The SAS GLMCURV9 Macro

Ellen Hertzmark, Ruifeng Li, Biling Hong, and Donna Spiegelman

October 28, 2014

Abstract

The %GLMCURV9 macro uses `SAS PROC GENMOD` and restricted cubic splines to test whether there is a nonlinear relation between a continuous exposure and an outcome variable. The macro can automatically select spline variables for a model. It produces a publication quality graph of the relationship.

Keywords: generalized linear models, PROC GENMOD, generalized estimating equations, GEE, nonlinearity, restricted cubic splines, automatic variable selection, graphics

Contents

1 Description 2

2 Invocation and Details 2

3 Examples 9

3.1 Example 1. A minimal call to the macro 9

3.2 Example 2. A more specified call with NK=3 13

3.3 Example 3. Using 5-knot splines, USERSPLV, making a file of plotting points, and showing the use of PRINTPOINTS 14

3.4 Example 4. An example with multiple observations per subject 16

3.5 Example 5. An example with multiple observations per person and MDEP(1) working covariance structure 20

3.6 Example 6. Using BYVAR, AXLABMULT, and other customizations, as well as automatic selection without specifying the number of knots 23

4 Computational Methods 30

4.1 Automatic knot placement, given a desired number of knot points 30

4.2 Computation of the spline variables: 30

4.3 Default bandwidth for smoothing: 31

4.4 Outline of stepwise model selection: 31

5 Including the graph in a MS-WORD document 31

6 How should I describe this in my Methods section? 32

7 WARNINGS 32

8 Frequently Asked Questions 32

8.1 Q: Why does a variable named MVARBL keep showing up in my output? 32

8.2 Q: If the p-value for nonlinearity is greater than .05 but the p-value for the overall significance of the curve is less than .05, why can't I say that the relationship is nonlinear? 33

8.3 Q: Why does the confidence cloud stop abruptly in the middle of the graph? 33

8.4 Q: Why is the confidence band so wide? 33

8.5 Q: Why are the values on the horizontal axis printing out vertically? 33

8.6 Q: I want to plot the smoothed histogram, but the SAS .log says that the Sheather-Jones plug-in did not converge 33

8.7 Q: Why are the coefficients of the spline variables so large in absolute value? 34

8.8 Q: I got an error saying the x-origin did not leave enough space for the text. 34

8.9 Q: How do I make ADJDAT? 34

9 References 34

10 Credits 35

1 Description

The %GLMCURV9 macro uses `SAS PROC GENMOD` and restricted cubic splines to test whether there is a nonlinear relation between a continuous exposure and an outcome variable. The outcome variable can be dichotomous (binary) or continuous. The user can control the variance distribution, the link function, and the working covariance structure. The macro can automatically select spline variables for a model. It produces a publication quality graph of the relationship. It can also produce a file to use in other plotting programs.

