Title

stata.com

mkspline — Linear and restricted cubic spline construction

Description

`mkspline` creates variables containing a linear spline or a restricted cubic spline of an existing variable. For linear splines, knots can be user specified, equally spaced over the range of the variable, or placed at percentiles. For restricted cubic splines, also known as natural splines, knot locations are based on Harrell's (2001) recommended percentiles or user-specified points.

Quick start

Linear spline of `v1` with knots at 30, 40, and 50

```
mkspline knot1 30 knot2 40 knot3 50 knot4 = v1
```

Add knots at 20 and 60

```
mkspline knot1 20 knot2 30 knot3 40 knot4 50 knot5 60 knot6 = v1
```

Define knots by quintiles

```
mkspline knot 5 = v1, pctile
```

As above, but apply frequency weight `wvar` before calculating quintiles

```
mkspline knot 5 = v1 [fweight=wvar], pctile
```

Restricted cubic spline of `v2` with default 5 knots

```
mkspline knot = v2, cubic
```

As above, but place knots at 30, 40, and 50

```
mkspline knot = v2, cubic knots(30, 40, 50)
```

Menu

Data > Create or change data > Other variable-creation commands > Linear and cubic spline construction

Syntax

Linear spline with knots at specified points

```
mkspline newvar1 #1 [newvar2 #2 [...]] newvark = oldvar [if] [in] [, marginal
    displayknots]
```

Linear spline with knots equally spaced or at percentiles of data

```
mkspline stubname # = oldvar [if] [in] [weight] [, marginal pctile
    displayknots]
```

Restricted cubic spline

```
mkspline stubname = oldvar [if] [in] [weight], cubic [nknots(#) knots(numlist)
    displayknots]
```

`fweights` are allowed with the second and third syntax; see **[U] 11.1.6 weight**.

Options

Options

`marginal` is allowed with the first or second syntax. It specifies that the new variables be constructed so that, when used in estimation, the coefficients represent the change in the slope from the preceding interval. The default is to construct the variables so that, when used in estimation, the coefficients measure the slopes for the interval.

`displayknots` displays the values of the knots that were used in creating the linear or restricted cubic spline.

`pctile` is allowed only with the second syntax. It specifies that the knots be placed at percentiles of the data rather than being equally spaced over the range.

`nknots(#)` is allowed only with the third syntax. It specifies the number of knots that are to be used for a restricted cubic spline. This number must be between 3 and 7 unless the knot locations are specified using `knots()`. The default number of knots is 5.

`knots(`*numlist*`)` is allowed only with the third syntax. It specifies the exact location of the knots to be used for a restricted cubic spline. The values of these knots must be given in increasing order. When this option is omitted, the default knot values are based on Harrell's recommended percentiles with the additional restriction that the smallest knot may not be less than the fifth-smallest value of *oldvar* and the largest knot may not be greater than the fifth-largest value of *oldvar*. If both `nknots()` and `knots()` are given, they must specify the same number of knots.

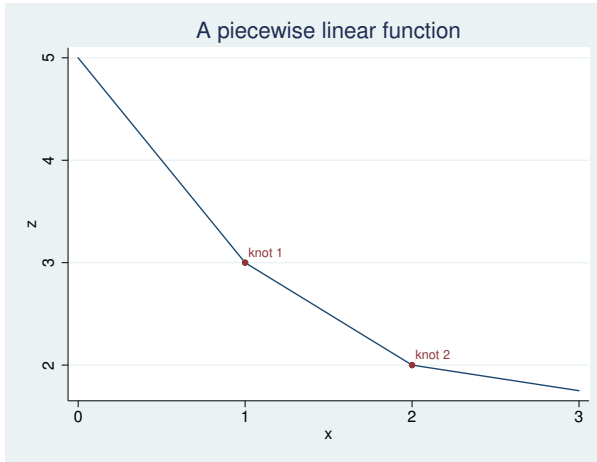
Remarks and examples

stata.com

Remarks are presented under the following headings:

Linear splines

Linear splines allow estimating the relationship between y and x as a piecewise linear function, which is a function composed of linear segments—straight lines. One linear segment represents the function for values of x below x_0, another linear segment handles values between x_0 and x_1, and so on. The linear segments are arranged so that they join at x_0, x_1, ..., which are called the knots. An example of a piecewise linear function is shown below.



