Title

stata.com

smooth — Robust nonlinear smoother

Syntax

`smooth` *smoother* [`,` `twice`] *varname* [*if*] [*in*] `,` `generate(`*newvar*`)`

where *smoother* is specified as *Sm* [*Sm* [. . .]] and *Sm* is one of

```
{ 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 } [ R ]
3 [ R ] S [ S | R ] [ S | R ] ...
E
H
```

Letters may be specified in lowercase if preferred. Examples of *smoother* [`,twice`] include

```
3RSSH     3RSSH,twice     4253H     4253H,twice     43RSR2H,twice
3rssh     3rssh,twice     4253h     4253h,twice     43rsr2h,twice
```

Menu

Statistics > Nonparametric analysis > Robust nonlinear smoother

Description

`smooth` applies the specified resistant, nonlinear smoother to *varname* and stores the smoothed series in *newvar*.

Option

`generate(`*newvar*`)` is required; it specifies the name of the new variable that will contain the smoothed values.

Remarks and examples

stata.com

Smoothing is an exploratory data-analysis technique for making the general shape of a series apparent. In this approach (Tukey 1977), the observed data series is assumed to be the sum of an underlying process that evolves smoothly (the smooth) and of an unsystematic noise component (the rough); that is,

$$\text{data} = \text{smooth} + \text{rough}$$

Smoothed values z_t are obtained by taking medians (or some other location estimate) of each point in the original data y_t and a few of the points around it. The number of points used is called the span of the smoother. Thus a span-3 smoother produces z_t by taking the median of y_{t-1}, y_t, and y_{t+1}. smooth provides running median smoothers of spans 1 to 9—indicated by the digit that specifies their span. Median smoothers are resistant to isolated outliers, so they provide robustness to spikes in the data. Because the median is also a nonlinear operator, such smoothers are known as robust (or resistant) nonlinear smoothers.

smooth also provides the Hanning linear, nonrobust smoother, indicated by the letter H. Hanning is a span-3 smoother with binomial weights. Repeated applications of H—HH, HHH, etc.— provide binomial smoothers of span 5, 7, etc. See Cox (1997, 2004) for a graphical application of this fact.

Because one smoother usually cannot adequately separate the smooth from the rough, compound smoothers—multiple smoothers applied in sequence—are used. The smoother 35H, for instance, then smooths the data with a span-3 median smoother, smooths the result with a span-5 median smoother, and finally smooths that result with the Hanning smoother. smooth allows you to specify any number of smoothers in any sequence.

Three refinements can be combined with the running median and Hanning smoothers. First, the endpoints of a smooth can be given special treatment. This is specified by the E operator. Second, smoothing by 3, the span-3 running median, tends to produce flat-topped hills and valleys. The splitting operator, S, "splits" these repeated values, applies the endpoint operator to them, and then "rejoins" the series. Finally, it is sometimes useful to repeat an odd-span median smoother or the splitting operator until the smooth no longer changes. Following a digit or an S with an R specifies this type of repetition.

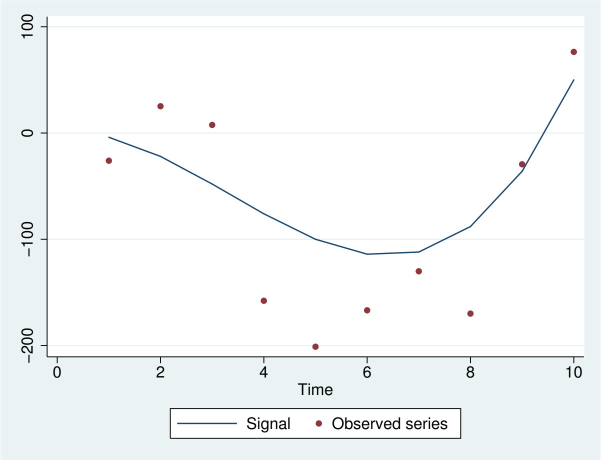
Even the best smoother may fail to separate the smooth from the rough adequately. To guard against losing any systematic components of the data series, after smoothing, the smoother can be reapplied to the resulting rough, and any recovered signal can be added back to the original smooth. The twice operator specifies this procedure. More generally, an arbitrary smoother can be applied to the rough (using a second smooth command), and the recovered signal can be added back to the smooth. This more general procedure is called reroughing (Tukey 1977).

