

SESSION 1: DESCRIBING DATA AND DATA MANIPULATION

Manuel V. Montesinos

Computation Brush-Up Course
Competition and EPP Master Programs

Fall 2021



NLSY Dataset

The dataset used in this session is a subsample of the **National Longitudinal Survey of Youth (NLSY)**.

More data and explanations [here](#).

Home

F.A.Q.s

Site Map

Bibliography

Contact Us

National Longitudinal Surveys

A Program of the U.S. Bureau of Labor Statistics

search site

Advanced Search

Getting Started

Cohorts

Access Data / Investigator

Suggest Questions for Future NLSY Surveys

About the National Longitudinal Surveys (NLS)

The NLS, sponsored by the U.S. Bureau of Labor Statistics, are nationally representative surveys that follow the same sample of individuals from specific birth cohorts over time. The surveys collect data on labor market activity, schooling, fertility, program participation, health, and much, much more. Choose a cohort below to learn more.

NLS Youth 1987 (NLSY97)

Men and women born in the years 1980-84

NLS Youth 1979 (NLSY79)

Men and women born in the years 1957-64

NLSY Child and Young Adult (NLSYCA)

Biological children of women in the NLSY79

NLS Mature and Young Women (NLSMW)

Mature women born in the years 1922-37 and young women born in the years 1943-53.

NLS Older and Young Men (NLSOM)

Older men born in the years 1909-21 and young men born in the years 1941-52.

Who Are You?

Please indicate your NLS experience so we can direct you to a page that focuses on your needs:

✖ New NLS User

✖ Experienced NLS User

✖ Press, Policymakers and Other Interested Parties

Need Help Using NLS Data?

Explore the Tutorials

Learn how to approach research projects, search for and extract data, and then program with NLS data.

Introduction to the Investigator

Learn how to use the online NLS data search and extraction site to create your own data sets.

NLS Learning Opportunities: Booth Schedule at Conferences

Going to a professional conference soon? The NLS User Services booth makes an appearance at several conferences throughout the year. Stop by to get some hands-on instruction (or just to say hi).

Latest Information from the NLS

RE-RELEASE OF THE NLSY97 DATA. We have re-released the NLSY97 public-use Rounds 1-18 data with some updates. A comprehensive review of the created wage variables, `CV_HIRLY_COMPENSATION` and `CV_HIRLY_PAY` led to updates to a number of these created variables and select underlying raw data through all rounds. See the [Errata](#) for details. The re-release also includes a small number of corrections to the created incarceration variables and to invalid dates for teenage job start dates; these are also detailed on the errata page. The dataset can be accessed at [NLS Investigator](#).

NEW RELEASE DECEMBER 6, 2019: NLSY97 Dataset. Includes data from 1967 through 2017 (Rounds 1-18). To access public data, use NLS Investigator ([www.nlsinfo.org/nlsinvestigator](#)).

NLSY97 TIMINGS DATASET NOW AVAILABLE. A new separate dataset of variable timings has been released for the NLSY97. The timings, which measure the time of the respondent's interaction on a particular question during the interview, are mainly for questions related to sexual activity, pregnancy, smoking, drug use, delinquent behaviors, criminal activity, and expectations. To access the dataset, go to [www.nlsinfo.org/nlsinvestigator](#) and choose "NLSY97" in the cohort selection list, then choose "Select Timings." More details can be found at [https://www.nlsinfo.org/content/cohortsnlsy97timing-and-understanding-the-dataset](#).

National Longitudinal Surveys | Bureau of Labor Statistics

Postal Square Building | 2 Massachusetts Ave., NE
Washington, DC 20212 | [enr339@bls.gov](#)
TEL 1 (202) 861-7410 TDD 1 (800) 877-4339

NLS User Services
[nlsuser@bls.gov](#) | 1 (814) 442-7555
[Program Use Notations](#) | [Browse Data](#) | [Site Map](#)
[Problems with the Site?](#)

Navigation icons and page number 2

Describing Data

To get a **general description** of the dataset in memory and the format of each variable, type **describe**.

Contains data from data1.dta				
obs:	540	Random subsample of data from NLSY79		
vars:	26	30 Aug 2017 19:21		
variable name	storage type	display format	value label	variable label
id	int	%8.0g		respondent's identifier
female	byte	%8.0g		
date_birth	int	%d		
ethnicity	str8	%9s		
ethhisp	byte	%8.0g	yn	hispanic
ethblack	byte	%8.0g		black
ethwhite	byte	%8.0g		white
age	byte	%8.0g		age in 2002
height	byte	%8.0g		height in 1985 (in inches)
weight85	int	%8.0g		weight in 1985 (in pounds)
weight02	int	%8.0g		weight in 2002 (in pounds)
sm	byte	%8.0g	yn	year of schooling of respondent's mother
sf	byte	%8.0g		year of schooling of respondent's father
pov78	byte	%8.0g		living in poverty in 1978
s	byte	%8.0g		years of schooling (as of 2002)
asvab_mean	float	%9.0g		aptitude test, composite score
asvab02	byte	%8.0g		arithmetic reasoning
asvab03	byte	%8.0g		word knowledge
asvab04	byte	%8.0g		paragraph comprehension
earnings	double	%9.0g		hourly earnings in 2002 (in \$)
hours	byte	%8.0g		number of hours worked per week in 2002

Note that **describe using filename** describes a stored Stata format dataset. You can also **describe a subset of a dataset** by specifying the variables you are interested in: **describe varlist**.