2 Invocation and Details

Parameters with default values are listed as OPTIONAL below.

```
PARAMETERS RELATING TO THE DATA SET
===================================
DATA=,  The name of the dataset to use
        REQUIRED
EXPOSURE=,  The continuous variable you are testing for
            nonlinearity
            REQUIRED
OUTCOME=,  The dependent variable.
           This can be dichotomous/binary or continuous
           REQUIRED
HPCT=,  The percentile of the EXPOSURE at which to trim the data at the high end
        If the data are highly skewed, the highest values may not
        be helpful in determining what the relationship is in the
        bulk of the sample.
        Typically at least 95.
        OPTIONAL
LPCT=,  The percentile of the EXPOSURE at which to trim the data  at the low end.
        Similar to HPCT.
        Typically at most 5
        OPTIONAL
HICUT=,  A value of the EXPOSURE at which to trim the data at the high end.
         This may be based on biological plausibility for values of the
         exposure, but you don't want to delete too much of your sample
         OPTIONAL
LOWCUT=,  A value of the EXPOSURE at which to trim the data at the low end.
          Similar to HICUT
          OPTIONAL
WHERE=,  A conditional clause to restrict the analysis to a subset of the data
         examples:  sex eq male
                    %str (age lt 75)
         NOTE:  No commas or = signs allowed, unless using %str
         OPTIONAL
SUBJECT=,
         REQUIRED if you have repeated measures.

PARAMETERS RELATED TO THE MODEL AND THE REPORT
==============================================
DIST=,  The distribution to use.
        Options are N (normal), B (binomial), P (poisson), G (gamma),
        NB (negative binomial), MULT (multinomial), IG (inverse gaussian).
        REQUIRED
LINK=,  The link function to use.
        Options are ID, LOG, LOGIT, Power(-1), Power(-2), CLOGIT
        If you do not specify the LINK, the macro will use the canonical
        link function for the specified distribution, and will tell you so
        in a diagnostic
        OPTIONAL (if you want to use the canonical link function)
ADJ=,  The list of covariates in the model.
       This can be written using the macro variables made by %INDIC3,
       but CANNOT use notation like xq1-xq4 or xq:
```

```
       OPTIONAL (i.e. crude models are allowed)
ADJDAT=,  The name of the SAS dataset containing the values of the
          covariates at which you want to evaluate the outcome for the
          graph.  These should usually be near the center of the
          distribution for each variable.
          ADJDAT has one observation.
          REQUIRED if there are covariates and the plot is not the
          odds ratio or the risk ratio (i.e. PLOTOR=F,  see below)
BYVAR=,  The name of a variable by which to stratify the results.
         e.g. SEX, AGEGP
         NOTE:  If BYVAR has more than 2 values, the prediction curves
         will be plotted on the same graph without not the confidence limits.
         OPTIONAL
WEIGHTVAR=,  The name of a variable in DATA to be used as a weight
             for the model
             OPTIONAL
REFVAL=MIN,  The value to use as a reference value for the splines.
             Can be a user-given number or a statistic such as
             MIN, MAX, MEAN, MEDIAN.
             OPTIONAL and not applicable for continuous outcome variables.
NK=,  The number of knots to use.
      Can be 3, 4, 5, 6, 7, 8, 9, 10, 17, 21, 25, 50.
      If you do not give NK or KNOT (see below) and have SELECT=3,
      (see below), the macro will set NK=21.  If SELECT is not 3,
      the macro will set NK=4 by default.
      Among the diagnostics the macro prints out are lines stating
      the fractions of the data range below the lowest knot or
      above the highest knot.  If these are above 20%, it is
      advisable to do one of the following:  (1)  trim the data using
      HPCT, LPCT, HICUT, or LOWCUT (as appropriate);
      (2) use KNOT (see next parameter);
      (3) adjust the horizontal axis so that not too much of the visible
      graph is beyond the last knot (see AXORDH).
      and add at least one more knot location
      OPTIONAL
KNOT=,  a list of values of the EXPOSURE to be the knots for the splines.
        IF KNOT is not null, it overrides NK
        OPTIONAL
USEGEE=F,  Whether you want to use the empirical variance when there
           is one record per subject.
           USEGEE=T by default if you have any subject with more
           than one observation.
           OPTIONAL
REPTYPE=IND,  The working correlation structure to use.
              IND for one observation per person;
              EXCH or CS for exchangeable;
              MDEP(1), MDEP(2) for Toeplitz types;
              UN
              REQUIRED
WITHINVAR=,  If REPTYPE is not IND or EXCH, the variable that tells PROC GENMOD
```

```
             the order of the observations within each SUBJECT,
             REQUIRED (if REPTYPE is not IND or EXCH)
SELECT=1,  Whether you want to use automatic selection to make your model.
           1=No selection (i.e. use all spline variables)
             If NK is large, this may result in a graph with many
