How do I standardize variables in Stata?

Authored by

stats writer

June 30, 2024

To standardize variables in Stata, one can use the command "standardize" followed by the list of variables to be standardized. This command will transform the selected variables to have a mean of 0 and a standard deviation of 1, making them comparable on the same scale. This process is particularly useful when working with datasets that have variables with different scales or units of measurement, as it allows for a more accurate comparison and analysis. Additionally, Stata provides options to specify the mean and standard deviation to be used for the standardization process, giving users more control over the transformation.

How do I standardize variables in Stata? | Stata FAQ

A standardized variable (sometimes called a z-score or a standard score) is a variable that has been rescaled to have a mean of zero and a standard deviation of one. For a standardized variable, each case's value on the standardized variable indicates it's difference from the mean of the original variable in number of standard deviations (of the original variable). For example, a value of 0.5 indicates that the value for that case is half a standard deviation above the mean, while a value of -2 indicates that a case has a value two standard deviations lower than the mean. Variables are standardized for a variety of reasons, for example, to make sure all variables contribute evenly to a scale when

items are added together, or to make it easier to interpret results of a regression or
other analysis.

Standardizing a variable is a relatively straightforward procedure. First, the mean is subtracted
from the value for each case, resulting in a mean of zero. Then, the difference
between the individual's score and the mean is divided by the standard deviation,
which results in a standard deviation of one.
If we start with a variable x, and generate a variable x*, the process is:

x* = (x-m)/sd

Where m is the mean of x, and sd is the standard deviation of
x.

To illustrate the process of standardization, we will use the High School
and Beyond dataset (hsb2). We will create standardized versions of three variables,
math, science, and socst. These variables contain

students'
scores on tests of knowledge of mathematics (math), science (science),
social studies (socst). First, we will use the summarize command (abbreviated
as sum below) to get the mean and standard deviation for each variable.

```
use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear
(highschool and beyond (200 cases))

sum math science socst

Variable | Obs Mean Std. Dev. Min Max
-------------+--------------------------------------------------------
math | 200 52.645 9.368448 33 75
science | 200 51.85 9.900891 26 74
socst | 200 52.405 10.73579 26 71
```

The mean of math is 52.645, and it's standard deviation is 9.368448. Based
on this information, we can generate a standardized version of math called
z1math. The code below does this with the generate

command (abbreviated to gen), then uses summarize to confirm that the mean of z1math is very close to zero (due to rounding error, the mean of a standardized variable will rarely be exactly 0), and the standard deviation is one.

```
gen z1math = (math-52.645)/9.368448
sum z1math

Variable | Obs Mean Std. Dev. Min Max
-------------+--------------------------------------------------------
z1math | 200 -8.51e-09 1 -2.096932 2.386201
```

Below we do the same for science and socst, creating two new variables, z1science and z1socst, using their respective means and standard deviations taken from first table of summary statistics. The table of summary statistics shown below demonstrates that both variables are indeed standardized.

```
gen z1science = (science-51.85)/9.900891
```

gen z1socst = (socst-52.405)/10.73579

sum z1science z1socst

Variable | Obs Mean Std. Dev. Min Max

----------+----------

z1science | 200 -4.95e-09 1 -2.610876 2.237172

z1socst | 200 4.02e-09 1 -2.45953 1.732057

Standardizing variables is not difficult, but to make this process easier,
and less error prone,
you can use the egen command to make standardized variables. The commands
below standardize the values of math, science, and socst,
creating three new variables, z2math, z2science, and z2socst.

egen z2math = std(math)
egen z2science = std(science)
egen z2socst = std(socst)

Again we can look at a table of summary statistics to confirm that these variables are

standardized. Note that the means are not exactly zero, nor do they match the
means from the set of standardized variables created above using the generate
command, in both cases, this is due to very slight rounding error.

```
sum z2math z2science z2socst

Variable | Obs Mean Std. Dev. Min Max
-------------+--------------------------------------------------------
z2math | 200 3.41e-09 1 -2.096932 2.386201
z2science | 200 -2.94e-09 1 -2.610876 2.237172
z2socst | 200 -7.38e-09 1 -2.459529 1.732056
```