



Introduction to NLS Investigator

(ver. 1.5)

Oscar Torres-Reyna
Data Consultant
otorres@princeton.edu



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/>

NLS Investigator

Welcome, Guest | [LOGIN](#) | [Register](#) | [Search](#) | [Help](#)

Log In

Username:

Password:

Login

[I cannot access my account](#)

Welcome to Investigator

Sponsored by the Bureau of Labor Statistics, the National Longitudinal Surveys (NLS) are a family of surveys dedicated to tracking the labor market and other life experiences of American men and women.

The seven NLS cohorts are:

- National Longitudinal Survey of Youth 1997 (NLSY97)
- National Longitudinal Survey of Youth 1979 (NLSY79)
- NLSY79 Child and Young Adult
- Older Men
- Mature Women
- Young Men
- Young Women

To access data for any of the seven NLS cohorts use the login box to the left or [begin searching](#) as guest.

NLSY User-Initiated Questions: We're soliciting suggestions for new questions to add to the NLSY97, NLSY79, and child/young adult surveys. Please visit the [NLSY User-Initiated Questions](#) page to learn how to make an informal suggestion or submit a formal proposal.

New with updated NLSY79 release: Beta version of [Employer History roster](#) now available.

Attention

In the event that Investigator does not appear to be working correctly, first please try to clear your browser cache. If you continue to have issues, please contact usersvc@chrr.osu.edu

You can search for data as a guest

If you have an account login here

Selecting a data source

NLS Investigator

Welcome, Guest | [Login](#) | [Register](#) | [SEARCH](#) | [Help](#)

Select the study you want to work with:

(Choose One) ▼

- (Choose One)
- NLSY79 (1979-2010)
- NLSY79 Child/Young Adult 1986-2010
- NLSY97 1997-2010 (rounds 1-14)**
- Original Cohorts - Mature Women & Young Women
- Original Cohorts - Older Men & Young Men

Select the study you want. For this exercise we will work with NLSY97

Tagsets tab

NLS Investigator

Welcome, Guest | [Login](#) | [Register](#) | [SEARCH](#) | [Help](#)

Select the study you want to work with:

NLSY97 1997-2010 (rounds 1-14) ▼

Released September 17, 2012

Additional Resources:

[Errata](#), [Documentation](#) (user's guide, questionnaires and other materials)

[Custom Weights](#)

To start a new search [click here](#)

Choose Tagsets

Variable Search

Review Selected Variables (6)

Codebook

Save / Download

Required / Recommended Variables:

- Required ID Variable - PUBID will always be selected (1 variable)
- Recommended Demographic Variables (5 variables)

Saved Tagsets (on server):

None Available

Upload Tagset (from PC):

No file selected.

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 fist 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” tab

NLS Investigator

Welcome, Guest | [Login](#) | [Register](#) | [SEARCH](#) | [Help](#)

Select the study you want to work with:

NLSY97 1997-2010 (rounds 1-14) ▼

Released September 17, 2012

Additional Resources:

[Errata, Documentation](#) (user's guide, questionnaires and other materials)

[Custom Weights](#)

To start a new search [click here](#)

Choose Tagsets

Variable Search

Review Selected Variables (6)

Codebook

Save / Download

Browse Index

Browse Index with Search

Search

Index of Selected Variables

Options ▲

- + Education, Training & Achievement Scores (16972)
- + Employment (21210)
- + Household, Geography & Contextual Variables (7047)
- + Dating, Marriage & Cohabitation (1281)
- + Sexual Activity, Pregnancy & Fertility (1427)
- + Children (1508)
- + Parents, Family Process & Childhood (169)
- + Income, Assets & Program Participation (3959)
- + Health (1313)
- + Attitudes, Expectations & Non-cognitive Tests (372)
- + Crime & Substance Use (5863)
- + Survey Methodology (268)

Please browse the index on the left to display variables.

This index contains a set of NLSY97 variables commonly used in research and is not the full data set.

“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

Searching for data (1)

The general search offers a variety of options. From the dropdown menu select the type of search you want. See the next slide

The screenshot shows the NLS Investigator search interface. At the top, there is a blue header with the text "NLS Investigator" and navigation links for "Welcome, Guest", "Login", "Register", "SEARCH", and "Help". Below the header, there is a section for selecting a study, with a dropdown menu currently showing "NLSY97 1997-2010 (rounds 1-14)" and the release date "Released September 17, 2012". To the right of this section, there are "Additional Resources" including links for "Errata, Documentation" and "Custom Weights", and a link to "click here" to start a new search. Below these elements is a horizontal navigation bar with buttons for "Choose Targets", "Variable Search", "Review Selected Variables (6)", "Codebook", and "Save / Download". Underneath this bar are three buttons: "Browse Index", "Browse Index with Search", and "Search". A red arrow points from the text box above to the "Search" button. Below the "Search" button, there is a section titled "Create search criteria below:" which contains a search form. The form has a dropdown menu with "(Choose One)" selected, and a text input field. A red arrow points from the text box above to this dropdown menu. The dropdown menu is open, showing a list of search criteria options: "(Choose One)", "Area of Interest (pick from list)", "Word in Title (pick from list)", "Word in Title (enter search term)", "Question Text (enter search term)", "Question Name (pick from list)", "Question Name (enter search term)", "Reference Number (pick from list)", "Reference Number (enter search term)", "Survey Year (pick from list)", "Codebook (enter search term)", and "Variable Type (pick from list)". To the right of the search form, there are "Add" and "Reset" buttons, and a "Display Variables" button. At the bottom of the page, there is a footer with links for "NLS Home", "NLS Bibliography", and "Privacy Policy", and a contact email "usersvc@chrr.osu.edu".