             unimportant/extraneous 'wiggles.'
           2=Use spline variables specified by user (See USERSPLV below)
           3=Use automatic selection of spline variables
           OPTIONAL
USERSPLV=,  If SELECT=2 and you are not using BYVAR, the list of spline variables to
            include in the model.
            This would normally come from a previous run of the macro
            in a situation where the selection takes a lot of time and
            you are just trying to improve the graph.
USERSPLV1=,  Like USERSPLV for the first level of the BYVAR, if you have one
USERSPLV2=,  Like USERSPLV for the second level of the BYVAR, if you have one
...
...
USERSPLV10=,  Like USERSPLV for the tenth level of the BYVAR, if you have one
SLE=.05,  If SELECT=3, the p-value at which a variable enters the model
          OPTIONAL
SLS=.05,  If SELECT=3, the p-value at which a variable leaves the model
          OPTIONAL
HEADER1=,  A description of the analysis
           The default is GRAPHTIT (See below), and if that is empty,
           OUTCOME and EXPOSURE.
           OPTIONAL
TESTREP=LONG,  The type of report you want for the tests of the 3
               hypotheses:
               1.  Whether the relationship between the EXPOSURE
                   and the OUTCOME, if it is significant, is
                   nonlinear;
               2.  If the p-value for the first test is small, the
                   p-value for the overall significance of the curve
                   (i.e., is the nonlinear relationship significant?)
               3.  If the p-value for the first test is large, the
                   p-value for the linear relationship (i.e. is the
                   linear relationship significant?)
               The LONG report (default) gives explicit directions
               about how to read the report.  The SHORT report is
               more terse.
              OPTIONAL
MODPRINT=T,  Whether to print the results of the 3 models:
             Adjusters only
             Adjusters plus linear EXPOSURE
             Adjusters plus linear EXPOSURE plus any splines used
             (based on SELECT=1 or 2, or on automatic selection)
             OPTIONAL
PVALUEFORMAT=pvalue6.4,  The format for printing the p-values in the
                         report.
                         OPTIONAL

PARAMETERS RELATING TO THE GRAPH
================================
PLOT=2,  The type of plot output you want:
         0.  No plot
         1.  PROC PLOT in the SAS .lst
         2.  Some publication-quality format
             Options are PS (encapsulated postscript), PDF, JPEG, CGM (see
             OUTPLOT below).
         3.  PROC PLOT and a publication-quality format
         4.  Just a text file of the plotting points to use in some other
             plotting program
         OPTIONAL
OUTPLOT=PS,  The type of publication-quality format you want.
             If OUTPLOT is not 2 or 3, this will be ignored.
             OPTIONAL
GRAPHTIT=  The title of the graph.
           If no GRAPHTIT is specified, HEADER1 will be used.
           IF no title is wanted (e.g. for publication), use GRAPHTIT=NONE (upper case requi
           OPTIONAL
PICTNAME=&data..&exposure..&outcome..sel&select._&nk..&outplot ,
             The name of the graphics file.
             We strongly suggest that you use a mnemonic name,
             rather than relying on the default (which can get
             overwritten if another macro call with the same
             parameters is run.
             OPTIONAL

FOOTER=DEFAULT,  The footnote for the graph.
                 The default footnote lists the first few variables in ADJ,
                 followed by 'and other variables'.
                 If you want a footnote, it is more informative to make up
                 your own.
                 If no footnote is desired, set FOOTER=NONE.
                 OPTIONAL
PLOTOR=F,  Whether to plot the odds ratio or risk ratio, rather than
           the direct outcome of the model.  Usually models with logit
           or log links will have PLOTOR=T, while other models will not.
           OPTIONAL
CI=T,  Whether to plot confidence band as cloud
       If you say 'F,' the confidence band will be bounded by
       dot-dash curves.
       OPTIONAL
PWHICH=SPLINE,  Whether to plot the spline graph or the linear graph.
                It is not usually of interest to plot the linear graph
                if the relationship is not nonlinear,
                OPTIONAL
DISPLAYX=T  Whether to show the smoothed histogram of the EXPOSURE
            OPTIONAL
```

```
VLABEL=  The label for the vertical axis.
         Be specific.  Give a human-interpretable label with units, if
         appropriate (e.g. Predicted BMI (kg per sq m), or
         %str(Predicted BMI (kg/sq. m)) ).
         If PLOTOR=T, Predicted Odds Ratio (or Predicted Risk Ratio) for
         {some human-interpretable description of the outcome}
         OPTIONAL
AXORDV=  description of the vertical axis given as
         {lowest value} to {highest value} by {distance between
         major tick marks}.
         If you do not specify AXORDV, the macro can automatically
