

## NAME

sas2stata - Convert SAS datasets into Stata datasets.

## SYNOPSIS

sas2stata [-options] ssdname ...

## DESCRIPTION

Sas2stata converts one or more SAS datasets into a Stata file. If a sas file is specified with an extension of .ssd01, then the V6 engine is used to read the file. If it is specified with an extension of sas7bdat, then the V8 engine is used to read the file.

If no extension is used, and the -sasv option is specified, sas2stata will look for a file with the appropriate extension, e.g., if -sasv=6 it will look for an ssd01 file.

If no extension is used and -sasv is not specified, sas2stata uses whichever file is found, e.g., if an ssd01 file exists, it assumes V6. If both versions are found with the same first name, then an error message is displayed and no conversion is done.

The output data set will have the same name, but with extension .dta. It is created in the current working directory, even though the SAS data set may be located elsewhere. If the SAS data set (version 7 or above) contains long variable names (greater than 8 characters), sas2stata will keep the long names unless the -vn8 option is used. If using version 6 or less of Stata, the -vn8 option should be specified to truncate long variable names to 8 characters.

By default, sas2stata attempts the creation of a .dta file containing all observations and every variable in the .sas7bdat or .ssd01 file. The procedure is as follows: (1) sas2stata creates a SAS program, (2) which is executed to 'put' SAS variables into ascii format and determine the optimal storage type in Stata; (3) sas2stata creates a Stata 'dictionary' file to (4) read in the ascii data; and (5) Stata saves the data into Stata's internal (.dta) format.

The resulting Stata data set will contain variable names in lowercase; variable labels maintain their case. If the SAS data set is sorted, Stata will also sort before saving.

## OPTIONS

-s/-silent

be silent; in this case, sas2stata does not print any output to the screen, except for error messages. By default, sas2stata tells what stage of the conversion process is currently being executed, and it reports number of variables, number of observations, and more.

-a/-ascii

create ascii data only: by default, sas2stata executes all five steps outlined above. Option -a aborts this process after step (3). The user then needs to read in

- the data ``manually'` -- sas2stata writes instructions on how to do this to standard output.
- `-abc` creates the Stata data set with variables in alphabetical (ascii) order. Variables whose names start with an underscore (``_'`) appear first. By default, sas2stata maintains the variable order of the SAS data set. The ``-abc'` option is often convenient if one wishes to list (array) variables that have the same stem, but are not contiguous in the SAS data set.
- `-f/-float`  
prevents the use of Stata's variable type ``double'`. All variables whose SAS precision would require Stata's ``double'` type are created as ``float'`. This option may lead to a loss of precision, but saves space: a ``float'` takes four, a ``double'` eight bytes. See option ``-q'` for details on storage types.
- `-q/-quick`  
prevents sas2stata from determining the optimal variable type in Stata. Stata is capable of storing numerical variables in ``byte'`, ``int'`, ``long'`, ``float'` or ``double'` format, taking 1, 2, 4, 4, and 8 bytes, respectively. The first three types are for integers (``byte'` for -127 to +101; ``int'` for -32767 to +32740; ``long'` for -2147483647 to +2147483620), the other two for floating point variables. By default, SAS stores numerical variables in eight bytes, i.e., like Stata's ``double'` type. SAS users may reduce the required storage by using the ``length'` statement, which truncates the mantisse and/or exponent of the internal representation. By default, sas2stata stores numerical variables with `length<=4` as ``float'` and otherwise as ``double'`, and then figures out whether they may be compressed into one of the integer types. The latter feature often leads to dramatic storage space reductions. The ``-q'` option prevents sas2stata from performing this space optimization, which slightly (<25%) increases its speed of operation. (Note: users who do not wish variables to be stored in integer format, for example because they want to append data containing fractional values, may be better off using Stata's ``recast'` command.)
- `-r/-replace`  
By default, sas2stata warns the user if its output data set already exists, and asks permission to overwrite it. Option ``-replace'` suppresses this interactive behavior and replaces any existing output data set without warning.
- `-m/-messy`  
sas2stata creates many intermediary files during its operation. Option ``-m'` prevents sas2stata from cleaning up after it has finished. This option is mostly

useful for debugging purposes, to find out where something went wrong. Most intermediary files have a name starting with the process ID.