Searching for data (2)

NLS Investigator
Welcome, Guest | [Login](#) | [Register](#) | [SEARCH](#) | [Help](#)

Select the study you want to work with:
 NLSY97 1997-2010 (rounds 1-14)
Released September 17, 2012

Additional Resources:
[Errata, Documentation](#) (user's guide, questionnaires and other materials)
[Custom Weights](#)

To start a new search [click here](#)

Choose Tagsets
Variable Search
Review Selected Variables (20)
Codebook
Save / Download

Browse Index

Browse Index with Search

Search

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

Create search criteria below: Include only intersecting (AND) ▾

Area of Interest (pick from list)	equals	DEMOGRAPHIC INDICATORS	Remove
Word in Title (enter search term)	contains	age	Remove
Word in Title (enter search term)	contains	at interview date	Add Reset

You can add more terms if needed

14 Variables Display Variables

Options ▾ Showing 14 of 14 filtered by All Variables ▾

	RNUM ▲	QUESTION NAME	VARIABLE TITLE	YEAR
1	<input checked="" type="checkbox"/>	R11941.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	1997
2	<input checked="" type="checkbox"/>	R25535.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	1998
3	<input checked="" type="checkbox"/>	R38763.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	1999
4	<input checked="" type="checkbox"/>	R54537.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2000
5	<input checked="" type="checkbox"/>	R72160.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2001
6	<input checked="" type="checkbox"/>	S15314.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2002
7	<input checked="" type="checkbox"/>	S20010.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2003
8	<input checked="" type="checkbox"/>	S38011.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2004
9	<input checked="" type="checkbox"/>	S54010.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2005
10	<input checked="" type="checkbox"/>	S75012.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2006
11	<input checked="" type="checkbox"/>	T00085.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2007
12	<input checked="" type="checkbox"/>	T20111.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2008
13	<input checked="" type="checkbox"/>	T36015.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2009
14	<input checked="" type="checkbox"/>	T52014.00 CV AGE INT DATE	RS AGE AT INTERVIEW DATE	2010

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)

NLS Investigator

Select the study you want to work with:
 NLSY97 1997-2010 (rounds 1-14) Released September 17, 2012

Additional Resources:
[Errata](#), [Documentation](#) (user's guide)
[Custom Weights](#)

To start a new search [click here](#)

Choose Tagsets
Variable Search
Review Selected Variables (20)
Codebook

Browse Index
Browse Index with Search
Search

Create search criteria below.

Area of Interest (pick from list)	equals	DEMOGRAPHIC INDICATORS
Word in Title (enter search term)	contains	age
Word in Title (enter search term)	contains	at interview date

R11941.00 [CV_AGE_INT_DATE] Survey Year: 1997
 PRIMARY VARIABLE

RS AGE AT INTERVIEW DATE

Age as of interview date.

0	0 TO 11: LESS THAN 12
1169	12
1726	13
1858	14
1877	15
1719	16
614	17
21	18
0	19 TO 999: GREATER THAN 18

8984	

Refusal(-1) 0
 Don't Know(-2) 0
 TOTAL =====> 8984 VALID SKIP(-4) 0 NON-INTERVIEW(-5) 0

Min: 12 Max: 18 Mean: 14.35

Hard Minimum: [0] Hard Maximum: [25]

Lead In: R11940.00[Default]
 Default Next Question: R11980.00

Options x Showing 14 of 14 filtered by All Variables

	RNUM	QUESTION NAME	VARIABLE TITLE	YEAR
1	R11941.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	1997
2	R25535.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	1998
3	R38763.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	1999
4	R54537.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2000
5	R72160.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2001
6	S15314.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2002
7	S20010.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2003
8	S38011.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2004
9	S54010.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2005
10	S75012.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2006
11	T00085.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2007
12	T20111.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2008
13	T36015.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2009
14	T52014.00	CV_AGE_INT_DATE	RS AGE AT INTERVIEW DATE	2010

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

“Codebook” tab

NLS Investigator

Welcome, Guest | [Login](#) | [Register](#) | [SEARCH](#) | [Help](#)

Select the study you want to work with:

NLSY97 1997-2010 (rounds 1-14) ▼

Released September 17, 2012

Additional Resources:

[Errata, Documentation](#) (user's guide, questionnaires and other materials)

[Custom Weights](#)

To start a new search [click here](#)

Choose Tagsets

Variable Search

Review Selected Variables (20)

Codebook

Save / Download

R00001.00 [PUBID] Survey Year: 1997
PRIMARY VARIABLE

PUBID, YOUTH CASE IDENTIFICATION CODE

COMMENT: YOUTH CASE IDENTIFICATION CODE

0	0
998	1 TO 999
999	1000 TO 1999
997	2000 TO 2999
996	3000 TO 3999
998	4000 TO 4999
996	5000 TO 5999
994	6000 TO 6999
994	7000 TO 7999
989	8000 TO 8999
23	9000 TO 9999

8984

Refusal (-1)	0
Don't Know (-2)	0
TOTAL =====>	8984
VALID SKIP (-4)	0
NON-INTERVIEW (-5)	0

Min: 1 Max: 9022 Mean: 4504.3

Hard Minimum: [0] Hard Maximum: [99999999]

Lead In: [R72976.00](#) [Default]

Default Next Question: [R05363.00](#)

Here you can get additional information on your variables along with some basic stats

Lead In

[R72976.00](#) [Default]

Default Next

[R05363.00](#)

Selected Variables

R00001.00



Stats as Graphs



Questionnaire Links

[NLSY97 Round 1 Parent and Screener Questionnaires](#)

[NLSY97 Round 1 Youth Questionnaire](#)

Areas of Interest

SYMBOLS

TYPE: SYMBOLS

Downloading data (1)

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

The screenshot shows the NLS Investigator web interface. At the top, there is a blue header with the text "NLS Investigator" and navigation links for "Welcome, Guest", "Login", "Register", "SEARCH", and "Help". Below the header, there is a section for selecting a study, with a dropdown menu showing "NLSY97 1997-2010 (rounds 1-14)" and a release date of "Released September 17, 2012". To the right, there are "Additional Resources" links for "Errata, Documentation" and "Custom Weights". A navigation bar contains buttons for "Choose Tagsets", "Variable Search", "Review Selected Variables (20)", "Codebook", and "Save / Download". The "Save Tagset" button is highlighted, and a sub-menu is open with options for "Basic Download", "Advanced Download", and "Manage Downloads". The "Advanced Download" option is selected. Below this, there is a section titled "Choose where to save the tagset of your selected variables:" with radio buttons for "Save to PC" (selected) and "Save on our server". There is also a "Tagset Type" section with radio buttons for "By Rnum" (selected) and "By Qname with Year". A "Filename:" field contains the text "Age" and a "Save" button. A note at the bottom of the sub-menu states: "Tagsets saved on your PC or our server can be reloaded at a later time through the 'Choose Tagsets' tab. Tagsets stored on our server with no activity may be deleted after 90 days."