         specify one, but it may not be as good looking or intuitive
         as one you specify yourself.
         There should be 8-12 major tick marks.
         Giving too few makes the graph hard to read.
         Giving too many may result in values overwriting each other.
         OPTIONAL
VLABELSTYLE=V,  Whether you want the label for the vertical axis to
                print parallel to the axis or horizontally
                (vlabelstyle=H, easier
                to read, but it usually takes up too much of the
                graphics area and squeezes the actual graph).
                OPTIONAL
HLABEL,   The label for the horizontal axis.
          Be specific.  Give a human-interpretable label with units,
          e.g. Weight (kg), not Weight,  Years since randomization, not Time.
          NOTE:  If there is punctuation (or a percent sign),
          you may need to use %str to
          keep the macro interpreter from misunderstanding the label
          OPTIONAL
AXORDH=,  description of the horizontal axis.
          All remarks relating to AXORDV apply here too.
          OPTIONAL
PRINTLEGEND=T,  Whether to print the legend for the graph.
                OPTIONAL
LEGLAB1=, LEGLAB2=,...LEGLAB10=,   A set of (short) labels for the levels
                                   of BYVAR
                                   OPTIONAL
FONT=swiss  A font for the graph.
            It is best to use a true-type font, rather than one that
            has to be simulated by your software
            OPTIONAL
TITLEMULT=1,  A multiplier for the print of the graph title, if any.
             The interaction between fonts and output devices is complex,
             and this allows you to optimize the look of the title.
             OPTIONAL
AXLABMULT=1,  A similar multiplier for the print of the axis labels.
              OPTIONAL
AXVALMULT=1,  A similar multiplier for the print of the numeric
              values on the axis
```

```
               OPTIONAL
CUTOFF=F,  Whether to cut off the vertical axis at a specified value.
           Sometimes the upper confidence bound gets very high in areas
           of sparse data, to the extent that the prediction curve is
           squashed at the bottom of the graph, and its shape cannot be
           discerned.  To make the confidence-cloud-making apparatus happy,
           the macro needs to know that values will not go higher than some
           level.
           If a cutoff is used, it should be in the form
           CUTOFF=2 x,
           where x is a numeric value (e.g. 10).
           OPTIONAL

BELOW ARE LINETYPES  and COLORS for up to 10 lines (including
confidence bounds.
=============================================================
LINETYPE1 (solid line)=1,
LINETYPE2=20 (small dashed line),
LINETYPE3=35 (far apart dots),
LINETYPE4=4 (medium dashed line),
LINETYPE5=8 (dot and dashed line),
LINETYPE6=40 (2- and 3- groups of dots line),
LINETYPE7 (very long dashed line)=7,
LINETYPE8=2 (medium dashed and dot line),
LINETYPE9=30 (long, short, and medium patterns of dots line),
LINETYPE10=34 (dotted line),   The types of lines used for the plot.
                 See SAS GRAPH documentation
                 (search sas graph line types and scroll to Specifying Line Types).
COLOR1=black,
COLOR2=red,
COLOR3=tan,
COLOR4=lib,  (light blue)
COLOR5=gold,
COLOR6=violet,
COLOR7=pink,
COLOR8=ligr,  (light gray)
COLOR9=libr,  (light brown)
COLOR10=pagr,  (pale gray)  The colors for the lines used.
               NOTE:  Because of red-green color-blindness, we
               have not included any shades of green.
               NOTE:  The actual shades of pale gray and light gray vary
               by device used.
PRINTCV=F,  Whether to print the variance-covariance matrix of the coefficients
            OPTIONAL
PRINTPOINTS=,  A space-delimited list of values for EXPOSURE for which you want to print ou
               the estimates and 95% confidence limits.
               You might use this for values you want to describe in
               the text of a paper or to put in a table.
               OPTIONAL
NG=500, The number of points to make for plotting.
```

```
         This works fine.  no reason to change it.
         OPTIONAL

PARAMETERS RELATING TO THE SMOOTHED HISTOGRAM
=============================================
The 2 parameters below relate to the level of smoothing.
If there are no problems making the graph, these should be left at their defaults.
BWM=1,  The band width multiplier.  Raising this makes the graph
        more smoothe
        OPTIONAL
DISTMETH=SJPI,  The smoothing method for making the smoothed histogram.
                Other options are OS and SROT.
                OPTIONAL

PARAMETERS RELATING TO TEXT OUTPUT OF PLOTTING POINTS
=====================================================
PLOTDATA=&EXPOSURE..&OUTCOME..SEL&SELECT._&NK,
              The name of the file of plotting points.
              We strongly suggest that you give this file a
              mnemonic name to avoid overwriting.
              On a UNIX system, this file will be in the same
              directory as the program.
              OPTIONAL
COMMTYPE=1,
             OPTIONAL
FILEMODE=mod,
             OPTIONAL
```