The details of each of the smoothers and operators are explained in *Methods and formulas* below.

▷ Example 1

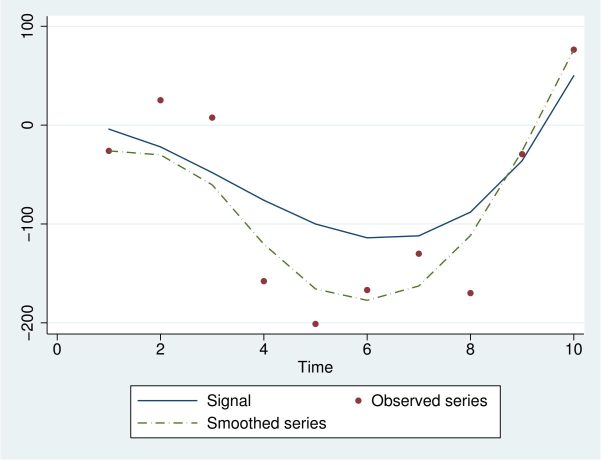
smooth is designed to recover the general features of a series that has been contaminated with noise. To demonstrate this, we construct a series, add noise to it, and then smooth the noisy version to recover an estimate of the original data. First, we construct and display the data:

```
. drop _all
. set obs 10
. set seed 123456789
. generate time = _n
. label variable time "Time"
. generate x = _n^3 - 10*_n^2 + 5*_n
. label variable x "Signal"
. generate z = x + 50*rnormal()
. label variable z "Observed series"
. scatter x z time, c(l .) m(i o) ytitle("")
```



Now we smooth the noisy series, z, assumed to be the only data we would observe:

```
. smooth 4253eh,twice z, gen(sz)
. label variable sz "Smoothed series"
. scatter x z sz time, c(l . l) m(i o i) ytitle("") || scatter sz time,
> c(l . l) m(i o i) ytitle("") clpattern(dash_dot)
```

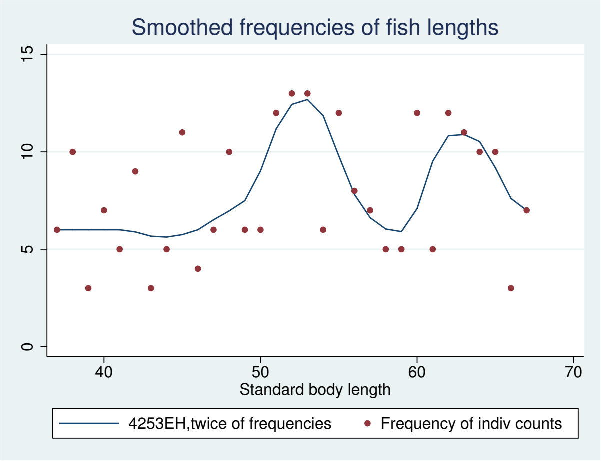


◁

▷ Example 2

Salgado-Ugarte and Curts-García (1993) provide data on the frequencies of observed fish lengths. In this example, the series to be smoothed—the frequencies—is ordered by fish length rather than by time.

```
. use http://www.stata-press.com/data/r13/fishdata, clear
. smooth 4253eh,twice freq, gen(sfreq)
. label var sfreq "4253EH,twice of frequencies"
. scatter sfreq freq length, c(l .) m(i o)
> title("Smoothed frequencies of fish lengths") ytitle("") xlabel(#4)
```



◁

❑ Technical note

smooth allows missing values at the beginning and end of the series, but missing values in the middle are not allowed. Leading and trailing missing values are ignored. If you wish to ignore missing values in the middle of the series, you must drop the missing observations before using smooth. Doing so, of course, would violate smooth's assumption that observations are equally spaced—each observation represents a year, a quarter, or a month (or a 1-year birth-rate category). In practice, smooth produces good results as long as the spaces between adjacent observations do not vary too much.