**-obs=n**

converts only the first n observations. By default, sas2stata converts all observations of the SAS data set.

**-varfile=filename**

may be used to select only a subset of variables to be included in the Stata data set. This will speed up the conversion process, and is also useful in situations where the number of variables is too large (generally speaking, more than 2047), or if truncating variable names to 8 characters will result in duplicate variable names. The filename is the name of a file whose contents are variable names only. These variable names are case-insensitive. Multiple variables with the same stem may be specified as ranges according to general SAS rules. For example, var1-var20 is understood as var1, var2, ..., var20.

**-sasv=n**

specifies the version for SAS data sets to be converted. If n is 6, for version 6 or less of SAS, then sas2stata will convert .ssd01 files. If n is 7 or 8, for version 7 or 8 of SAS, then sas2stata will convert .sas7bdat files. By default, long variable names are converted as is. If using Stata version 6 or less, use the -vn8 option to truncate long variable names in SAS data sets to 8 characters.

**-vn8** forces sas2stata to truncate SAS variable names to 8 characters during conversion to Stata. (Warning: This may cause duplicate variable names if two names are the same in the first 8 characters. In this case, the SAS variables should be renamed or converted separately.) Use this option if there are long SAS variable names and Stata is version 6 or less. If running Stata version 7, long variable names are allowed in Stata. If running Stata version 7 with -vn8, Stata file is saved as a version 6 file.

## FEATURES

sas2stata attempts to transfer SAS variable formats to Stata. Fixed SAS formats (Fw.d) translate into Stata's %w.df format. SAS date formats are translated as closely as possible. (Date formats are available in Stata version 4.0 and later only.) Unformatted variables get Stata's default formats for the appropriate data type (%8.0g for bytes and ints, %9.0g for floats, and %10.0g for doubles), except for long variables, which sas2stata formats as %12.0g.

Stata data sets may be named; sas2stata stamps the SAS creation date and time on the Stata data set name, so that the Stata user knows not only when the Stata data set was

created, but also the original SAS creation date and time.

Not all SAS variable names are acceptable in Stata. sas2stata attempts to prevent conflicts by using uppercase names for reserved names. These names are ``byte'`, ``int'`, ``long'`, ``float'`, and ``double'`, as well as names starting with ``str'` and followed by an integer. (For example, name ``street'` does not pose any problems, but SAS name ``str10'` will be translated into Stata name ``STR10'`.) SAS name ``_n'` translates into ``_N000001'` (and a warning is issued.)

(Stata version 6 or less users, please see the `-vn8` option.)

#### FILES

`/homer/a/bin/sas2stata`  
the (cshell) program itself;

`/homer/a/bin/Sas2stata/Scripts`  
directory in which two sas2stata gawk scripts reside.

In addition, numerous standard Unix utilities are used. sas2stata also needs `'gawk'`, the GNU version of `'awk'`.

#### AUTHOR

Originally written by Susan Thomas. Adapted and maintained by Constantijn Panis (panis@rand.org), with valuable input from William H. Rogers and others. Updated to accommodate Stata version 5.0 and above by Patricia St.Clair. Sponsored by the RAND Center for the Study of Aging, with funds from the National Institute on Aging. RAND Corporation, Santa Monica, California.

#### VERSION

The current version is 1.10 and was installed on 2 October 2004.

#### BUGS

May or may not work with SAS versions older than 6.03.

SAS character variables may be up to 200 characters in length, or more for SAS V7 or later; Stata limits its string variables to 80 characters. sas2stata will truncate such variables and write out a warning.

Intercooled Stata data sets are limited to 2048 variables with a maximum width (number of bytes) of 8191. This version of Sas2stata assumes the use of intercooled Stata.