▷ Example 1

We wish to fit a model of log income on education and age by using a piecewise linear function for age:

$$\texttt{lninc} = b_0 + b_1\,\texttt{educ} + f(\texttt{age}) + u$$

The knots are to be placed at 10-year intervals: 20, 30, 40, 50, and 60.

```
. use http://www.stata-press.com/data/r15/mksp1
. mkspline age1 20 age2 30 age3 40 age4 50 age5 60 age6 = age, marginal
. regress lninc educ age1-age6
```
(*output omitted*)

Because we specified the `marginal` option, we could test whether the age effect is the same in the 30–40 and 40–50 intervals by asking whether the `age4` coefficient is zero. With the `marginal` option, coefficients measure the change in slope from the preceding group. Specifying `marginal` changes only the interpretation of the coefficients; the same model is fit in either case. Without the `marginal` option, the interpretation of the coefficients would have been

$$\frac{dy}{d\texttt{age}} = \begin{cases} a_1 & \text{if } \texttt{age} < 20 \\ a_2 & \text{if } 20 \le \texttt{age} < 30 \\ a_3 & \text{if } 30 \le \texttt{age} < 40 \\ a_4 & \text{if } 40 \le \texttt{age} < 50 \\ a_5 & \text{if } 50 \le \texttt{age} < 60 \\ a_6 & \text{otherwise} \end{cases}$$

With the marginal option, the interpretation is

$$\frac{dy}{d\texttt{age}} = \begin{cases} a_1 & \text{if } \texttt{age} < 20 \\ a_1 + a_2 & \text{if } 20 \le \texttt{age} < 30 \\ a_1 + a_2 + a_3 & \text{if } 30 \le \texttt{age} < 40 \\ a_1 + a_2 + a_3 + a_4 & \text{if } 40 \le \texttt{age} < 50 \\ a_1 + a_2 + a_3 + a_4 + a_5 & \text{if } 50 \le \texttt{age} < 60 \\ a_1 + a_2 + a_3 + a_4 + a_5 + a_6 & \text{otherwise} \end{cases}$$

◁

▷ Example 2

Say that we have a binary outcome variable called outcome. We are beginning an analysis and wish to parameterize the effect of dosage on outcome. We wish to divide the data into five equal-width groups of dosage for the piecewise linear function.

```
. use http://www.stata-press.com/data/r15/mksp2, clear
. mkspline dose 5 = dosage, displayknots
```

	knot1	knot2	knot3	knot4
dosage	20	40	60	80

```
. logistic outcome dose1-dose5
  (output omitted)
```

mkspline dose 5 = dosage creates five variables—dose1, dose2, . . . , dose5—equally spacing the knots over the range of dosage. Because dosage varied between 0 and 100, the mkspline command above has the same effect as typing

```
. mkspline dose1 20 dose2 40 dose3 60 dose4 80 dose5 = dosage
```

The pctile option sets the knots to divide the data into five equal sample-size groups rather than five equal-width ranges. Typing

```
. mkspline pctdose 5 = dosage, pctile displayknots
```

	knot1	knot2	knot3	knot4
dosage	16	36.4	55.6	82

places the knots at the 20th, 40th, 60th, and 80th percentiles of the data.

◁

Restricted cubic splines

A linear spline can be used to fit many functions well. However, a restricted cubic spline may be a better choice than a linear spline when working with a very curved function. When using a restricted cubic spline, one obtains a continuous smooth function that is linear before the first knot, a piecewise cubic polynomial between adjacent knots, and linear again after the last knot.

▷ Example 3

Returning to the data from example 1, we may feel that a curved function is a better fit. First, we will use the `knots()` option to specify the five knots that we used previously.

```
. use http://www.stata-press.com/data/r15/mksp1, clear
. mkspline agesp = age, cubic knots(20 30 40 50 60)
. regress lninc educ agesp*
```
(*output omitted*)

Harrell (2001, 23) recommends placing knots at equally spaced percentiles of the original variable's marginal distribution. If we do not specify the `knots()` option, variables will be created containing a restricted cubic spline with five knots determined by Harrell's default percentiles.

```
. use http://www.stata-press.com/data/r15/mksp1, clear
. mkspline agesp = age, cubic displayknots
. regress lninc educ agesp*
```
(*output omitted*)