Downloading data (2)

NLS Investigator

Welcome, Guest | [Login](#) | [Register](#) | [SEARCH](#) | [Help](#)

Select the study you want to work with:
NLSY97 1997-2010 (rounds 1-14) Released September 17, 2012

Additional Resources:
[Errata, Documentation](#) (user's guide, questionnaires and other materials)
[Custom Weights](#)

To start a new search [click here](#)

[Choose Tagsets](#) | [Variable Search](#) | [Review Selected Variables \(20\)](#) | [Codebook](#) | [Save / Download](#)

[Save Tagset](#) | [Basic Download](#) | [Advanced Download](#) | [Manage Downloads](#)

Customize your advanced download:

- Create Download of Data**
 - Tagset (list of selected variables)
 - SAS® control file (includes the datafile of selected variables)
 - SPSS® control file (includes the datafile of selected variables)
 - STATA® dictionary file of selected variables
 - Codebook of selected variables
 - Short Description File
 - Comma-delimited datafile of selected variables (to be read in Excel, etc.)
 - Column headers -- Use Reference Number Question Name (does not guarantee uniqueness)
- Create Frequency / Table**
- Apply Universe Restrictors** ([How to use Universe Restrictors](#))

Filename:
Filename must only contain alpha, numeric, hyphen or underscore characters.

*Download status appears under 'Manage Downloads' tab.
Downloads may be deleted after 10 days of inactivity.*

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.

NLS Investigator Welcome, Guest | [Login](#) | [Register](#) | [SEARCH](#) | [Help](#)

Select the study you want to work with:
NLSY97 1997-2010 (rounds 1-14)
Released September 17, 2012

Additional Resources:
[Errata, Documentation](#) (user's guide, questionnaires and other materials)
[Custom Weights](#)

To start a new search [click here](#)

Choose Tagsets | Variable Search | Review Selected Variables (20) | Codebook | Save / Download

Save Tagset | Basic Download | **Advanced Download** | Manage Downloads

Download Status:
All downloads are available. Please click a download link below to begin downloading.

All Available Downloads:

	Date	Study Name	Size	Download
1		NLSY97 Age	152.5K	download

Delete Selected Files

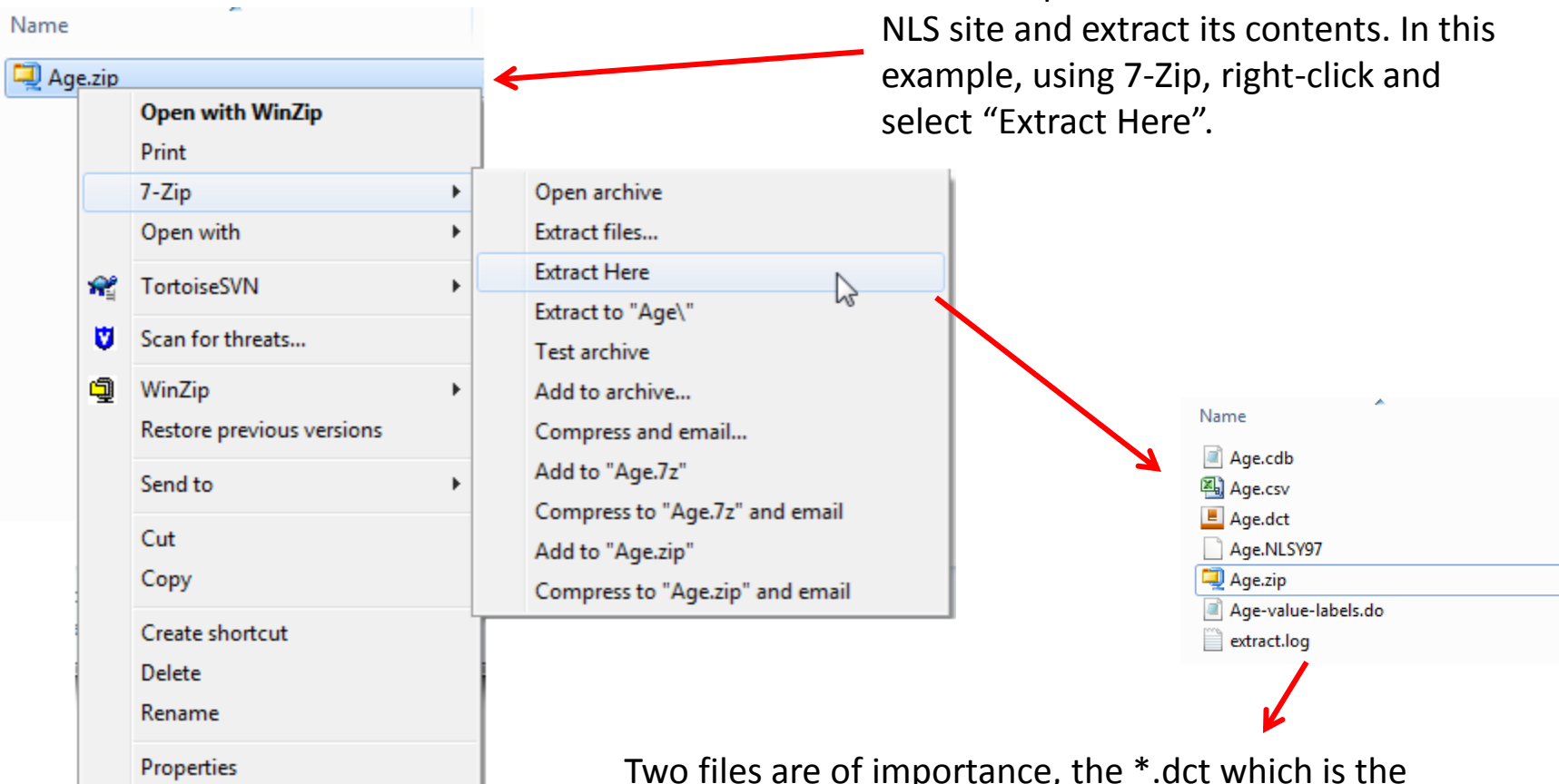
Opening Age.zip
You have chosen to open:
Age.zip
which is a: Compressed (zipped) Folder
from: https://www.nlsinfo.org
What should Firefox do with this file?
 Open with WinZip (default)
 Save File
 Do this automatically for files like this from now on.
OK Cancel

Click here to start downloading the data

NLS Home | NLS Bibliography | Privacy Policy For help, email usersvc@chrr.osu.edu

