to detail and replicability in the practical implementation of quantitative research tasks in STATA. Topics include basic programming, data cleaning, plotting graphs, and making regression tables; best practices for working collaboratively will also be covered.

There is *no course credit* from taking this training course, and thus there is *no grading*. The main payoff from this course is in the skills you will develop to work as a research assistant in development economics and the preparation it will provide for jobs that require quantitative skills.

Course Materials and Exercises

1. All of the class materials, including data, codes, slides and readings, will be shared with the class via Google Drive. You should review slides before class. The link to the Google drive is: https://drive.google.com/drive/folders/1SW3B4HDNgSZSh_-X17Ae-EBcynJgXXx1?usp=share_link
2. The main resource for this class covers important principles that good research assistants should follow in their work. We hope this will be a resource you can come back to again and again.

   Gentzkow, M., & Shapiro, J. M. (2014). Code and data for the social sciences: A practitioner's guide. *Chicago, IL: University of Chicago.*

3. Readings (include notes and blog posts) for each section will be the class Google Drive (**"Readings"** folder). These will be helpful for understanding the class materials and we recommend you read them after each component of the class.

4. Accessing STATA: You will need to download a short-term licensed copy of STATA, the main statistically packaged used by applied economics researchers. The instructions on the STATA installment (including the license) will be provided right before the class.

5. Exercises that are assigned in class should be submitted to the shared class Google Drive under the folder of **"Exercise"**. You are encouraged to discuss exercises with your group members, but everyone should have an individual submission. The primary purpose of these practice problems is to let you get familiar with the commands we talk about in class. So just try your best and play with the codes!

6. The final exercise will involve generating a short research paper in Overleaf, using output produced in STATA. More details on this project will follow.

Course schedule

1) **<u>Introduction and Preparation [Day 1]</u>**

    a) **Programming languages**
        i) why we use Stata
        ii) get access to Stata at ND
    b) **General Workflow management**
        i) what makes a good research assistant
        ii) examples for folder structure
    c) **Introduction to Randomised Controlled Trials**

2) **<u>Intro to Stata [Days 2, 3, 4 and 5]</u>**

    a) **Stata quick starts**
        i) get to know Stata interface, and Stata windows
        ii) how to import data into Stata, what are the data types
        iii) understand data structure in Stata
    b) **Basic commands in Stata for data management**
    c) **Stata files**
        i) introduction to log file and do file
        ii) write your first do-file
    d) **More advanced commands**
    e) **Basic programming for exploration**
        i) conditional expressions
        ii) scalar and matrices
        iii) macros: global vs. local
        iv) algorithmic structures: Loops

3) **<u>Making Effective Graphs [Day 6]</u>**

    a) **Examples for useful and effective graphs**
    b) **Graphs in two-way families**
        i) basic grammar
        ii) fine tuning graphs: shapes, colors, axis titles and labels, tick marks, graphs in secondary axes.
    c) **Other graph types**
        i) basic plot types: bars, lines, histograms, scatter plots, pie charts
        ii) Stata help document

4) **<u>Linear Regression [Day 7]</u>**

    a) **Understand OLS**
        i) why regression analysis?
        ii) how to interpret regression results

b) **Export regression tables**
    i) how to make regression table in Stata
    ii) export regression tables and why do you want to automate the process of table-making

5) <u>Working with relational databases [Day 8]</u>

a) **How data are related**
    i) time series, panel data, and cross-sectional data
b) **Commands to work with relational databases**
    i) reshape and collapse
    ii) preserve/restore
    iii) merge and append

6) <u>Preparations for your own research! [Day 9]</u>

a) **Generating summary statistics tables**
    i) what is a summary statistics table? why do we need one?
    ii) how to extract summary statistics from Stata and moving the output into other environment such as Latex
b) **LaTex templates and Overleaf**

7) <u>Wrap up [Day 10]</u>

a) **String commands (optional)**
b) **Organization and replicability:** Automating every step of quantitative research production
    i) what does replication mean
    ii) importance of replicability in economic research
        (1) Review AEA data replicability requirements
        (2) Look at a good example of a replication archive
c) **Discuss more about trouble-shooting**