Variables

Names: variable names can be 1-32 characters long and can be formed only by letters, numbers, and underscore.

Types: there are two types of variables in Stata, *numeric* and *string*. A third type, *date*, is a special type of numeric. Numeric variables contain numbers. String variables contain text which can contain any character on the keyboard (letters, numbers, and special characters). Numeric calculations and statistical analysis can be done on numeric variables but not on string variables.

Numeric: there are five numeric types for storing variables, three of them are integer types (*byte*, *int*, *long*), and two of them floating point (*double* and *float*). For further info, type `help data types`.

Variables

Stata has a **color-coded system** for each type:

- ▶ **Black** is for numbers.
- ▶ **Red** is for text or strings.
- ▶ **Blue** is for labeled variables (a short description is inputted to each value).

Missing values are represented by a dot (“.”).

When **string variables have missing values**, these are blank, and are represented in commands by two double quotes with nothing in between (“”).

female	date_birth	ethnicity	ethhisp	ethblack
0	05sep1961	white	0	no
0	20mar1958	white	0	no
1	17mar1962	white	0	no
0	19oct1959	white	0	no
1	21jul1958	white	0	no
0	04sep1964	white	0	no
0	19jun1959	black	0	yes
0	02may1963	black	0	yes

Summary Statistics

To calculate **summary statistics**, such as means, standard deviations, and so on, use the command `summarize`.

You can also use `summarize, detail`. This will also report percentiles, the variance, measures of skewness and kurtosis.

age in 2002				
	Percentiles	Smallest		
1%	37	37		
5%	37	37		
10%	38	37	Obs	540
25%	39	37	Sum of Wgt.	540
50%	40		Mean	40.68519
		Largest	Std. Dev.	2.2317
75%	42	45		
90%	44	45	Variance	4.980485
95%	44	45	Skewness	.1884904
99%	45	45	Kurtosis	2.003487

To calculate **correlation** or **covariance** matrices, use the command `correlate`.

Frequencies

Use the command `tabulate varname` to get the absolute frequency (*Freq.*), the relative frequency for each value (*Percent*) and the cumulative frequency (*Cum.*).

`table varname, contents()` allows you to choose the contents of the table. You can select up to five statistics (e.g. `table female, c(freq mean age mean s)`). If table cell contents are not specified, Stata provides a raw count of each value by typing `table varname`.

Using `bysort` as a prefix, you can also generate threeway crosstabs. Try: `bysort female: tab ethblack ethwhite, column row`.

For continuous data, try `tabstat`. For instance, `tabstat age s, statistics(mean median sd var count range min max)`.

Questions

1. Is respondents' education positively correlated with their parents' education?
2. Compute average test scores by ethnicity and gender.
3. Are respondents who were living in poverty in 1978 more likely to earn less in 2002?

Data Manipulation

Sorting: if you have to sort your data observations (rows), type `sort varname`. This puts the observations in ascending order. To put them in descending order, use `gsort -varname`.

Ordering: you can re-order the variables (columns) by typing `order varlist`, writing the variables in `varlist` in the desired order.

Renaming: to change the name of a variable, type `rename varnameold varnamenew`.

Labeling: add labels to the values. For example, `label define sex 0 "male" 1 "female"`, and then `label values female sex`, or a short description of the variable, such as `label variable sex "Sex (0 if male, 1 if female)"`. If you type `labelbook`, Stata will list all the labels in the data.

Data Manipulation

Table 1: Operators in Stata

Operator	Meaning	Operator	Meaning
==	equal to	&	and
>	greater than		or
=>	at least as big as	~	is not
<=	at most as large as	!=	not equal to
<	less than	~=	not equal to

Drop and keep: it works with both subsets of observations and variables.

Data Manipulation

Generating new variables: generate vs. egen

- ▶ `generate` create unique values for each observation (e.g. `gen LS = log(s)`).
- ▶ `egen` can create summary statistics (e.g. `egen MEDS = median(s)`).

Other examples:

```
gen s2 = s^2
gen x = 1
gen name = "A"
gen surname = "B"
gen fullname = name+" "+surname
gen abb = "Mr." if female==0
gen id = _n
```

Questions

1. Are male respondents more likely to work full-time (i.e. 40 hours or more)?
2. In the data, non-white respondents have lower earnings on average. Find evidence suggesting that less favorable socioeconomic conditions when they were children may be one of the reasons.
3. Explore the relationship between change in weight (from 1985 to 2002) and wages earned in 2002. Is there any difference by gender?