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

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

Review window
(anything typed in the command window will appear here)

```
2. New update available; type -update all-  
# Command _rc  
1 cd "H:\MyData\NLS"  
2 infile using Age.dct
```

Command window

```
. cd "H:\MyData\NLS"  
H:\MyData\NLS  
  
. infile using Age.dct  
  
infile dictionary {  
R0000100 "PUBID - YTH ID CODE 1997"  
R0536300 "KEY!SEX (SYMBOL) 1997"  
R0536401 "KEY!BDATE M/Y (SYMBOL) 1997"  
R0536402 "KEY!BDATE M/Y (SYMBOL) 1997"  
R1194100 "CV_AGE_INT_DATE 1997"  
R1235800 "CV_SAMPLE_TYPE 1997"  
R1482600 "KEY!RACE_ETHNICITY (SYMBOL) 1997"  
R2553500 "CV_AGE_INT_DATE 1998"  
R3876300 "CV_AGE_INT_DATE 1999"  
R5453700 "CV_AGE_INT_DATE 2000"  
R7216000 "CV_AGE_INT_DATE 2001"  
S1531400 "CV_AGE_INT_DATE 2002"  
S2001000 "CV_AGE_INT_DATE 2003"  
S3801100 "CV_AGE_INT_DATE 2004"  
S5401000 "CV_AGE_INT_DATE 2005"  
S7501200 "CV_AGE_INT_DATE 2006"  
T0008500 "CV_AGE_INT_DATE 2007"  
T2011100 "CV_AGE_INT_DATE 2008"  
T3601500 "CV_AGE_INT_DATE 2009"  
T5201400 "CV_AGE_INT_DATE 2010"  
}  
  
(8984 observations read)  
  
.
```

Variables window

Variable	Label
R0000100	PUBID - YTH ID CODE 1997
R0536300	KEY!SEX (SYMBOL) 1997
R0536401	KEY!BDATE M/Y (SYMBOL) 1997
R0536402	KEY!BDATE M/Y (SYMBOL) 1997
R1194100	CV_AGE_INT_DATE 1997
R1235800	CV_SAMPLE_TYPE 1997
R1482600	KEY!RACE_ETHNICITY (SYMBOL) 1997
R2553500	CV_AGE_INT_DATE 1998
R3876300	CV_AGE_INT_DATE 1999
R5453700	CV_AGE_INT_DATE 2000
R7216000	CV_AGE_INT_DATE 2001
S1531400	CV_AGE_INT_DATE 2002
S2001000	CV_AGE_INT_DATE 2003
S3801100	CV_AGE_INT_DATE 2004
S5401000	CV_AGE_INT_DATE 2005
S7501200	CV_AGE_INT_DATE 2006

Properties window

Variables	
Name	
Label	
Type	
Format	
Value Label	
Notes	

Data	
Filename	
Label	
Notes	
Variables	20
Observations	8,984
Size	701.88K
Memory	32M

Output window

Opening Stata's do-file editor

1

Type `doedit` in the command window or click on the little 'notebook' icon.

```
. cd "H:\MyData\NLS"  
H:\MyData\NLS  
  
. infile using Age.dct  
  
infile dictionary {  
R0000100 "PUBID - YTH ID C  
R0536300 "KEY!SEX (SYMBOL)  
R0536401 "KEY!BDATE M/Y (S  
R0536402 "KEY!BDATE M/Y (S  
R1194100 "CV_AGE_INIT_DATE  
R1235800 "CV_SAMPLE_TYPE 1  
R1482600 "KEY!RACE_ETHNICI  
R2553500 "CV_AGE_INIT_DATE  
R3876300 "CV_AGE_INIT_DATE  
R5453700 "CV_AGE_INIT_DATE  
R7216000 "CV_AGE_INIT_DATE  
S1531400 "CV_AGE_INIT_DATE
```

2

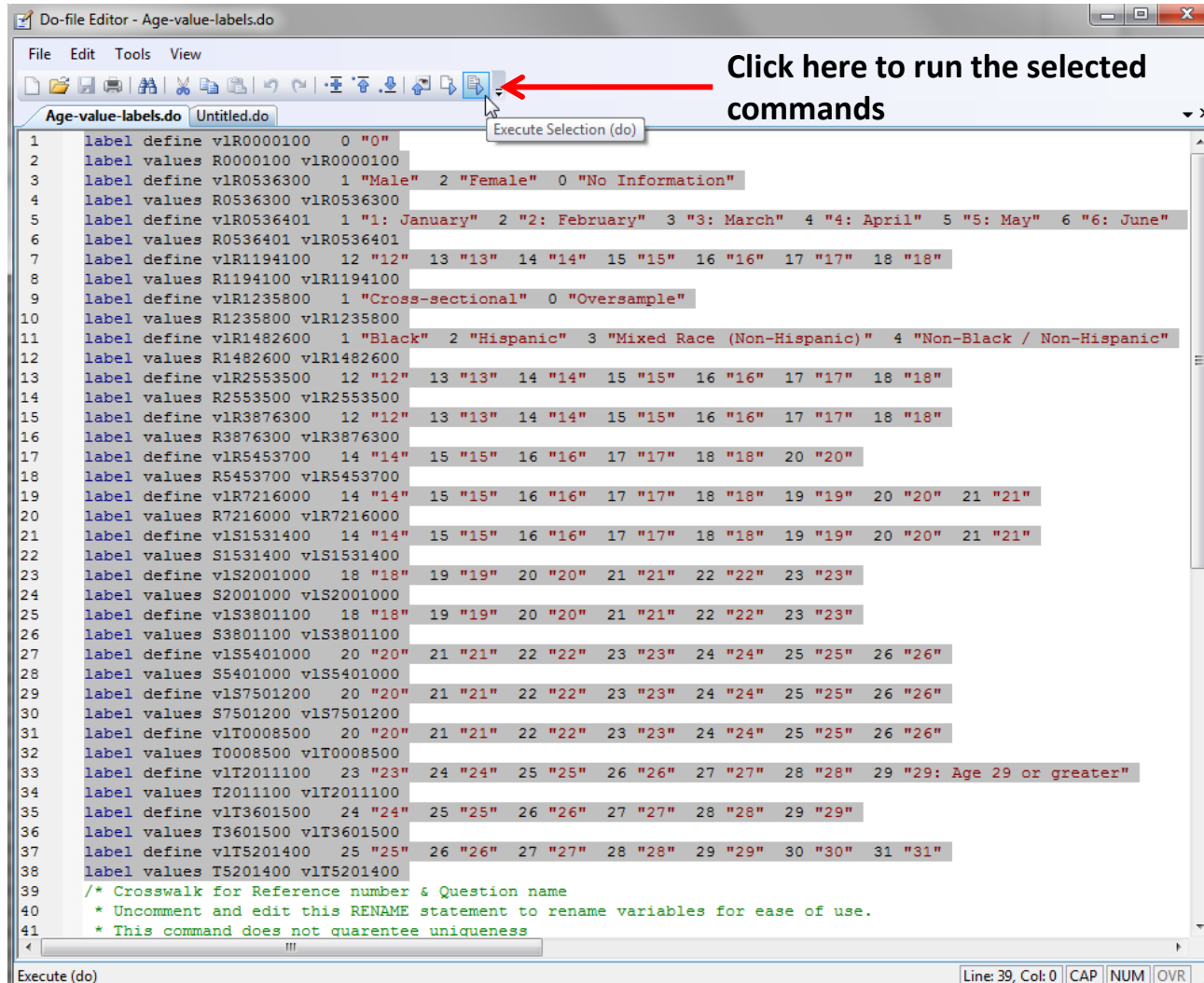
The do-file editor will pop-up. Go to File->Open and look for the do-file from the NLS download. In this example is called "Age-value-labels.do"

