Working with NCES Datasets

Will Doyle

2021-09-08

Intro

In this lesson we'll go over how to access and work with various NCES datasets in Stata. I diverge from the intended use from NCES, as there's a simpler way to get access to the data than the one that they lay out.

```
. capture log close

. log using "nces_datasets.log", replace
-------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/doylewr/lpo_prac/lessons/s1-03-nces_datasets/nces_datasets.log
  log type:  text
 opened on:   8 Sep 2021, 10:59:22
```

Directory Structure

Data files (particularly large ones) should be stored in their own subdirectories. While it's possible to simply dump everything in one big directory, you may find that over time, as the folder grows, it becomes very difficult to find what you need and almost impossible to share your work with others. Yes, your computer can search really well. An organized directory structure is for you, the human. Get into the habit now, and you'll be thankful later.

Creating Directory Structures

In programming, we many times need to move around in directories on a computer. Sometimes we use fixed paths, which specify exactly where something is on the computer, other times we use relative paths. An example of a fixed path would be something like:

/Users/doylewr/lpo_prac/lessons/s1-03-nces_datasets

That path specifies the exact directory on my computer. In general, you really should avoid fixed paths, because everyone's computer is different. However, you might have something like a clone of a GitHub repository on your computer. Within that repository, you can specify relative paths to clarify where you want the program to look. A standard directory structure for a statistical programming project is something like this:

____'project_directory/

```
----data/

        -----source/

        -----analysis/

----scripts/

----output/

        -----tables/

        -----graphics/

----paper`___
```

Since our code exists in the ./scripts directory, to access the source data names source_data.dta we would need to go up one level to the main project directory and then down into the source data directory. The command for this in Stata would be:

```
use ../data/source/source_data.dta
```

The `../` means to go up one level. Using `./` means to go into a subdirectory, or down one level.

In my github repository, I store large data files in the data directory. To access that directory from the current lesson I need to go up two levels and then into the data directory, so the relative path is: `../../data/`.

Working with globals

We're going to set the structure for ourselves using global macros. In Stata, a macro is a variable that can be set to some value. There are two types of macros in Stata– global and local. A global macro is persistent during a session (when Stata is open). A local macro is forgotten as soon as a script (do file) is run. In general, I will encourage you to use local macros, as their use enables better programming practices and replication. However, using global macros is a good idea for data management.

First we tell Stata what a macro will represent:

```
. global ddir "../../data/"
```

What the above means is that every time I call that macro, Stata will know I means the directory in question. We can test this by asking Stata to display the global ...

```
. display "$ddir"
../../data/
```

. . . and there you have it.

One big takeaway from all of this is that you should *never* include a `cd` statement in a do file that references a specific spot on your computer. Either don't include a cd command at all, or include a cd command that makes use of a relative directory strucure. The easiest (but not necessarialy the best) way to do this is to assume that the do file and the

I'm also going to get the information for my current directory so I can easily return to it.

```
. global cdir `c(pwd)'

. di "$cdir"
/Users/doylewr/lpo_prac/lessons/s1-03-nces_datasets
```

Working with HSLS

The high school longitudinal study of 2009 tracks a set of students who began high school in 2009. It has been updated in 2012 and again in 2016. It's a great source of information about how students navigate high school and make the transition to college or the workforce (and in many cases both).

HSLS can be accessed using the nces electronic codebook: https://nces.ed.gov/OnlineCodebook Once variables have been selected from the codebook, they can be accessed using the ***use*** . . .***using*** approach below:

```
. use  ///
         STU_ID ///
         X1SES ///
         using "${ddir}hsls_17_student_pets_sr_v1_0.dta", ///
         clear


. save "${ddir}hsls_analysis.dta",replace
file ../../data/hsls_analysis.dta saved
```

Working with NHES

The National Household Education Survey collects data on the education activities of children and adults in the United States. The NHES has varying emphases in different years.

Because of the different emphases, NHES will include different data files in each year. It's important to know which data file a given variable comes from. Below, I open up the early childhood program data file and extract a few variables regarding the child's participation in early childhood programs.

```
. use ///
 BASMID ///
```

```
 CPNNOWX ///
 CPTYPE ///
 CPHRS ///
 using "${ddir}nhes_16_ecpp_v1_0.dta", clear


. renvars *, lower

. save "${ddir}nhes_analsyis.dta", replace
file ../../data/nhes_analsyis.dta saved
```

Working with ECLS 2011

ECLS 2011 uses a nationally represntative sample of students that were in kindergarten as of 2011. This study is excellent for tracking younger students as they progress through early grades.

ECLS is a bit different in that NCES doesn't have any equivalent of the online codebook for it. Instead we have to navigate it using some other tools.

NCES provides a do file, a dictionary file, and a data file (zipped) for ECLS 2011. The code below assumes that you have downloaded the zip file `ChildK5p.zip` from the ECLS data products website.

```
. clear

. capture confirm file "${ddir}ChildK5p.zip"

. if _rc==601 {
.         copy https://nces.ed.gov/ecls/data/2019/ChildK5p.zip "${ddir}ChildK5p.zip"
. }

. capture confirm file "${ddir}childK5p.dat"

. if _rc==601 {
.         unzipfile ChildK5p.zip
. }

. capture confirm file "${ddir}ECLSK2011_K5PUF.do"

. if _rc==601 {
.         copy https://nces.ed.gov/ecls/data/2019/ECLSK2011_K5PUF.do ${ddir}ECLSK2011_
. }

. capture confirm file "${ddir}ECLSK2011_K5PUF.dct"

. if _rc==601 {
```

```
.         copy https://nces.ed.gov/ecls/data/2019/ECLSK2011_K5PUF.do "${ddir}ECLSK2011
. }

. capture confirm file "${ddir}ECLSK2011_K5PUF.dta"

. if _rc==601{
. cd $ddir
. do ECLSK2011_K5PUF.do
. cd $cdir
. }

. use   ///
  CHILDID ///
  X9SESL_I ///
  X9INCCAT_I ///
  using "${ddir}ECLSK2011_K5PUF.dta", clear


. renvars *, lower

. save "${ddir}ecls_analsyis.dta", replace
file ../../data/ecls_analsyis.dta saved

. exit
```