The SAS MAKESPL Macro
Ellen Hertzmark, Ruifeng Li, and Donna Spiegelman

Abstract

The %MAKESPL macro is a SAS macro that makes restricted cubic spline variables to be used in procedures. It is incorporated in several of the macros that test for non-linearity, but can also be used on its own to create spline variables for covariates (allowing better control for the covariates usually using up fewer degrees of freedom).

Keywords: SAS, macro, restricted cubic splines, nonlinear relation

Contents

1. Description
2. Invocation and Details
3. Examples
4. Computational Methods
5. References
6. Credits
7. See Also

1: Description

%MAKESPL is a SAS macro that makes restricted cubic spline variables for a given number of knot points or a list of knots. This macro is useful when you wish to control for confounding by a continuous variable with a possibly non-linear relationship with the outcome.

2: Invocation and Details
To call %MAKESPL, your program must know where to look for it. The most efficient way is to include the following statement (or its equivalent) at the top of your program.

```options mautosource sasautos='/usr/local/channing/sasautos';%
```

%MAKESPL has very few parameters.

**DATA**  
The name of data set on which the regression is to be run

**REQUIRED**

**SPLVBL**  
The name of the variable to be "splined"

**REQUIRED**

**NK**  
The number of knots, if you want %MAKESPL to place them automatically. The numbers of knots for which automatic knot placement can be done are 3, 4, 5, 6, 7, 8, 9, 10, 17, 21, 25, and 50. See COMPUTATIONAL METHODS for details about knot placement. default is 4

**KNOT1**  
The list of knots if you want to determine them. If you give a value for KNOT1, this takes precedence over any value NK may have.

**REFVAL**  
A value of SPLVBL to use as the reference level for procedures that need reference levels (LOGISTIC, log-binomial regression, PHREG).
The macro prints the values of all the spline variables when SPLVBL = REFVAL. To do this, it adds one observation to the dataset.

default is the minimum value of SPLVBL

**NOTE:** Sometimes the best way to control for covariates that are originally continuous is to use 3 or 4-knot splines, rather than sets of indicators for categorical levels of the covariates.
For many purposes (such as plotting and estimation), you will need to know the values of the spline variables at a specified value for the covariate. You can find this out by using that specified value as REFVAL.

If you wish to make spline variables for several covariates, you must do them one at a time, using the OUTDAT of one run as the DATA of the next (See Example 1). If you set REFVAL=(value SPLVBL will have in ADJDAT), the macro will print out the values.

**NOTE:** If the value of REFVAL is outside the range of your data, the macro will give you the following warning message in the .log file.

**WARNING:** Your REFVAL, <value of REFVAL>, is lower than all values of the exposure in of the data.
The macro will continue, but the graph may look strange.

**OR**

**WARNING:** Your REFVAL, <value of REFVAL>, is
higher than all values of the exposure in of the data.
The macro will continue, but the graph may look strange.

'The graph may look strange' means that the point for
which SPLVBL = REFVAL will be included in the plotting points.
For example, if the range of SPLVBL in the data is 16-40,
and REFVAL=0, and the predicted value for SPLVBL=0 is a lot
higher than that for SPLVBL=16, your graph will have a large
drop from 0 to 16, and all the real information may be squeezed
into a small part of the graph.

OUTDAT The name of the dataset %MAKESPL will make.
This dataset includes all the variables in the original
dataset, plus the newly made spline variables.
It has the extra observation made for SPLVBL = REFVAL.
If MAKEPTS = T, it includes the plotting points (See below).
default=_splstuf

MAKEPTS Whether you want to make 501 plotting points for a graph.
If MAKEPTS = T, the plotting points in the OUTDAT dataset have
non-missing values for SPLVBL and COVAR (See below).
Values the outcome variable in your analysis are missing.
Because the outcome variable is missing, the plotting points
will not be used in the estimation routine, but those procedures
that predict values of the outcome based on the model will
produce predicted values for the graph.
The macro makes a variable ({\tt ine}) in OUTDAT to distinguish
the plotting points.
default is F

COVAR The list of covariates in the model other than SPLVBL.
This is only required if MAKEPTS=T and there are covariates
in your model and the vertical axis of your graph will be an
absolute quantity (predicted probability, value of a continuous
variable), rather than a relative quantity (relative risk, odds
ratio).