◁

Methods and formulas

Methods and formulas are presented under the following headings:

Linear splines
Restricted cubic splines

Linear splines

Let V_i, $i = 1, \ldots, n$, be the variables to be created; k_i, $i = 1, \ldots, n-1$, be the corresponding knots; and $\mathcal{V}$ be the original variable (the command is `mkspline` V_1 k_1 V_2 k_2 ... V_n `=` $\mathcal{V}$). Then

$$
\begin{aligned}
V_1 &= \min(\mathcal{V}, k_1) \\
V_i &= \max\Big\{ \min(\mathcal{V}, k_i), k_{i-1} \Big\} - k_{i-1} \quad i = 2, \ldots, n-1 \\
V_n &= \max(\mathcal{V}, k_{n-1}) - k_{n-1}
\end{aligned}
$$

If the `marginal` option is specified, the definitions are

$$
\begin{aligned}
V_1 &= \mathcal{V} \\
V_i &= \max(0, \mathcal{V} - k_{i-1}) \quad i = 2, \ldots, n
\end{aligned}
$$

In the second syntax, `mkspline` *stubname* `#` `=` $\mathcal{V}$, so let m and M be the minimum and maximum of $\mathcal{V}$. Without the `pctile` option, knots are set at $m + (M - m)(i/n)$ for $i = 1, \ldots, n-1$. If `pctile` is specified, knots are set at the $100(i/n)$ percentiles, for $i = 1, \ldots, n-1$. Percentiles are calculated by `centile`; see [R] **centile**.

Restricted cubic splines

Let k_i, $i = 1, \ldots, n$, be the knot values; V_i, $i = 1, \ldots, n-1$, be the variables to be created; and $\mathcal{V}$ be the original variable. Then

$$V_1 = \mathcal{V}$$

$$V_{i+1} = \frac{(\mathcal{V}-k_i)_+^3 - (k_n-k_{n-1})^{-1}\{(\mathcal{V}-k_{n-1})_+^3(k_n-k_i) - (\mathcal{V}-k_n)_+^3(k_{n-1}-k_i)\}}{(k_n-k_1)^2}$$

$$i = 1, \ldots, n-2$$

where

$$(u)_+ = \begin{cases} u, \text{ if } & u > 0 \\ 0, \text{ if } & u \leq 0 \end{cases}$$

Without the `knots()` option, the locations of the knots are determined by the percentiles recommended in Harrell (2001, 23). These percentiles are based on the chosen number of knots as follows:

No. of knots	Percentiles						
3	10	50	90				
4	5	35	65	95			
5	5	27.5	50	72.5	95		
6	5	23	41	59	77	95	
7	2.5	18.33	34.17	50	65.83	81.67	97.5

Harrell provides default percentiles when the number of knots is between 3 and 7. When using a number of knots outside this range, the location of the knots must be specified in `knots()`.

Acknowledgment

The restricted cubic spline portion of `mkspline` is based on the `rc_spline` command by William Dupont of the Department of Biostatistics at Vanderbilt University.

References

Gould, W. W. 1993. sg19: Linear splines and piecewise linear functions. *Stata Technical Bulletin* 15: 13–17. Reprinted in *Stata Technical Bulletin Reprints*, vol. 3, pp. 98–104. College Station, TX: Stata Press.

Greene, W. H. 2018. *Econometric Analysis*. 8th ed. New York: Pearson.

Harrell, F. E., Jr. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Newson, R. B. 2000. sg151: B-splines and splines parameterized by their values at reference points on the x-axis. *Stata Technical Bulletin* 57: 20–27. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 221–230. College Station, TX: Stata Press.

——. 2012. Sensible parameters for univariate and multivariate splines. *Stata Journal* 12: 479–504.

Orsini, N., and S. Greenland. 2011. A procedure to tabulate and plot results after flexible modeling of a quantitative covariate. *Stata Journal* 11: 1–29.

Panis, C. 1994. sg24: The piecewise linear spline transformation. *Stata Technical Bulletin* 18: 27–29. Reprinted in *Stata Technical Bulletin Reprints*, vol. 3, pp. 146–149. College Station, TX: Stata Press.

Also see

[R] **fp** — Fractional polynomial regression