

Data management — Introduction to data management commands[Description](#) [References](#) [Also see](#)

Description

This manual, called [D], documents Stata's data management features. See [Mitchell \(2020\)](#) 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] sysuse	Use shipped dataset
[D] webuse	Use dataset from Stata website
[D] save	Save Stata dataset
[D] describe	Describe data in memory or in a 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
[D] frames intro	Introduction to frames

To work with multiple datasets in memory, see

[D] frames intro	Introduction to frames
[D] frames	Data frames
[D] frame change	Change identity of current (working) frame
[D] frame copy	Make a copy of a frame
[D] frame create	Create a new frame
[D] frame drop	Drop frames from memory
[D] frame prefix	The frame prefix command
[D] frame put	Copy selected variables or observations to a new frame
[D] frame pwf	Display name of current (working) frame
[D] frame rename	Rename existing frame
[D] frames dir	Display names of all frames in memory
[D] frames reset	Drop all frames from memory
[D] frames save	Save a set of frames on disk
[D] frames use	Load a set of frames from disk
[D] frames describe	Describe frames in memory or in a file
[D] frget	Copy variables from linked frame
[D] frlink	Link frames

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

[D] generate	Create or change contents of variable
[D] egen	Extensions to generate
[D] drop	Drop 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] import	Overview of importing data into Stata
[D] import dbase	Import and export dBase files
[D] import delimited	Import and export delimited text data
[D] import excel	Import and export Excel files
[D] import fred	Import data from Federal Reserve Economic Data
[D] import haver	Import data from Haver Analytics databases
[D] import sas	Import SAS files
[D] import sasxport5	Import and export data in SAS XPORT Version 5 format
[D] import sasxport8	Import and export data in SAS XPORT Version 8 format
[D] import spss	Import and export SPSS files
[D] infile (fixed format)	Import text data in fixed format with a dictionary
[D] infile (free format)	Import unformatted text data
[D] infix (fixed format)	Import text data in fixed format
[D] jdbc	Load, write, or view data from a database with a Java API
[D] odbc	Load, write, or view data from ODBC sources
[D] hexdump	Display hexadecimal report on file
[D] icd9	ICD-9-CM diagnosis codes
[D] icd9p	ICD-9-CM procedure codes
[D] icd10	ICD-10 diagnosis codes
[D] icd10cm	ICD-10-CM diagnosis codes
[D] icd10pcs	ICD-10-PCS 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] import dbase	Import and export dBase files
[D] import delimited	Import and export delimited text data
[D] import excel	Import and export Excel files
[D] import sasxport5	Import and export data in SAS XPORT Version 5 format
[D] import sasxport8	Import and export data in SAS XPORT Version 8 format
[D] import spss	Import and export SPSS files
[D] jdbc	Load, write, or view data from a database with a Java API
[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] append	Append datasets
[D] merge	Merge datasets
[D] frlink	Link frames
[D] frget	Copy variables from linked frame
[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] splitsample	Split data into random samples
[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 ASCII or binary patterns in a file
[D] checksum	Calculate checksum of file
[D] zipfile	Compress and uncompress files and directories in zip archive format

For handling Unicode strings, see

[D] unicode	Unicode utilities
[D] unicode translate	Translate files to Unicode
[D] unicode encoding	Unicode encoding utilities
[D] unicode locale	Unicode locale utilities
[D] unicode collator	Language-specific Unicode collators
[D] unicode convertfile	Low-level file conversion between encoding

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] dyngen	Dynamically generate new values of variables
[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 conversion	String to numeric date conversion functions
[D] Datetime durations	Obtaining and working with durations
[D] Datetime relative dates	Datetime relative dates
[D] Datetime values from other software	Date and time conversion from other software
[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] assertnested	Verify variables nested
[D] clonevar	Clone existing variable
[D] compare	Compare two variables
[D] corr2data	Create dataset with specified correlation structure
[D] ds	Compactly list variables with specified properties
[D] duplicates	Report, tag, or drop duplicate observations
[D] insobs	Add or insert 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
[D] vl	Manage variable lists
[D] vl create	Create and modify user-defined variable lists
[D] vl drop	Drop variable lists or variables from variable lists
[D] vl list	List contents of variable lists
[D] vl rebuild	Rebuild variable lists
[D] vl set	Set system-defined variable lists

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

References

- Hoffmann, J. P. 2017. *Principles of Data Management and Presentation*. Oakland, CA: University of California Press.
- Mitchell, M. N. 2020. *Data Management Using Stata: A Practical Handbook*. 2nd ed. College Station, TX: Stata Press.

Also see

- [D] **Intro** — Introduction to data management reference manual
- [R] **Intro** — Introduction to base reference manual

Stata, Stata Press, and Mata are registered trademarks of StataCorp LLC. Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations. Other brand and product names are registered trademarks or trademarks of their respective companies. Copyright © 1985–2023 StataCorp LLC, College Station, TX, USA. All rights reserved.