Name

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.



Do-file Editor - Age-value-labels.do

File Edit Tools View

Age-value-labels.do Untitled.do

Execute Selection (do)

Click here to run the selected commands

```
1 label define v1R0000100 0 "0"
2 label values R0000100 v1R0000100
3 label define v1R0536300 1 "Male" 2 "Female" 0 "No Information"
4 label values R0536300 v1R0536300
5 label define v1R0536401 1 "1: January" 2 "2: February" 3 "3: March" 4 "4: April" 5 "5: May" 6 "6: June"
6 label values R0536401 v1R0536401
7 label define v1R1194100 12 "12" 13 "13" 14 "14" 15 "15" 16 "16" 17 "17" 18 "18"
8 label values R1194100 v1R1194100
9 label define v1R1235800 1 "Cross-sectional" 0 "Oversample"
10 label values R1235800 v1R1235800
11 label define v1R1482600 1 "Black" 2 "Hispanic" 3 "Mixed Race (Non-Hispanic)" 4 "Non-Black / Non-Hispanic"
12 label values R1482600 v1R1482600
13 label define v1R2553500 12 "12" 13 "13" 14 "14" 15 "15" 16 "16" 17 "17" 18 "18"
14 label values R2553500 v1R2553500
15 label define v1R3876300 12 "12" 13 "13" 14 "14" 15 "15" 16 "16" 17 "17" 18 "18"
16 label values R3876300 v1R3876300
17 label define v1R5453700 14 "14" 15 "15" 16 "16" 17 "17" 18 "18" 20 "20"
18 label values R5453700 v1R5453700
19 label define v1R7216000 14 "14" 15 "15" 16 "16" 17 "17" 18 "18" 19 "19" 20 "20" 21 "21"
20 label values R7216000 v1R7216000
21 label define v1S1531400 14 "14" 15 "15" 16 "16" 17 "17" 18 "18" 19 "19" 20 "20" 21 "21"
22 label values S1531400 v1S1531400
23 label define v1S2001000 18 "18" 19 "19" 20 "20" 21 "21" 22 "22" 23 "23"
24 label values S2001000 v1S2001000
25 label define v1S3801100 18 "18" 19 "19" 20 "20" 21 "21" 22 "22" 23 "23"
26 label values S3801100 v1S3801100
27 label define v1S5401000 20 "20" 21 "21" 22 "22" 23 "23" 24 "24" 25 "25" 26 "26"
28 label values S5401000 v1S5401000
29 label define v1S7501200 20 "20" 21 "21" 22 "22" 23 "23" 24 "24" 25 "25" 26 "26"
30 label values S7501200 v1S7501200
31 label define v1T0008500 20 "20" 21 "21" 22 "22" 23 "23" 24 "24" 25 "25" 26 "26"
32 label values T0008500 v1T0008500
33 label define v1T2011100 23 "23" 24 "24" 25 "25" 26 "26" 27 "27" 28 "28" 29 "29: Age 29 or greater"
34 label values T2011100 v1T2011100
35 label define v1T3601500 24 "24" 25 "25" 26 "26" 27 "27" 28 "28" 29 "29"
36 label values T3601500 v1T3601500
37 label define v1T5201400 25 "25" 26 "26" 27 "27" 28 "28" 29 "29" 30 "30" 31 "31"
38 label values T5201400 v1T5201400
39 /* Crosswalk for Reference number & Question name
40 * Uncomment and edit this RENAME statement to rename variables for ease of use.
41 * This command does not guarantee uniqueness
```