3 Examples

The first three examples are from the same study of BMI and ovulatory infertility used in the documentation for %LGTPHCURV9. Since this is a case-control study, we need to use the LOGIT link and plot the odds ratio. Also, since the unit of analysis is the pregnancy, and we do not want to use the empirical variance, even though some subjects may have more than one pregnancy included in the study, we have to 'fool' the macro into using ordinary `PROC GENMOD` by not telling it the *SUBJECT* variable name.

3.1 Example 1. A minimal call to the macro

```
title2 'example 1--bare bones';
%glmcurv9(data=merge0, exposure=BMI, outcome=case,
          dist=bin, link=logit, plotor=T,
          reptype=ind,
  adj= age2 age3 age4 period2 period3);
```

You get a WARNING from SAS,

```
WARNING:  The variable SCALE in the DROP, KEEP, or RENAME list has never been
```

```
referenced.
```

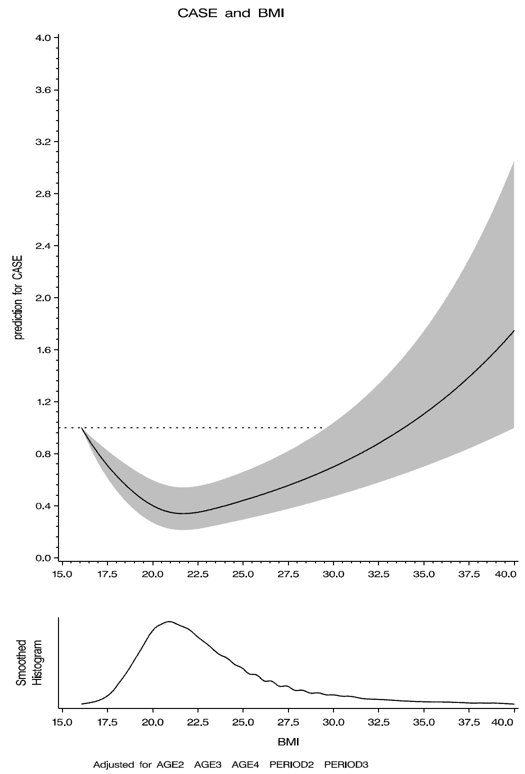
You can safely ignore this.

The results are

```
--------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv Program example1-6 07OCT2014 18:03                    1
example 1--bare bones
Percent of range of BMI below the first knot is 11  .
Percent of range of BMI above the last knot  is 38  .

--------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv Program example1-6 07OCT2014 18:03                    2
example 1--bare bones
    Knots for BMI:
    18.64 21.14 23.52 31.02

--------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv Program example1-6 07OCT2014 18:03                    3
example 1--bare bones


values of spline variables when BMI is 16.07000000 and for extrapoints, if any

  Obs     BMI     BMI1    BMI2

27103    16.07      0       0
27604    16.07      0       0

--------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv Program example1-6 07OCT2014 18:03                    4
example 1--bare bones
27102 subjects with 27102 observations
model with adjusters only

Obs VARNAME   DF    COEFF   STDERR LOWERWALDCL UPPERWALDCL   CHISQ      P

 1  Intercept  1  -3.3729   0.0659   -3.5021     -3.2437   2618.30 <.0001
 2  AGE2       1  -0.2041   0.0813   -0.3634     -0.0448      6.31 0.0120
 3  AGE3       1  -0.2336   0.1075   -0.4443     -0.0230      4.73 0.0297
 4  AGE4       1   0.5816   0.1774    0.2339      0.9293     10.75 0.0010
 5  PERIOD2    1   0.1406   0.0798   -0.0159      0.2971      3.10 0.0782
 6  PERIOD3    1  -0.0415   0.0988   -0.2353      0.1522      0.18 0.6744
 7  Scale      0   1.0000   0.0000    1.0000      1.0000        _    _

--------------------------------------------------------------------------------
```

```
/udd/stleh/doctn/glmcurv Program example1-6 07OCT2014 18:03                    5
example 1--bare bones
27102 subjects with 27102 observations
model with linear exposure

Obs VARNAME   DF    COEFF   STDERR LOWERWALDCL UPPERWALDCL   CHISQ      P

 1  Intercept  1  -4.6925   0.1989   -5.0823     -4.3026    556.59 <.0001
 2  BMI        1   0.0573   0.0080    0.0416      0.0730     51.30 <.0001
 3  AGE2       1  -0.2243   0.0814   -0.3839     -0.0648      7.59 0.0059
 4  AGE3       1  -0.2798   0.1078   -0.4910     -0.0686      6.74 0.0094
 5  AGE4       1   0.5219   0.1779    0.1732      0.8705      8.61 0.0034
 6  PERIOD2    1   0.1280   0.0799   -0.0287      0.2847      2.56 0.1094
 7  PERIOD3    1  -0.0780   0.0991   -0.2722      0.1162      0.62 0.4311
 8  Scale      0   1.0000   0.0000    1.0000      1.0000       _    _

--------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv Program example1-6 07OCT2014 18:03                    6
example 1--bare bones
27102 subjects with 27102 observations
model with splines, if any

Obs VARNAME   DF    COEFF   STDERR LOWERWALDCL UPPERWALDCL   CHISQ      P

  1 Intercept  1   1.2414   1.0421   -0.8010      3.2839      1.42 0.2335
  2 BMI        1  -0.2376   0.0527   -0.3410     -0.1343     20.31 <.0001
  3 BMI1       1   1.4023   0.3517    0.7129      2.0916     15.89 <.0001
  4 BMI2       1  -2.8876   0.8168   -4.4884     -1.2867     12.50 0.0004
  5 AGE2       1  -0.2091   0.0816   -0.3690     -0.0492      6.57 0.0104
  6 AGE3       1  -0.2545   0.1081   -0.4663     -0.0426      5.54 0.0185
  7 AGE4       1   0.5610   0.1785    0.2111      0.9109      9.87 0.0017
  8 PERIOD2    1   0.1342   0.0801   -0.0227      0.2911      2.81 0.0936
  9 PERIOD3    1  -0.0748   0.0993   -0.2693      0.1198      0.57 0.4514
 10 Scale      0   1.0000   0.0000    1.0000      1.0000       _    _

--------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv Program example1-6 07OCT2014 18:03                    7
example 1--bare bones
27102 subjects with 27102 observations
    Number of observations in the whole data set:  27102

    Dependent variable: CASE
    Exposure: BMI
    Range of exposure in data used:  16.07 to 39.99
    Number of knots: 4
    You chose to use all 2 spline variables: BMI1 BMI2
    Adjusted for:
         AGE2  AGE3  AGE4  PERIOD2  PERIOD3

    The DISTRIBUTION (DIST) is B and the LINK function is LOGIT

    CASE and BMI
    Name of graph file:  MERGE0.BMI.CASE.sel1_4.PS
    Graph option:  SPLINE



    Line Test Name     Description                          P value
    -------------------------------------------------------------


    1    Test for      If the P value is small, the
         curvature     relationship between the
         (i.e. non-    exposure and the outcome, if any,
          linear       is non-linear.
          relation)    SEE LINE 2.
                       If the P value is large, the
                       relationship between the
                       exposure and the outcome, if any
                       is linear
                       SEE LINE 3.
              Using likelihood ratio test, p-value is:     <.0001
    -------------------------------------------------------------
    2    Test for      If LINE 1 indicated a possible
         overall sig-  non-linear relation between the
         nificance     exposure and the outcome,
         of the curve  use this P value to express the
                       the significance of the exposure-outcome
              Using likelihood ratio test, p-value is:     <.0001
    -------------------------------------------------------------
    3    Test for      If the result of LINE 1 indicated
         linear        no significant nonlinearity between
         relation      the exposure and the outcome,
                       use this P value AND rerun your
                       model with the parameter
                       PWHICH=LINEAR, to get the graph
                       corresponding to the model of
                       interest (if you intend to use
                       the graph).
              Using likelihood ratio test, p-value is:     <.0001
```

