Introduction to NLS Investigator
(v. 1.5)

Oscar Torres-Reyna
otorres@princeton.edu

http://www.princeton.edu/~otorres/
Introduction

This document offers a quick introduction to the NLS Investigator. It follows a basic approach and focus on searching, downloading and putting the data into Stata.

If you are not familiar at all with the site, I strongly recommend to follow the example in this document.

It is important to clarify that this document does not cover all the complexities of the NLS site. For more details I suggest to look at the following links:

Getting Started: How to Get the Most from This Site
•  https://www.nlsinfo.org/content/getting-started

How to Use the NLS Investigator
•  https://www.nlsinfo.org/InvestigatorGuide/investigator_guide_TOC.html

To start using the NLS Investigator, please go to the following page:
•  https://www.nlsinfo.org/investigator/
NLS investigator
https://www.nlsinfo.org/investigator/

You can search for data as a guest

If you have an account login here
Selecting a data source

Select the study you want. For this exercise we will work with NLSY97
After selecting the study you will see a series of tabs. The first one is to choose a tagset. Tagsets are basically saved searches, if you have not save them or it is your first time you will not need this.

Notice that by the default six variables will be added to your data: id (can’t remove this) and the following demographics (optional): gender, age, race/ethnicity and birthday (month and year).

Select “Variable Search” and go to the next slide.
“Variable Search” offers three search modes: by topics (index), by searching the index and a general search. Notice that the site does not offer access to the full data sets (see the note above). For this exercise we will use the “Search” tab, go to the next slide.
The general search offers a variety of options. From the dropdown menu select the type of search you want. See the next slide.
Searching for data (2)

This is a ‘nested’ search, we are looking for ‘age’ within ‘demographic indicators’

You can do conditional search with “AND”/”OR”

Results will show here, sometimes you will see a “+” next to the ‘RNUM’ meaning that there are more variables available (as a subset of responses). Check the square next to ‘RNUM’ to select all variables or check the ones you need.
Searching for data (3)

If you hover the cursor over the question name a window will pop-up with detail information about that variable.
“Codebook’ tab

Here you can get additional information on your variables along with some basic stats.
Once you are satisfied with your search, you can save it as a “tagset’. You can keep the defaults, select a name (in this case we choose ‘Age’) and click on “Save”. Go to “Advanced Download”, see next slide.
Select the data format you want. For this example we will select Stata and comma-delimited along with the codebook. If you are an R user, R can easily read Stata, SPSS or comma-delimited files.

1. Select the data format you want. For this example we will select Stata and comma-delimited along with the codebook. If you are an R user, R can easily read Stata, SPSS or comma-delimited files.

2. Select a name for the file and click “Download”
Downloading data (3)

If you set your browser to ask you where to save, a pop-up window will prompt you to select a folder location for the zip file, click “OK” and select the folder. If it gets downloaded automatically, then the file should be in the “Downloads” folder or the default place for downloads.
Unzipping the files downloaded from the NLS site

Find the zip file downloaded from the NLS site and extract its contents. In this example, using 7-Zip, right-click and select “Extract Here”.

Two files are of importance, the *.dct which is the dictionary file that has the layout to read the data. And the do-file (*.do) which has additional commands to format the data once in Stata.
The *.csv file can be read directly by Excel, Stata, SPSS, SAS or R.
Reading the NLS data into Stata

1. Change the working directory to the location of the NLS files.
   (type in the command window below)

2. To read the data use the `infile` command using the dictionary file (*.dct) downloaded from the NLS site.
   (type in the command window below)

3. After running `infile`, the variables window will populate with information about your dataset.

Notice the working directory here.
Opening Stata’s do-file editor

1. Type `doedit` in the command window or click on the little ‘notebook’ icon.

2. The do-file editor will pop-up. Go to File->Open and look for the do-file from the NLS download. In this example is called “Age-value-labels.do”.
Adding value labels to the original variables

Once the file `Age-value-labels.do` is open, select all the commands starting with “label…” and run them by clicking on the last icon at the top.
Renaming variables

In the file Age-value-labels.do you need to remove the “/*” in row 44 and “*/” in row 65. This will uncomment the commands between them.

One important edit is adding the years to the names of the variables that change over time. In the example below, rows 49, 52-64 had originally the same name except that now each has its corresponding year, this makes them unique. Other minor edits were done in rows 46-48 and 51 (compare to the original).