Smoothing is usually applied to time series, but any variable with a natural order can be smoothed. For example, a smoother might be applied to the birth rate recorded by the age of the mothers (birth rate for 17-year-olds, birth rate for 18-year-olds, and so on).

❑

Methods and formulas

Methods and formulas are presented under the following headings:

Running median smoothers of odd span
Running median smoothers of even span
Repeat operator
Endpoint rule
Splitting operator
Hanning smoother
Twicing

Running median smoothers of odd span

The smoother 3 defines

$$z_t = \text{median}(y_{t-1}, y_t, y_{t+1})$$

The smoother 5 defines

$$z_t = \text{median}(y_{t-2}, y_{t-1}, y_t, y_{t+1}, y_{t+2})$$

and so on. The smoother 1 defines $z_t = \text{median}(y_t)$, so it does nothing.

Endpoints are handled by using smoothers of shorter, odd span. Thus for 3,

$$\begin{aligned} z_1 &= y_1 \\ z_2 &= \text{median}(y_1, y_2, y_3) \\ &\vdots \\ z_{N-1} &= \text{median}(y_{N-2}, y_{N-1}, y_N) \\ Z_N &= y_N \end{aligned}$$

For 5,

$$\begin{aligned} z_1 &= y_1 \\ z_2 &= \text{median}(y_1, y_2, y_3) \\ z_3 &= \text{median}(y_1, y_2, y_3, y_4, y_5) \\ z_4 &= \text{median}(y_2, y_3, y_4, y_5, y_6) \\ &\vdots \\ z_{N-2} &= \text{median}(y_{N-4}, y_{N-3}, y_{N-2}, y_{N-1}, y_N) \\ z_{N-1} &= \text{median}(y_{N-2}, y_{N-1}, y_N) \\ Z_N &= y_N \end{aligned}$$

and so on.

Running median smoothers of even span

Define the median() function as returning the linearly interpolated value when given an even number of arguments. Thus the smoother 2 defines

$$z_{t+0.5} = (y_t + y_{t+1})/2$$

The smoother 4 defines $z_{t+0.5}$ as the linearly interpolated median of $(y_{t-1}, y_t, y_{t+1}, y_{t+2})$, and so on. Endpoints are always handled using smoothers of shorter, even span. Thus for 4,

$$\begin{aligned}
z_{0.5} &= y_1 \\
z_{1.5} &= \text{median}(y_1, y_2) = (y_1 + y_2)/2 \\
z_{2.5} &= \text{median}(y_1, y_2, y_3, y_4) \\
&\vdots \\
z_{N-2.5} &= \text{median}(y_{N-4}, y_{N-3}, y_{N-2}, y_N) \\
z_{N-1.5} &= \text{median}(y_{N-2}, y_{N-1}) \\
z_{N-0.5} &= \text{median}(y_{N-1}, y_N) \\
z_{N+0.5} &= y_N
\end{aligned}$$

As defined above, an even-span smoother increases the length of the series by 1 observation. However, the series can be recentered on the original observation numbers, and the "extra" observation can be eliminated by smoothing the series again with another even-span smoother. For instance, the smooth of `4` illustrated above could be followed by a smooth of `2` to obtain

$$\begin{aligned}
z_1^* &= (z_{0.5} + z_{1.5})/2 \\
z_2^* &= (z_{1.5} + z_{2.5})/2 \\
z_3^* &= (z_{2.5} + z_{3.5})/2 \\
&\vdots \\
z_{N-2}^* &= (z_{N-2.5} + z_{N-1.5})/2 \\
z_{N-1}^* &= (z_{N-1.5} + z_{N-0.5})/2 \\
z_N^* &= (z_{N-0.5} + z_{N+0.5})/2
\end{aligned}$$

`smooth` keeps track of the number of even smoothers applied to the data and expands and shrinks the length of the series accordingly. To ensure that the final smooth has the same number of observations as *varname*, `smooth` requires you to specify an even number of even-span smoothers. However, the pairs of even-span smoothers need not be contiguous; for instance, `4253` and `4523` are both allowed.