The output gives the graph name, which is the default.
The graph is



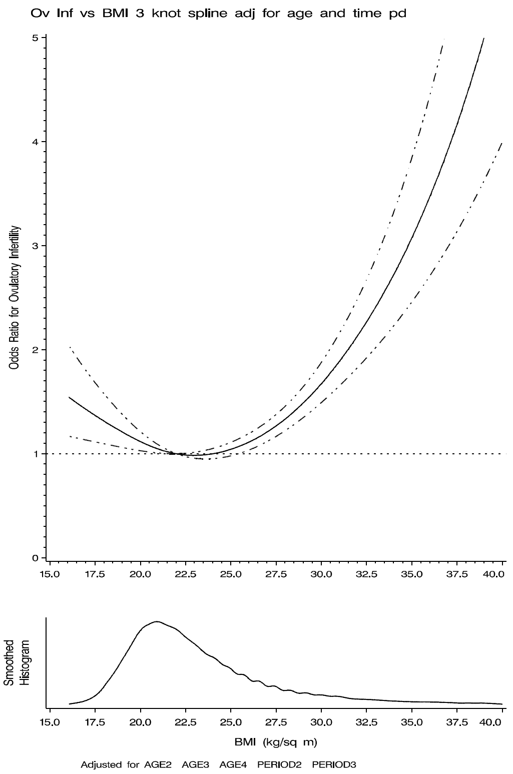
Since we did not give a reference value, the macro used the default, *MIN*.

3.2 Example 2. A more specified call with NK=3

Here we changed the reference value to 22. The macro call is

```
%glmcurv9(data=merge0, refval=22, exposure=BMI, outcome=case,
          reptype=ind,
          dist=bin, link=logit, plotor=t,
          pictname=example2.ps, ci=F,
          axordv=0 to 5 by 1,
          hlabel=%quote(BMI (kg/sq m)), nk=3,
          vlabel=Odds Ratio for Ovulatory Infertility,
          adj= age2 age3 age4 period2 period3,
          graphtit=Ov Inf vs BMI 3 knot spline adj for age and time pd,
          testrep=short,
          modprint=f);
```

and the graph is



The graph is (roughly) the same shape as in Example 1, but has moved up because we set *REFVAL*=22. Since *CI* is not equal to T, we got dotted lines for the confidence limits instead of a confidence cloud.

