

Title

data management — Introduction to data-management commands

Description

This manual, called [D], documents Stata's data-management features. See Mitchell (2010) for additional information and examples on data management in Stata.

Data management for statistical applications refers not only to classical data management—sorting, merging, appending, and the like—but also to data reorganization because the statistical routines you will use assume that the data are organized in a certain way. For example, statistical commands that analyze longitudinal data, such as `xtreg`, generally require that the data be in long rather than wide form, meaning that repeated values are recorded not as extra variables, but as extra observations.

Here are the basics everyone should know:

[D] use	Load Stata dataset
[D] save	Save Stata dataset
[D] describe	Describe data in memory or in file
[D] codebook	Describe data contents
[D] inspect	Display simple summary of data's attributes
[D] count	Count observations satisfying specified conditions
[D] data types	Quick reference for data types
[D] missing values	Quick reference for missing values
[D] datetime	Date and time values and variables
[D] list	List values of variables
[D] edit	Browse or edit data with Data Editor
[D] varmanage	Manage variable labels, formats, and other properties
[D] rename	Rename variable
[D] format	Set variables' output format
[D] label	Manipulate labels

You will need to create and drop variables, and here is how:

[D] generate	Create or change contents of variable
[D] functions	Functions
[D] egen	Extensions to generate
[D] drop	Eliminate variables or observations
[D] clear	Clear memory

For inputting or importing data, see

[D] use	Load Stata dataset
[D] sysuse	Use shipped dataset
[D] webuse	Use dataset from Stata website
[D] input	Enter data from keyboard
[D] insheet	Read text data created by a spreadsheet
[D] import	Overview of importing data into Stata
[D] import excel	Import and export Excel files
[D] import sasxport	Import and export datasets in SAS XPORT format
[D] infile (fixed format)	Read text data in fixed format with a dictionary
[D] infile (free format)	Read unformatted text data
[D] infix (fixed format)	Read text data in fixed format
[D] odbc	Load, write, or view data from ODBC sources
[D] xmlsave	Export or import dataset in XML format
[D] hexdump	Display hexadecimal report on file
[D] icd9	ICD-9-CM diagnostic and procedure codes

and for exporting data, see

[D] save	Save Stata dataset
[D] export	Overview of exporting data from Stata
[D] outfile	Export dataset in text format
[D] outsheet	Write spreadsheet-style dataset
[D] import excel	Import and export Excel files
[D] import sasxport	Import and export datasets in SAS XPORT format
[D] odbc	Load, write, or view data from ODBC sources

The ordering of variables and observations (sort order) can be important; see

[D] order	Reorder variables in dataset
[D] sort	Sort data
[D] gsort	Ascending and descending sort

To reorganize or combine data, see

[D] merge	Merge datasets
[D] append	Append datasets
[D] reshape	Convert data from wide to long form and vice versa
[D] collapse	Make dataset of summary statistics
[D] contract	Make dataset of frequencies and percentages
[D] fillin	Rectangularize dataset
[D] expand	Duplicate observations
[D] expandcl	Duplicate clustered observations
[D] stack	Stack data
[D] joinby	Form all pairwise combinations within groups
[D] xpose	Interchange observations and variables
[D] cross	Form every pairwise combination of two datasets

In the above list, we particularly want to direct your attention to [D] **reshape**, a useful command that beginners often overlook.

For random sampling, see

[D] sample	Draw random sample
[D] drawnorm	Draw sample from multivariate normal distribution

For file manipulation, see

[D] type	Display contents of a file
[D] erase	Erase a disk file
[D] copy	Copy file from disk or URL
[D] cd	Change directory
[D] dir	Display filenames
[D] mkdir	Create directory
[D] rmdir	Remove directory
[D] cf	Compare two datasets
[D] changeool	Convert end-of-line characters of text file
[D] filefilter	Convert text or binary patterns in a file
[D] checksum	Calculate checksum of file
[D] zipfile	Compress and uncompress files and directories in zip archive format

The entries above are important. The rest are useful when you need them:

[D] datasignature	Determine whether data have changed
[D] type	Display contents of a file
[D] notes	Place notes in data
[D] label language	Labels for variables and values in multiple languages
[D] labelbook	Label utilities
[D] encode	Encode string into numeric and vice versa
[D] recode	Recode categorical variables
[D] ipolate	Linearly interpolate (extrapolate) values
[D] destring	Convert string variables to numeric variables and vice versa
[D] mvencode	Change missing values to numeric values and vice versa
[D] pctile	Create variable containing percentiles
[D] range	Generate numerical range
[D] by	Repeat Stata command on subsets of the data
[D] statsby	Collect statistics for a command across a by list
[D] compress	Compress data in memory
[D] recast	Change storage type of variable
[D] datetime display formats	Display formats for dates and times
[D] datetime translation	String to numeric date translation functions
[D] bcal	Business calendar file manipulation
[D] datetime business calendars	Business calendars
[D] datetime business calendars creation	Business calendars creation

[D] assert	Verify truth of claim
[D] clonevar	Clone existing variable
[D] compare	Compare two variables
[D] corr2data	Create dataset with specified correlation structure
[D] ds	List variables matching name patterns or other characteristics
[D] duplicates	Report, tag, or drop duplicate observations
[D] isid	Check for unique identifiers
[D] lookfor	Search for string in variable names and labels
[D] memory	Memory management
[D] putmata	Put Stata variables into Mata and vice versa
[D] obs	Increase the number of observations in a dataset
[D] rename group	Rename groups of variables
[D] separate	Create separate variables
[D] shell	Temporarily invoke operating system
[D] snapshot	Save and restore data snapshots
[D] split	Split string variables into parts

There are some real jewels in the above, such as [D] **notes**, [D] **compress**, and [D] **assert**, which you will find particularly useful.

Reference

Mitchell, M. N. 2010. *Data Management Using Stata: A Practical Handbook*. College Station, TX: Stata Press.

Also see

[D] **intro** — Introduction to data-management reference manual

[R] **intro** — Introduction to base reference manual