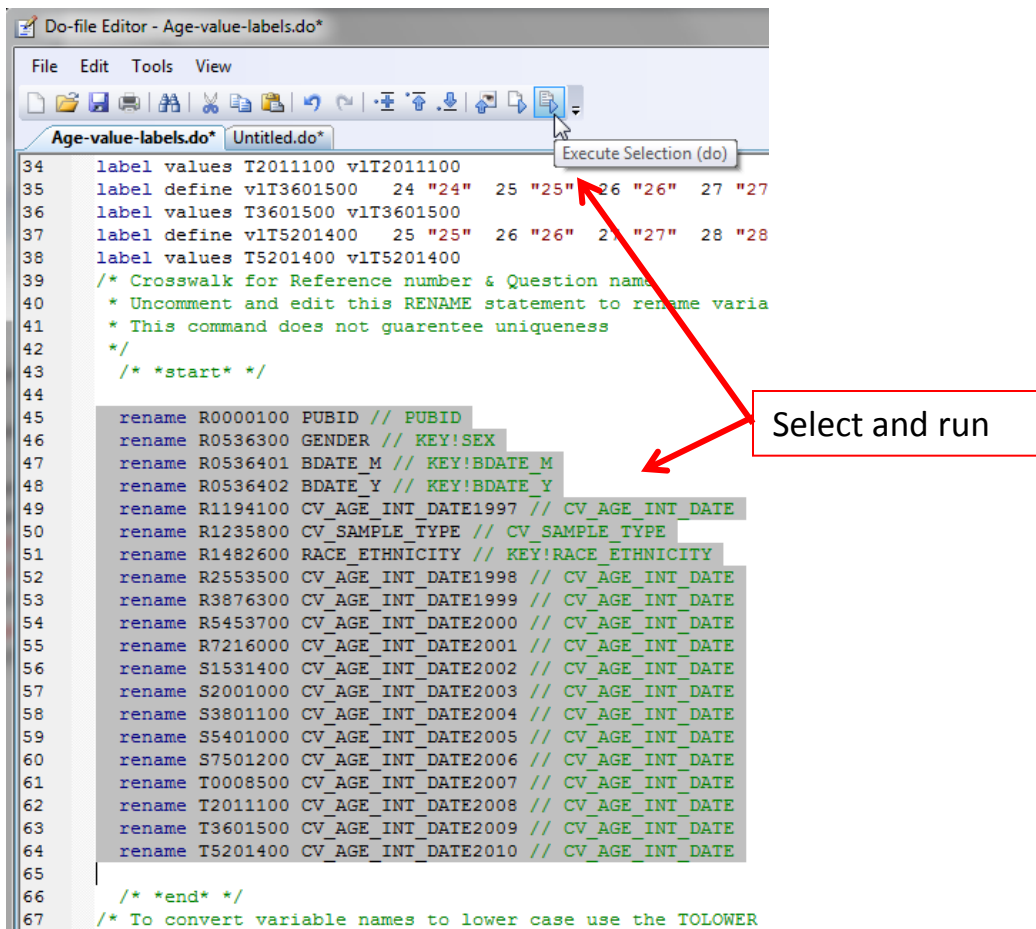
Execute (do) Line: 39, Col: 0 | CAP | NUM | OVR

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.

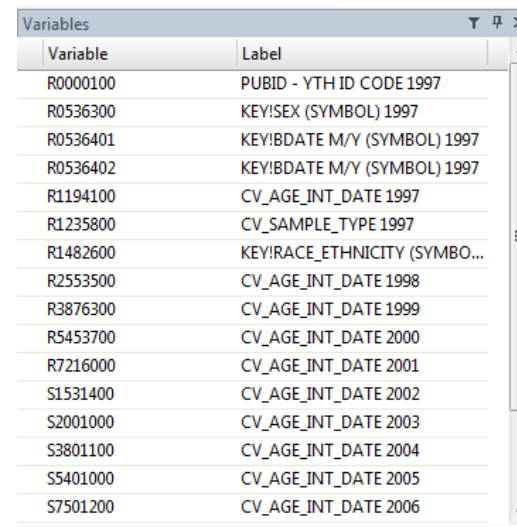
One important edit is adding the years to the names of the variables that change over time. In the example below, rows 49, 52-64 had originally the same name except that now each has its corresponding year, this makes them unique. Other minor edits were done in rows 46-48 and 51 (compare to the original)

2



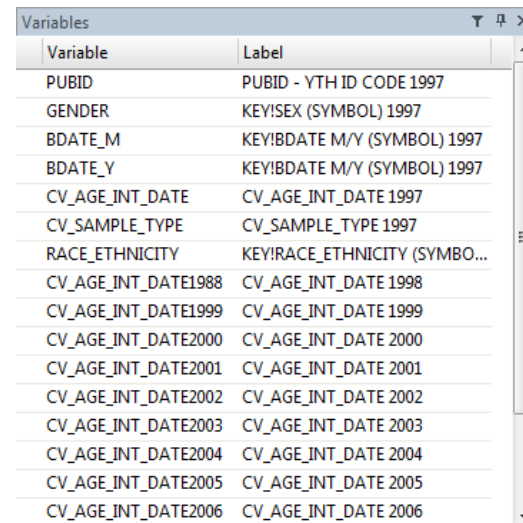
```
34 label values T2011100 v1T2011100
35 label define v1T3601500 24 "24" 25 "25" 26 "26" 27 "27"
36 label values T3601500 v1T3601500
37 label define v1T5201400 25 "25" 26 "26" 27 "27" 28 "28"
38 label values T5201400 v1T5201400
39 /* Crosswalk for Reference number & Question name
40 * Uncomment and edit this RENAME statement to rename varia
41 * This command does not guarantee uniqueness
42 */
43 /* *start* */
44
45 rename R0000100 PUBID // PUBID
46 rename R0536300 GENDER // KEY!SEX
47 rename R0536401 BDATE_M // KEY!BDATE_M
48 rename R0536402 BDATE_Y // KEY!BDATE_Y
49 rename R1194100 CV_AGE_INT_DATE1997 // CV_AGE_INT_DATE
50 rename R1235800 CV_SAMPLE_TYPE // CV_SAMPLE_TYPE
51 rename R1482600 RACE_ETHNICITY // KEY!RACE_ETHNICITY
52 rename R2553500 CV_AGE_INT_DATE1998 // CV_AGE_INT_DATE
53 rename R3876300 CV_AGE_INT_DATE1999 // CV_AGE_INT_DATE
54 rename R5453700 CV_AGE_INT_DATE2000 // CV_AGE_INT_DATE
55 rename R7216000 CV_AGE_INT_DATE2001 // CV_AGE_INT_DATE
56 rename S1531400 CV_AGE_INT_DATE2002 // CV_AGE_INT_DATE
57 rename S2001000 CV_AGE_INT_DATE2003 // CV_AGE_INT_DATE
58 rename S3801100 CV_AGE_INT_DATE2004 // CV_AGE_INT_DATE
59 rename S5401000 CV_AGE_INT_DATE2005 // CV_AGE_INT_DATE
60 rename S7501200 CV_AGE_INT_DATE2006 // CV_AGE_INT_DATE
61 rename T0008500 CV_AGE_INT_DATE2007 // CV_AGE_INT_DATE
62 rename T2011100 CV_AGE_INT_DATE2008 // CV_AGE_INT_DATE
63 rename T3601500 CV_AGE_INT_DATE2009 // CV_AGE_INT_DATE
64 rename T5201400 CV_AGE_INT_DATE2010 // CV_AGE_INT_DATE
65
66 /* *end* */
67 /* To convert variable names to lower case use the TOLOWER
```