ADJDAT The name of a data set with one (1) observation containing
the values of each covariate to be used for plotting,
IF there are covariates in your model and MAKEPTS=T.
This is REQUIRED for PROC MIXED and PROC GENMOD based
procedures, as well as for the PLOTPROB=T option in
LGTPHCURV8.

3: Examples

Using a data set from HPFS, we are planning to examine the relationship of a number of risk
factors to BMI, cross-sectionally in 1986. We plan to use splines for all the continuous variables.

BMI86 is the individual's BMI in 1986
Example 1. Making splines for age86 and tfat86n.

To make "spline" variables for age86 and tfat86n, and to print out 3 observations from the output data set each time we ran the macro, we used the following code.

```
title2 'make 4 knot splines for age86';
%makespl(data=alld, splvbl=age86, nk=4, refval=55, outdat=a1);
title2 'first 3 observations of a1';
proc print data=a1 (obs=3);
  var age86 age861 age862 tfat86n alco86n bmi86;
title2 'make 5 knot splines for tfat86n';
%makespl(data=a1, splvbl=tfat86n, nk=5, refval=70, outdat=a2);
title2 'first 3 observations of a2';
proc print data=a2 (obs=3);
  var age86 age861 age862 tfat86n tfat86n1 tfat86n2 tfat86n3
    alco86n bmi86;
run;
```

By using the output data set (OUTDAT) from the first run of %MAKESPL as the input data set (DATA) for the second run of %MAKESPL, the output data set of the second run (a2) has the spline variables for age86 and for tfat86n.

Since we are not planning to do the analysis before running we left MAKEPTS=F (the default) both times.

Since NK for age86 was 4, the "spline" variables for age86 are named age861 and age862. Since NK for tfat86n was 5, the "spline" variables for tfat86n are named tfat86n1, tfat86n2, and tfat86n3. In general the names will be variable name1 ...variable name(nk-2), where NK is the number of knots, either given to the macro or counted from a list given in KNOT1.

The results of the above bit of program are shown below.
First the macro lists the knot locations.
Because we gave the value 55 for REFVAL in the first call to %MAKESPL and 70 for REFVAL in the second call to %MAKESPL, we get printouts of the values of the "spline" variables at the reference values. If we had not given a value for REFVAL, the macro would have used the
default value (the minimum value of age86), and all spline variables would have had value 0 (because the first knot is greater than the minimum). We will use these in the ADJDAT86 data set (see Example 2).

values of spline variables when age86 is 55

<table>
<thead>
<tr>
<th>OBS</th>
<th>AGE86</th>
<th>AGE861</th>
<th>AGE862</th>
</tr>
</thead>
<tbody>
<tr>
<td>11766</td>
<td>55</td>
<td>4.62678</td>
<td>0.99778</td>
</tr>
</tbody>
</table>

values of spline variables when tfat86n is 70

<table>
<thead>
<tr>
<th>OBS</th>
<th>TFAT86N</th>
<th>TFAT86N1</th>
<th>TFAT86N2</th>
<th>TFAT86N3</th>
</tr>
</thead>
<tbody>
<tr>
<td>11767</td>
<td>70</td>
<td>7.19314</td>
<td>0.75785</td>
<td>.0027852</td>
</tr>
</tbody>
</table>
Example 2. Making the data set with plotting points for ALCO86N

We begin where Example 1 left off.

/* make data set with one observation to give values of covariates for plotting points.
We got these values from the MAKESPL above,
because we gave values for REFVAL */

data adjdat86;
  age86=55;  age861=4.626778;  age862=.9978;  tfat86n=70;
  tfat86n1=7.19314;  tfat86n2=.75785;  tfat86n3=.002852;  smk86=0;
run;

title2 'make 3 knot splines for ALCO86N. make plotting points';
%makespl(data=a2, splvbl=alco86n, nk=3, makepts=T, refval=0,
covar=age86 age861 age862 tfat86n tfat86n1 tfat86n2 tfat86n3 smk86,
adjdat=adjdat86, outdat=a3);

title2 'first 3 observations in a3 ';
proc print data=a3 (obs=3);
  var age86 age861 age862 tfat86n tfat86n1 tfat86n2 tfat86n3 alco86n alco86n1 bmi86 smk86;
run;

title2 'first 5 plotting points';
data pltpts; set a3; where _ine_ eq 1;
proc print data=pltpts (obs=5);
var age86 age861 age862 tfat86n tfat86n1 tfat86n2 tfat86n3 alco86n alco86n1 bmi86 smk86;
run;

The variable _ine_ is an indicator for being a plotting point (for the latest run of %MAKESPL). The results of this bit of code are below.

