The NLS Investigator

William Howard Beasley (Howard Live Oak LLC, Norman)
Joseph Lee Rodgers (University of Oklahoma, Norman)
Kelly Meredith (University of Oklahoma, Norman)
David Bard (University of Oklahoma Health Sciences Center, OKC)

March 10, 2012

Abstract

We describe how to use the NLS Investigator to select and download variables. These steps aren't specific to R, and are necessary regardless of the choice of statistical software.

1 Steps to Select Variables and Download Data

1. Browse to `http://www.nlsinfo.org/investigator/`. Select the 'REGISTER' link in the top right, and create a personal account.

2. Select your desired cohort, in the drop down box titled, "Select the study you want to work with:". In this screenshot, Gen2 of the Nlsy1979 sample is selected.
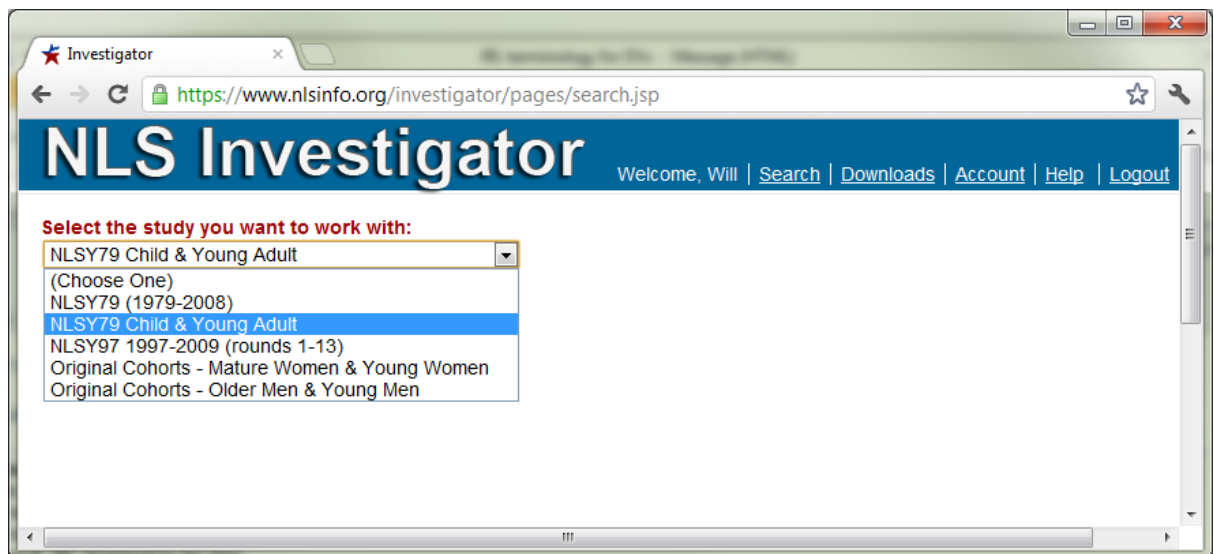
3. Select the variables. There are tens of thousands of variables in some cohorts, and selecting the correct ones can require careful attention and a few tricks. A thorough tutorial begins on the NLS Investigator page: `http://www.nlsinfo.org/InvestigatorGuide/investigator_guide_TOC.html`. To better leverage the NLSY's extensive variable set (and to avoid mistakes), we recommend that researchers dedicate time to this tutorial. However for the purposes of this vignette, we'll simply select a few easy variables.

   First, in the 'Variables Search' tab, select 'Word in Title (enter search term)'. Second, type "other symptom - f" in the textbox. Third, clicking the 'Display Variables' button should retrieve at least four Gen2 variables whose title starts with "other symptom - f". Fourth, supposing we care about only the fevers, click their two corresponding checkboxes. (If you're curious, the 'XRND' value for year stands for cross round; these values come from the most recent survey the subject completed).

4. Review your selected variables by clicking the corresponding tab. Notice that several important variables are automatically included in every dataset. In real research these steps are iterated many times, as you repeated select, then review, then save, then re-select, then re-review, then... But we'll move on, because these seven variables are good enough for an example.

5. When the dataset is complete, it is time to save the tagset. A **tagset file** contains the IDs of the desired variables, *but no data values*. The actual values are contained in the **data file**, which is discussed in the next two steps. These issues are covered further in the tutorial.

6. The first step of the download process is to create a zip file on the NLS server. Click the

'Save/Download' tab, and then the 'Advanced Download' tab. As far as the NlsyLinks package is concerned, only the the 'Comma-delimited datafile...' box needs to be checked. Then provide a 'Data filename'. Finally, click the 'Start Download' button.