1 Before



Variable	Label
R0000100	PUBID - YTH ID CODE 1997
R0536300	KEY!SEX (SYMBOL) 1997
R0536401	KEY!BDATE M/Y (SYMBOL) 1997
R0536402	KEY!BDATE M/Y (SYMBOL) 1997
R1194100	CV_AGE_INT_DATE 1997
R1235800	CV_SAMPLE_TYPE 1997
R1482600	KEY!RACE_ETHNICITY (SYMB...
R2553500	CV_AGE_INT_DATE 1998
R3876300	CV_AGE_INT_DATE 1999
R5453700	CV_AGE_INT_DATE 2000
R7216000	CV_AGE_INT_DATE 2001
S1531400	CV_AGE_INT_DATE 2002
S2001000	CV_AGE_INT_DATE 2003
S3801100	CV_AGE_INT_DATE 2004
S5401000	CV_AGE_INT_DATE 2005
S7501200	CV_AGE_INT_DATE 2006

3 After



Variable	Label
PUBID	PUBID - YTH ID CODE 1997
GENDER	KEY!SEX (SYMBOL) 1997
BDATE_M	KEY!BDATE M/Y (SYMBOL) 1997
BDATE_Y	KEY!BDATE M/Y (SYMBOL) 1997
CV_AGE_INT_DATE	CV_AGE_INT_DATE 1997
CV_SAMPLE_TYPE	CV_SAMPLE_TYPE 1997
RACE_ETHNICITY	KEY!RACE_ETHNICITY (SYMB...
CV_AGE_INT_DATE1988	CV_AGE_INT_DATE 1998
CV_AGE_INT_DATE1999	CV_AGE_INT_DATE 1999
CV_AGE_INT_DATE2000	CV_AGE_INT_DATE 2000
CV_AGE_INT_DATE2001	CV_AGE_INT_DATE 2001
CV_AGE_INT_DATE2002	CV_AGE_INT_DATE 2002
CV_AGE_INT_DATE2003	CV_AGE_INT_DATE 2003
CV_AGE_INT_DATE2004	CV_AGE_INT_DATE 2004
CV_AGE_INT_DATE2005	CV_AGE_INT_DATE 2005
CV_AGE_INT_DATE2006	CV_AGE_INT_DATE 2006

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.

The screenshot shows the Stata Data Editor (Browse) window. The main area displays a dataset with 26 rows and 11 variables. The variables are: PUBID, GENDER, BDATE_M, BDATE_Y, CV_AGE_INT, CV_SAMPLE, RACE_ETHNI, CV_AGE_1988, CV_AGE_1999, and two unlabeled variables. The data is as follows:

	PUBID	GENDER	BDATE_M	BDATE_Y	CV_AGE_INT	CV_SAMPLE	RACE_ETHNI	CV_AGE_1988	CV_AGE_1999	
1	1	2	9	1981	15	1	4	17	18	
2	2	1	7	1982	14	1	2	16	17	
3	3	2	9	1983	13	1	2	15	16	
4	4	2	2	1981	15	1	2	17	18	
5	5	1	10	1982	15	1	2	16	17	
6	6	2	1	1982	15	1	2	16	17	
7	7	1	4	1983	14	1	2	15	16	
8	8	2	6	1981	16	1	4	17	18	
9	9	1	10	1982	15	1	4	16	17	
10	10	1	3	1984	14	1	4	14	15	
11	11	2	6	1982	15	1	2	16	17	
12	12	1	10	1981	15	1	2	17	18	
13	13	1	11	1984	12	1	2	13	15	
14	14	1	7	1980	17	1	2	18	-5	
15	15	2	1	1983	15	1	2	15	17	
16	16	1	2	1982	15	1	2	16	17	
17	17	2	11	1981	15	1	2	17	18	
18	18	1	2	1982	15	1	1	16	17	
19	19	1	4	1984	12	1	1	14	15	
20	20	1	12	1980	16	1	1	17	19	
21	21	1	8	1982	14	1	2	16	17	
22	22	1	6	1982	14	1	2	16	17	
23	23	2	1	1983	14	1	2	15	16	
24	24	1	6	1984	12	1	2	14	15	
25	25	2	3	1983	13	1	2	15	16	
26	26	1	10	1980	16	1	1	18	19	

The right-hand panel shows the 'Variables' section with a list of variables and their labels. The 'Properties' section shows the properties for the selected variable 'PUBID': Name: PUBID, Label: PUBID - YTH, Type: float, Format: %9.0g, Value Label: , Notes: . The 'Data' section shows the filename, label, notes, variables (20), and observations (8,984).

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)
```

```
. reshape long CV_AGE_INT_DATE, i(PUBID) j(YEAR)
(note: j = 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010)