Repeat operator

`R` indicates that a smoother is to be repeated until convergence, that is, until repeated applications of the smoother produce the same series. Thus `3` applies the smoother of running medians of span 3. `33` applies the smoother twice. `3R` produces the result of repeating `3` an infinite number of times. `R` should be used only with odd-span smoothers because even-span smoothers are not guaranteed to converge.

The smoother `453R2` applies a span-4 smoother, followed by a span-5 smoother, followed by repeated applications of a span-3 smoother, followed by a span-2 smoother.

Endpoint rule

The endpoint rule `E` modifies the values z_1 and z_N according to the following formulas:

$$z_1 = \text{median}(3z_2 - 2z_3, z_1, z_2)$$
$$z_N = \text{median}(3z_{N-2} - 2z_{N-1}, z_N, z_{N-1})$$

When the endpoint rule is not applied, endpoints are typically "copied in"; that is, $z_1 = y_1$ and $z_N = y_N$.

Splitting operator

The smoothers `3` and `3R` can produce flat-topped hills and valleys. The split operator attempts to eliminate such hills and valleys by splitting the sequence, applying the endpoint rule `E`, rejoining the series, and then resmoothing by `3R`.

The `S` operator may be applied only after `3`, `3R`, or `S`.

We recommend that the `S` operator be repeated once (`SS`) or until no further changes take place (`SR`).

Hanning smoother

`H` is the Hanning linear smoother:

$$z_t = (y_{t-1} + 2y_t + y_{t+1})/4$$

Endpoints are copied in: $z_1 = y_1$ and $z_N = y_N$. `H` should be applied only after all nonlinear smoothers.

Twicing

A smoother divides the data into a smooth and a rough:

$$\text{data} = \text{smooth} + \text{rough}$$

If the smoothing is successful, the rough should exhibit no pattern. Twicing refers to applying the smoother to the observed, calculating the rough, and then applying the smoother to the rough. The resulting "smoothed rough" is then added back to the smooth from the first step.

Acknowledgments

`smooth` was originally written by William Gould (1992)—at which time it was named `nlsm`—and was inspired by Salgado-Ugarte and Curts-García (1992). Salgado-Ugarte and Curts-García (1993) subsequently reported anomalies in `nlsm`'s treatment of even-span median smoothers. `smooth` corrects these problems and incorporates other improvements but otherwise is essentially the same as originally published.

References

Cox, N. J. 1997. gr22: Binomial smoothing plot. *Stata Technical Bulletin* 35: 7–9. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 36–38. College Station, TX: Stata Press.

——. 2004. gr22_1: Software update: Binomial smoothing plot. *Stata Journal* 4: 490.

——. 2005. Speaking Stata: Smoothing in various directions. *Stata Journal* 5: 574–593.

Gould, W. W. 1992. sg11.1: Quantile regression with bootstrapped standard errors. *Stata Technical Bulletin* 9: 19–21. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, pp. 137–139. College Station, TX: Stata Press.

Royston, P., and N. J. Cox. 2005. A multivariable scatterplot smoother. *Stata Journal* 5: 405–412.

Salgado-Ugarte, I. H., and J. Curts-García. 1992. sed7: Resistant smoothing using Stata. *Stata Technical Bulletin* 7: 8–11. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, pp. 99–103. College Station, TX: Stata Press.

——. 1993. sed7.2: Twice reroughing procedure for resistant nonlinear smoothing. *Stata Technical Bulletin* 11: 14–16. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, pp. 108–111. College Station, TX: Stata Press.

Sasieni, P. D. 1998. gr27: An adaptive variable span running line smoother. *Stata Technical Bulletin* 41: 4–7. Reprinted in *Stata Technical Bulletin Reprints*, vol. 7, pp. 63–68. College Station, TX: Stata Press.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison–Wesley.

Velleman, P. F. 1977. Robust nonlinear data smoothers: Definitions and recommendations. *Proceedings of the National Academy of Sciences* 74: 434–436.

——. 1980. Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association* 75: 609–615.

Velleman, P. F., and D. C. Hoaglin. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury.

Also see

[R] **lowess** — Lowess smoothing

[R] **lpoly** — Kernel-weighted local polynomial smoothing

[TS] **tssmooth** — Smooth and forecast univariate time-series data