<p>| | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OBS</td>
<td>ALCO86</td>
<td>ALCO86N1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11768</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12269</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

values of spline variables when alco86n is 0

<table>
<thead>
<tr>
<th>OBS</th>
<th>AGE86</th>
<th>AGE861</th>
<th>AGE862</th>
<th>TFAT86N</th>
<th>TFAT86N1</th>
<th>TFAT86N2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69</td>
<td>23.8519</td>
<td>9.2181</td>
<td>41.2</td>
<td>0.0297</td>
<td>0.00000</td>
</tr>
<tr>
<td>2</td>
<td>73</td>
<td>30.0741</td>
<td>11.9835</td>
<td>34.1</td>
<td>0.0000</td>
<td>0.00000</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>28.5185</td>
<td>11.2922</td>
<td>79.2</td>
<td>14.6725</td>
<td>2.92104</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OBS</th>
<th>TFAT86N3</th>
<th>ALCO86N</th>
<th>ALCO86N1</th>
<th>BMI86</th>
<th>SMK86</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00000</td>
<td>0.0</td>
<td>0.0000</td>
<td>25.1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.00000</td>
<td>39.7</td>
<td>13.3333</td>
<td>24.3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.28549</td>
<td>0.0</td>
<td>0.0000</td>
<td>26.3</td>
<td>0</td>
</tr>
</tbody>
</table>
first 5 plotting points

15:17 Wednesday, July 24, 2002

<table>
<thead>
<tr>
<th>OBS</th>
<th>AGE86</th>
<th>AGE861</th>
<th>AGE862</th>
<th>TFAT86N</th>
<th>TFAT86N1</th>
<th>TFAT86N2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55</td>
<td>4.6268</td>
<td>0.9978</td>
<td>70</td>
<td>7.19314</td>
<td>0.75785</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>4.6268</td>
<td>0.9978</td>
<td>70</td>
<td>7.19314</td>
<td>0.75785</td>
</tr>
<tr>
<td>3</td>
<td>55</td>
<td>4.6268</td>
<td>0.9978</td>
<td>70</td>
<td>7.19314</td>
<td>0.75785</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>4.6268</td>
<td>0.9978</td>
<td>70</td>
<td>7.19314</td>
<td>0.75785</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
<td>4.6268</td>
<td>0.9978</td>
<td>70</td>
<td>7.19314</td>
<td>0.75785</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OBS</th>
<th>TFAT86N3</th>
<th>ALCO86N</th>
<th>ALCO86N1</th>
<th>BMI86</th>
<th>SMK86</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.002852</td>
<td>0.00</td>
<td>.000000</td>
<td>.</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>.002852</td>
<td>0.09</td>
<td>.000000</td>
<td>.</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>.002852</td>
<td>0.18</td>
<td>.000000</td>
<td>.</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>.002852</td>
<td>0.27</td>
<td>.000000</td>
<td>.</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>.002852</td>
<td>0.36</td>
<td>.000000</td>
<td>.</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that in the plotting points, bmi86, which is neither SPLVBL nor in the COVAR list, is missing. Also, values for the age86, tfat86n, and smk86 variables are all the same (= the plotting levels from ADJDAT86).

**Example 3.** Specifying the knots for ALCO86N in advance

Below we show the code and the results from specifying the knots for ALCO86N at 0, 2, and 35, instead of the automatic locations of 0, 6, and 35.1.

title2 'make spline variables for ALC86 (=ALCO86N) with knots given';
%makespl(data=a3, splvbl=alc86, knot1=0 2 35, makepts=T, covar=age86 age861 age862 tfat86n tfat86n1 tfat86n2 tfat86n3 smk86, adjdat=adjdat86, refval=0, outdat=a4);
title2 'first 6 observations of a4';
proc print data=a4 (obs=6);
var age86 age861 age862 tfat86n tfat86n1 tfat86n2 tfat86n3 alco86n alco86n1 alc86 alc861 bmi86 smk86;
run;

Knots for alc86
0 2 35
In this example, we printed out the values of the "spline" variables we got from the previous run (with automatic knot placement), as well as the values of the "spline" variables we got from this call to %MAKESPL. Because the knots are in different locations, the values of the "spline" variables are different unless the value of ALCO86N=ALC86 is 0.