3.3 Example 3. Using 5-knot splines, USERSPLV, making a file of plotting points, and showing the use of PRINTPOINTS

In this and future examples we will use *TESTREP*=SHORT.
The macro call is

```
title2 'example 3--5 knot spline, given knots';
%glmcurv9(data=merge0,  refval=22, exposure=BMI, outcome=case,
          reptype=ind,
          link=logit, dist=bin,
          plotor=t,
          pictname=example3a.ps, outplot=ps,
          hlabel=%quote(BMI (kg/sq m)), knot=18.64  21.14  23.52  31.02  36.5,
          vlabel=Odds Ratio for Ovulatory Infertility,
          vlabelstyle=h,
          printpoints=18.5 21.75 25 27.5 30 32.5 35 37.5,
          axlabmult=1.5, axvalmult=1.1,
          adj= age2 age3 age4 period2 period3, select=1, plot=4,
          graphtit=Ov Inf vs BMI 5 knot spline adj for age and time pd,
          testrep=short,
          axordv=0 to 5 by .5);
data plotdat;  infile '/udd/stleh/doctn/glmcurv/BMI.CASE.sel1_5' missover firstobs=2;
input mvarbl bmi estimate lower upper;
```

```
run;
title3 'first 5 observations of the plotting points file';
proc print data=plotdat (obs=5);  run;
```

and the first 5 plotting points are

```
--------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv Program example1-6 01DEC2011 15:22                    9
example 3--5 knot spline, given knots
first 5 observations of the plotting points file

 Obs    MVARBL      BMI      ESTIMATE          LOWER      UPPER

   1       1      16.0700    0.078232    .003016999    0.70417
   2       1      16.1178    0.077514    .002970151    0.70328
   3       1      16.1657    0.076802    .002924006    0.70238
   4       1      16.2135    0.076096    .002878556    0.70148
   5       1      16.2614    0.075396    .002833788    0.70058
```

The graphed values for the levels of BMI in *PRINTPONTS* are

```
--------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv Program example1-6 01DEC2011 15:22                    8
example 3--5 knot spline, given knots
20516 subjects with 27102 observations
Predicted values of CASE and 95% confidence bounds at designated values of BMI
for MVARBL = 1

                              lower 95%     upper 95%
                             confidence    confidence
                              bound for     bound for
                predicted     predicted     predicted
Obs     BMI       value         value         value

502    18.50     1.63251       1.30629       2.04018
503    21.75     1.00138       0.98549       1.01753
504    25.00     1.23759       1.04945       1.45946
505    27.50     1.64855       1.38231       1.96608
506    30.00     2.20710       1.78329       2.73162
507    32.50     2.76968       2.20173       3.48413
508    35.00     3.23774       2.55611       4.10113
509    37.50     3.68193       2.62679       5.16092
510    22.00     1.00000       1.00000       1.00000

--------------------------------------------------------------------------------
```

3.4 Example 4. An example with multiple observations per subject

This example comes from the Trial of Vitamins in HIV-positive pregnant women in Tanzania. We look at the relation of CD4 count (a measure of immune function) and age in their children. The macro call is

```
%glmcurv9(data=cd4kid, subject=idno2, outcome=cd4c, exposure=agemon,
 distmeth=os, displayx=t, bwm=5,
select=3, dist=n, link=id,
knot=0 6 12  24 36 48,
pictname=cd4.glmsel.6kn.ps,
testrep=short,
maxstep=2,
plotor=f);
```

Since we had trouble with the smoothed histogram when we used the default options, we switched to *DISTMETH*=OS. The output is

