README for "Introduction to Stata for Undergraduate Interns" course materials
By Micole De Vera

These materials were prepared for a 3-hour short course given to incoming interns in CEMFI's Undergraduate Summer Internship program. The target audience for this course are individuals with little or no background in statistical software, including Stata (the interns, for instance, are typically third-year undergraduate students). As such, I have chosen to focus on breadth and not depth, putting emphasis on understanding Stata syntax and basic usage of common commands. None of the commands are covered in-depth and students are encouraged to read the corresponding help files if they need some specific functionality during their internship.

There are three kinds of files in this folder:

- Slides ("Stata_Slides.pdf") – this was not meant to be followed in the class but was prepared as a reference for the students after the class. It includes the syntax for the common commands we studied, common options for each command, and a list of related commands. There are also short discussions on best practices in Stata and coding, in general.
- Do-file ("USI_2021.do") – during the course, I would focus on going through the do-file section-by-section to show different commands in action (using the data files also in this zip file). This is complemented with sometimes referring to the slides.
- Data files:
    - Main panel ("data_1981.dta", "data_1990.dta", etc.) – this is an adapted version of the data used in Blau and Kahn (2017), published in the Journal of Economic Literature (Link to main paper and posted data and code). The main empirical question that is answered by the data is the measurement of the gender wage gap. This is not exactly the data they posted as I edited some parts of the data to illustrate different issues faced in data management (e.g., cutting data by year, adding missing data, reducing variables).
    - Occupation data ("occ_info.dta") – this is a constructed dataset at the 1-digit occupation code level meant to mimic measures of occupational characteristics that could be derived using something like the O*NET Database.

I typically work in directories (and this is a practice I share with the students). For the do-file to work, one should also form directories. In any project folder, I usually include the following subfolders:

- raw – where the raw files are placed. Place all dta files from zip folder here.
- processed – where processed/manipulated data is saved. Distinguishes the raw data.
- scripts – where the do-files (or other code scripts) are placed. Place "USI_2021.do" here.
- out and figs – to store results tables and figures. Usually separated.
- log – to store log files.
- temp – a folder to store temporary datasets, especially in projects that work with large and complicated data.

I hope you find this helpful. If you have any comments and suggestions, feel free to contact me at micole.devera@cemfi.edu.es.