7. The second step of the download process is to transfer the zip file to your local computer. Click the brown 'download' hyperlink.

8. Open the zip file and extract the *.csv file to a location that the vignette examples have permission to read. Then try some of the vignette examples.

2   Using Multiple Tagsets

Tagsets were introduced in Step 5, but are discussed more thoroughly here. Tagsets make large projects more managable in two ways. First, they save effort and reduce errors because you don't have to re-select all the variables everytime you revist the NLS Investigator. You can save and load the tagsets during subsequent sessions. In fact, you can save multiple tagsets on the NLS server.

Second, using *multiple* tagsets provides a convenient approach to organize and compartmentalize your variables. It's not unusal for some complicated longitudinal studies to use hundreds of NLSY variables. We've found it easier to manage three tagsets of 100 variables, than one tagset of 300. Since all NLSY tagsets (and therefore their datasets) automatically include the subject ID later merging the multiple datasets is trivial with statistical software.

In R, use the merge function; the by argument should be 'R0000100' for NLSY79 Gen1 subjects, 'C0000100' for Nlsy79 Gen2 subjects, and 'R0000100' for NLSY97 subjects. For example, a study about intelligence

and teenage fertitity would have one tagset containing the intelligence variables, and a second tagset containing the fertility variables. These two tagsets eventually could produce two CSV datafiles called `IQ.csv` and `Fertility.csv`, located in the `BGResearch` directory. The R code to read and merge these two datasets could be as simple as

```
dsIQ <- read.csv('C:/BGResearch/IQ.csv', header=TRUE)
dsFertility <- read.csv('C:/BGResearch/Fertility.csv', header=TRUE)
ds <- merge(dsIQ, dsFertility, by="C0000100")
```

In SAS, the `merge` function (and its `BY` argument) behaves similarly as above. Notice that the SAS documentation states, "Before you can perform a match-merge, all data sets must be sorted by the variables that you want to use for the merge", which is `C0000100` in this example.

3   Tagset History

We recommend saving the tagsets on both the NLS server and your local machine. Tagsets are not guaranteed to be retained on the Server more than 90 days. Locally save when variables are added or removed from the tagset in a consistent location, and name the file to reflect the current date. The files can be read with any simple text editor. The directory below contains the partial evoluation of four distinct tagsets (i.e., 'Gen1Links', 'Gen2Links', 'Gen2ImplicitFather', 'Gen2LinksFromGen1').

Name	Date modified	Type	Size
BU 2010-07-12 Gen1Links.NLSY79	7/12/2010 7:58 PM	NLSY79 File	2 KB
BU 2010-10-26b Gen2Links.CHILDYA	10/26/2010 10:56 ...	CHILDYA File	1 KB
BU 2011-07-27 Gen1Links.NLSY79	7/27/2011 8:59 PM	NLSY79 File	2 KB
BU 2011-07-27 Gen2Links.CHILDYA	7/27/2011 9:03 PM	CHILDYA File	1 KB
BU 2011-07-27b Gen1Links.NLSY79	7/27/2011 9:45 PM	NLSY79 File	2 KB
BU 2011-07-29a Gen2Links.CHILDYA	7/29/2011 1:09 PM	CHILDYA File	2 KB
BU 2011-07-31 Gen1Links.NLSY79	7/31/2011 2:18 PM	NLSY79 File	2 KB
BU 2011-08-01a Gen2ImplicitFather.CHILDYA	8/1/2011 3:16 PM	CHILDYA File	1 KB
BU 2011-08-01a Gen2Links.CHILDYA	8/1/2011 3:19 PM	CHILDYA File	2 KB
BU 2011-08-01b Gen2ImplicitFather.CHILDYA	8/1/2011 5:17 PM	CHILDYA File	1 KB
BU 2011-08-01b Gen2Links.CHILDYA	8/1/2011 3:51 PM	CHILDYA File	3 KB
BU 2011-08-01c Gen2Links.CHILDYA	8/1/2011 3:56 PM	CHILDYA File	2 KB
BU 2011-08-12 Gen1Links.NLSY79	8/12/2011 5:05 PM	NLSY79 File	2 KB
BU 2011-08-15 Gen1Links.NLSY79	8/15/2011 10:11 PM	NLSY79 File	2 KB
BU 2011-08-15 Gen2LinksFromGen1.NLSY79	8/15/2011 10:08 PM	NLSY79 File	1 KB

4   Notes

These screenshots were taken Feb 23, 2012 with Google Chrome 17 and Windows 7 SP1 Enterprise.