For more on rename see http://dss.princeton.edu/training/StataTutorial.pdf#page=33
Looking at the data

If you type `browse` in the command line you will see the data set. As it is now, each row represents one individual. While you can start working with this format, it is not ideal for panel data analysis.
Preparing the data for panel analysis

To run panel regression you need to reshape the data so it looks like the example in this document:  
http://dss.princeton.edu/training/Panel101.pdf

For details on how to reshape data see here:  
http://dss.princeton.edu/training/DataPrep101.pdf#page=27

Since we already have a unique id for this dataset, in the command line we can just type

```
reshape long CV_AGE_INT_DATE, i(PUBID) j(YEAR)
```

Notice that reshape only applies to variables that are observed over time (i.e. have a year suffix), in this case `CV_AGE_INT_DATE`. If you have other variables you can add them to the list, for example:

```
reshape long CV_AGE_INT_DATE VAR2 VAR3 VAR4, i(PUBID) j(YEAR)
```
Looking at the reshaped data

If you type `browse` in the command line you will see that the dataset has only one `CV_AGE_INT_DATE` variable and all the years are in rows. Here, each row represents an individual per year. Data for individual 1 ends at row 14, data for individual 2 starts at row 15. You can analyze the data using the panel data techniques shown in this document: [http://dss.princeton.edu/training/Panel101.pdf](http://dss.princeton.edu/training/Panel101.pdf)
Adding value labels (...again, part 1)

Notice that reshaping the data removed the value labels added before (type `tab GENDER` to check it). If you go back to the do-file `Age-value-labels.do`, select and copy (Ctrl-C) rows 3-12 (or until you see categories). Copy the code at the end of the do-file (see next slide).
Adding value labels (…again, part 2)

Notice that the lines starting with “label values…” have the variables’ old names. To match the old and new names, look at the ‘rename’ section, rows 46-48, 50, 51.
Manually replace the first name appearing in the lines starting with “label values…” (in this example start with ‘R’) with the new names as they appear in the rename section (remember that Stata is case sensitive). Select the code (see the ‘After’ column below) and run it by clicking on the last icon in the do-file.

Before

```
lable define v1R0536300 1 "Male" 2 "Female" 0 "No Inf:
lable values R0536300 v1R0536300
lable define v1R0536401 1 "1: January" 2 "2: February"
lable values R0536401 v1R0536401
lable define v1R1194100 12 "12" 13 "13" 14 "14" 15 ":
lable values R1194100 v1R1194100
lable define v1R1235800 1 "Cross-sectional" 0 "Oversamp
lable values R1235800 v1R1235800
lable define v1R1482600 1 "Black" 2 "Hispanic" 3 "Mix
label values R1482600 v1R1482600
```

After

```
lable define v1R0536300 1 "Male" 2 "Female" 0 "No Inf:
label values v1R0536300 1 "Male" 2 "Female" 0 "No Inf:
label define v1R0536401 1 "1: January" 2 "2: February"
label values v1R0536401 1 "1: January" 2 "2: February"
label define v1R1194100 12 "12" 13 "13" 14 "14" 15 ":
label values v1R1194100 12 "12" 13 "13" 14 "14" 15 ":
label define v1R1235800 1 "Cross-sectional" 0 "Oversamp
label values v1R1235800 1 "Cross-sectional" 0 "Oversamp
label define v1R1482600 1 "Black" 2 "Hispanic" 3 "Mix
label values v1R1482600 1 "Black" 2 "Hispanic" 3 "Mix
```

For more on adding value labels see here [http://dss.princeton.edu/training/StataTutorial.pdf#page=34](http://dss.princeton.edu/training/StataTutorial.pdf#page=34)
The End

Do not forget to save the datafile by either using the menu, go to ‘File’->‘Save As’ or typing:

```
save name-of-your-file, replace
```

This will save the Stata file in the working directory specified at the beginning, it will have extension *.dta. The first time the ‘replace’ option is not necessary but if after saving you make changes to the dataset you will need to use it to update the file.

Now the data is ready for analysis, see here

- [http://dss.princeton.edu/training/Panel101.pdf](http://dss.princeton.edu/training/Panel101.pdf)
- [http://dss.princeton.edu/training/StataTutorial.pdf](http://dss.princeton.edu/training/StataTutorial.pdf)

Once again, for more details I suggest to look at the following links:

*Getting Started: How to Get the Most from This Site*

- [https://www.nlsinfo.org/content/getting-started](https://www.nlsinfo.org/content/getting-started)

*How to Use the NLS Investigator*