Data                wide  ->  long
-----
Number of obs.      8984  -> 125776
Number of variables    20  ->    8
j variable (14 values)    ->  YEAR
xij variables:
CV_AGE_INT_DATE1997 CV_AGE_INT_DATE1998 ... CV_AGE_INT_DATE2010->CV_AGE_INT_DATE
```

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>

The screenshot shows the Stata Data Editor (Browse) window. The main window displays a dataset with 26 observations and 8 variables. The variables are: PUBID, YEAR, GENDER, BDATE_M, BDATE_Y, CV_SAMPLE_~E, RACE_ETHNI~Y, and CV_AGE_INT~E. The data is reshaped with years in rows. The Variables panel on the right shows the list of variables and their properties.

Obs	PUBID	YEAR	GENDER	BDATE_M	BDATE_Y	CV_SAMPLE_~E	RACE_ETHNI~Y	CV_AGE_INT~E
1	1	1997	2	9	1981	1	4	15
2	1	1998	2	9	1981	1	4	17
3	1	1999	2	9	1981	1	4	18
4	1	2000	2	9	1981	1	4	19
5	1	2001	2	9	1981	1	4	20
6	1	2002	2	9	1981	1	4	21
7	1	2003	2	9	1981	1	4	22
8	1	2004	2	9	1981	1	4	23
9	1	2005	2	9	1981	1	4	24
10	1	2006	2	9	1981	1	4	25
11	1	2007	2	9	1981	1	4	26
12	1	2008	2	9	1981	1	4	27
13	1	2009	2	9	1981	1	4	28
14	1	2010	2	9	1981	1	4	29
15	2	1997	1	7	1982	1	2	14
16	2	1998	1	7	1982	1	2	16
17	2	1999	1	7	1982	1	2	17
18	2	2000	1	7	1982	1	2	18
19	2	2001	1	7	1982	1	2	19
20	2	2002	1	7	1982	1	2	20
21	2	2003	1	7	1982	1	2	21
22	2	2004	1	7	1982	1	2	22
23	2	2005	1	7	1982	1	2	23
24	2	2006	1	7	1982	1	2	-5
25	2	2007	1	7	1982	1	2	-5
26	2	2008	1	7	1982	1	2	26

Variables panel:

- Variable: PUBID, Label: PUBID - YTH
- Variable: YEAR
- Variable: GENDER, Key: SEX (SYM)
- Variable: BDATE_M, Key: BDATE M
- Variable: BDATE_Y, Key: BDATE M
- Variable: CV_SAMPLE_..., Key: CV_SAMPLE_...
- Variable: RACE_ETHNI..., Key: RACE_ET
- Variable: CV_AGE_INT_...

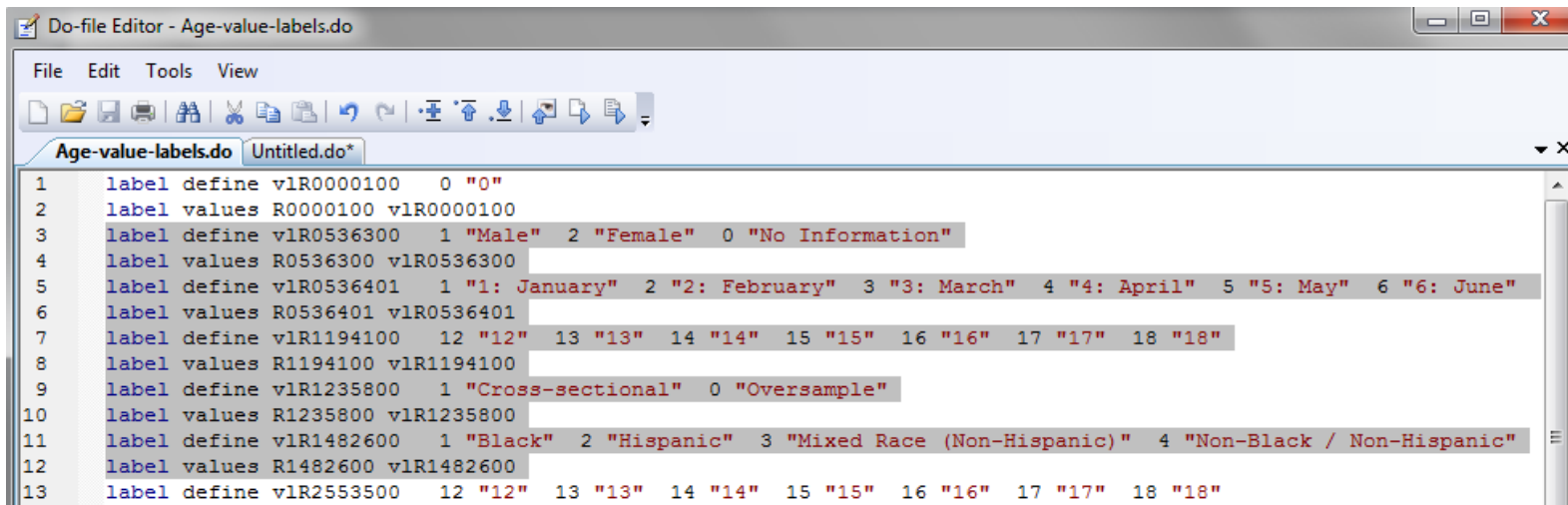
Properties panel:

- Variables: Name, Label, Type, Format, Value Label, Notes
- Data: Filename, Label, Notes, Variables: 8, Observations: 125,776 2 1

Ready Vars: 8 Order: Dataset Obs: 125,776 Filter: Off Mode: Browse CAP NUM

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)



```
Do-file Editor - Age-value-labels.do
File Edit Tools View
Age-value-labels.do Untitled.do*
1 label define v1R0000100 0 "0"
2 label values R0000100 v1R0000100
3 label define v1R0536300 1 "Male" 2 "Female" 0 "No Information"
4 label values R0536300 v1R0536300
5 label define v1R0536401 1 "1: January" 2 "2: February" 3 "3: March" 4 "4: April" 5 "5: May" 6 "6: June"
6 label values R0536401 v1R0536401
7 label define v1R1194100 12 "12" 13 "13" 14 "14" 15 "15" 16 "16" 17 "17" 18 "18"
8 label values R1194100 v1R1194100
9 label define v1R1235800 1 "Cross-sectional" 0 "Oversample"
10 label values R1235800 v1R1235800
11 label define v1R1482600 1 "Black" 2 "Hispanic" 3 "Mixed Race (Non-Hispanic)" 4 "Non-Black / Non-Hispanic"
12 label values R1482600 v1R1482600
13 label define v1R2553500 12 "12" 13 "13" 14 "14" 15 "15" 16 "16" 17 "17" 18 "18"
```

Adding value labels (...again, part 2)

Notice that the lines starting with "label values..." have the variables' old names.

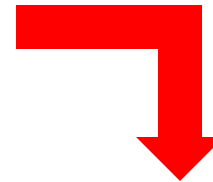
To match the old and new names, look at the 'rename' section, rows 46-48, 50, 51.

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

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



After

```
. tab GENDER
```

KEY:SEX (SYMBOL) 1997	Freq.	Percent	Cum.
Male	64,386	51.19	51.19
Female	61,390	48.81	100.00
Total	125,776	100.00	



```
label define v1R0536300 1 "Male" 2 "Female" 0 "No Information"
label values GENDER v1R0536300
label define v1R0536401 1 "1: January" 2 "2: February"
label values BDATE_M v1R0536401
label define v1R1194100 12 "12" 13 "13" 14 "14" 15 "15"
label values BDATE_Y v1R1194100
label define v1R1235800 1 "Cross-sectional" 0 "Oversampled"
label values CV_SAMPLE_TYPE v1R1235800
label define v1R1482600 1 "Black" 2 "Hispanic" 3 "Mixed"
label values RACE_ETHNICITY v1R1482600
```

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/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>

How to Use the NLS Investigator

- https://www.nlsinfo.org/InvestigatorGuide/investigator_guide_TOC.html