Example 4. Automatic knot placement does not yield distinct knots

Since alcohol consumption is skewed toward 0, requiring a large number of knots will lead to at least two of the knots having value 0. This causes an ERROR message.

```sas
title2 'try to make 8-knot splines for ALCO86N';
%makespl(data=a4, splvbl=alc86, nk=8, makepts=T, refval=0, covar=age86 age861 age862 tfat86n tfat86n1 tfat86n2 tfat86n3 smk86, adjdat=adjdat86, outdat=a5);
```
try to make 8-knot splines for ALCO86N

Knots for alc86
0 0 1.9 4.6 8.7 14 23.4 69.8
"

ERROR in macro run: The spline-making procedure did not generate 8 distinct knot points.
You will either have to change the number of knot points or supply values for the knots, i.e. KNOT1=
The macro will terminate.

NOTE: In the women's cohorts, it may be impossible to use automatic placement for 4 knot splines for alcohol intake variables.

4: Computational Methods

1. Automatic Knot Placement

   nk     percentile locations of the knots
   --     -----------------------------

       3  5  50  95
       4  5  35  65  95
       5  5  27.5  50  72.5  95
       6  5  23  41  59  77  95
       7  2.5  18.3  34.2  50  65.8  81.7  97.5
       8  1  15  29  43  57  71  85  99
       9  2  14  26  38  50  62  74  86  98
      10  2  12.7  23.3  34  44.7  55.3  68  76.7  87.3  98
      17  2  8  14  20  26  32  38  44  50  56  62  68  74  80  86  92  98
      21  1  4  9  14  19  24  29  34  39  44  49  54  59  64  69  74  79  84  89  94  99
      25  2  6  10  14  18  22  26  30  34  38  42  46  50  54  58  62  66  70  74  78  82  86  90  94  98
      50  1  3  5  7  9  11  13  15  17  19  21  23  25  27  29  31  33  35  37  39  41  43  45  47  49  51  53  55  57  59  61  63  65  67  69  71  73  75  77  79  81  83  85  87  89  91  93  95  97  99

2 Computation of the spline variables:
Let $t_j$ be the $j$th knot point.

Let $kd = (t_{nk} - t_j)^{2/3}$, where $kd$ is a normalizing parameter to get the spline variables back into the original units.

For a level of the exposure $x$, $x_j$, the value of the $j$th spline variable ($j$ runs from 1 to $NK-2$) is given by

$$x_j = \max((x - t_j)/kd, 0)^3$$

$$+ \left(t_{nk-1} - t_j\right) \cdot \max((x - t_{nk})/kd, 0)^3$$

$$- \left(t_{nk} - t_j\right) \cdot \max((x - t_{nk-1})/kd, 0)^3/(t_{nk} - t_{nk-1})$$

For $x < t_j$ the value of the $j$th spline variable is 0 (as are the 'higher' spline variables) (because all the 'max' values are 0, since $x < t_j < t_{nk-1} < t_{nk}$. As $x$ gets larger, it has more and more nonzero spline variables.

As an example, we show the values of the 3 "spline" variables for the 5 knot splines of tfat86n. The graph below was made using the 'plotting points' using the following code.

```r
%makespl( data=alld, splvbl=tfat86n, nk=5, makepts=T, outdat=a99);
data inest; set a99;
if _ine_ eq 1;
run;
```

This code uses the automatic _ine_ variable to select the 'plotting points.'
The reference lines are at the knot locations.
In fact, \( t_{fat86n1} \) becomes nonzero at the first knot, \( t_{fat86n2} \) becomes nonzero at the second knot, and \( t_{fat86n3} \) becomes nonzero at the third knot. Because of the large range of the vertical axis, this is not visible on the graph.

5: References


6: Credits

Written by Ellen Hertzmark and Donna Spiegelman for the Channing Laboratory. Based on a macro by Frank Harrell. Questions can be directed to Ellen Hertzmark, stleh@channing.harvard.edu, (617) 432-4957.

7: See Also

Channing SAS macro %LGTPHCURV9