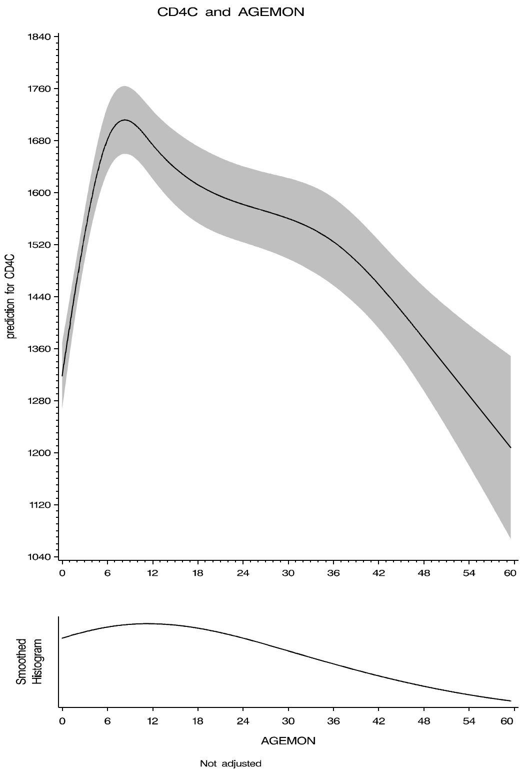
```
------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   1
Percent of range of AGEMON below first knot is 0  .
Percent of range of AGEMON above last knot is 19  .

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   2
    Knots for AGEMON:
    0 6 12 24 36 48

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   3



values of spline variables when AGEMON is 0.00000000 and for extrapoints, if a

 Obs    AGEMON    AGEMON1    AGEMON2    AGEMON3    AGEMON4

3571       0         0          0          0          0
4072       0         0          0          0          0

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   4

809 subjects with 3570 observations
model with adjusters only
```

```
Obs    VARNAME       COEFF     STDERR    LOWERCL    UPPERCL         Z        P

 1    Intercept   1533.097    18.8536   1496.145   1570.050     81.32   <.0001

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   5

809 subjects with 3570 observations
model with linear exposure

Obs   VARNAME        COEFF     STDERR    LOWERCL    UPPERCL         Z        P

 1    Intercept   1543.759    19.8856   1504.784   1582.734     77.63   <.0001
 2    AGEMON       -0.7977     0.9148    -2.5906     0.9952     -0.87   0.3832

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   6

809 subjects with 3570 observations
automatic gee selection
Step  0 :  variable AGEMON1  added

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   7

809 subjects with 3570 observations
automatic gee selection
Step 1 :  No variable dropped.

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   8

809 subjects with 3570 observations
automatic gee selection
Step  2 :  variable AGEMON2  added

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014   9

809 subjects with 3570 observations
automatic gee selection
Step 3 :  No variable dropped.

------------------------------------------------------------------------------
```

```
The SAS System                              13:55 Tuesday, October 7, 2014  10

809 subjects with 3570 observations
automatic gee selection
Step  4 :  variable AGEMON3  added

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014  11

809 subjects with 3570 observations
automatic gee selection
Step 5 :  No variable dropped.

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014  12

809 subjects with 3570 observations
automatic gee selection
Step  6 :   no variable added

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014  13

809 subjects with 3570 observations
automatic gee selection
Step 7 :  No variable dropped.

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014  14

809 subjects with 3570 observations
model with splines, if any

Obs   VARNAME        COEFF     STDERR    LOWERCL    UPPERCL          Z         P

 1    Intercept   1318.023    26.1281   1266.813   1369.234      50.44    <.0001
 2    AGEMON       76.3522     7.7337    61.1944    91.5099       9.87    <.0001
 3    AGEMON1     -1012.46   125.7902   -1259.00   -765.915      -8.05    <.0001
 4    AGEMON2     2120.408   286.3144   1559.242   2681.574       7.41    <.0001
 5    AGEMON3     -1148.09   176.2403   -1493.51   -802.662      -6.51    <.0001

------------------------------------------------------------------------------

The SAS System                              13:55 Tuesday, October 7, 2014  15

809 subjects with 3570 observations
```

```
Number of observations in the whole data set:  3570

Dependent variable: CD4C
Exposure: AGEMON
Range of exposure in data used:  0 to 59.539473684
Number of knots: 6
Spline variable(s) selected by stepwise: AGEMON1 AGEMON2 AGEMON3
Not adjusted

The DISTRIBUTION (DIST) is N and the LINK function is ID

CD4C and AGEMON
Name of graph file:  cd4.glmsel.6kn.ps
Graph option:  SPLINE



Line Test Name                                          P value
-------------------------------------------------------------
1    Test for curvature (i.e. non-linear relation)  <.0001

2    Test for overall significance of curve       <.0001
3    Test for linear relation                     0.3866
```

The graph is



3.5 Example 5. An example with multiple observations per person and MDEP(1) working covariance structure

This example and the next are from the trial of vitamins in HIV-negative pregnant women in Tanzania. It investigates whether the relationship between energy intake (ascertained from 24 hour recalls) and weight gain (in grams per 4 weeks) is nonlinear.
The covariate `difftimec4` is the number of weeks between the measurements -4 (since the average time between visits was 4 weeks).
The first time we ran the macro with nk=4, 34% of the range of the exposure was above the highest knot. We therefore added a knot at 3600.

```
%glmcurv9(data=final, subject=idno2, reptype=mdep(1), outcome=diffwt,
dist=n, link=id,
refval=2200,
knot=1185 1899 2344 3056 3600,
pictname=wtgainpreg.energy.ps,
axordv=600 to 1800 by 100,
vlabel=%str(Predicted 4-week weight gain during pregnancy (grams) ),
axordh=0 to 5000 by 500, hlabel=%str(Energy intake (kcal) ),
testrep=short,
cutoff=2 1800,
exposure=energy, adj=difftimec4, adjdat=adjdat);
```

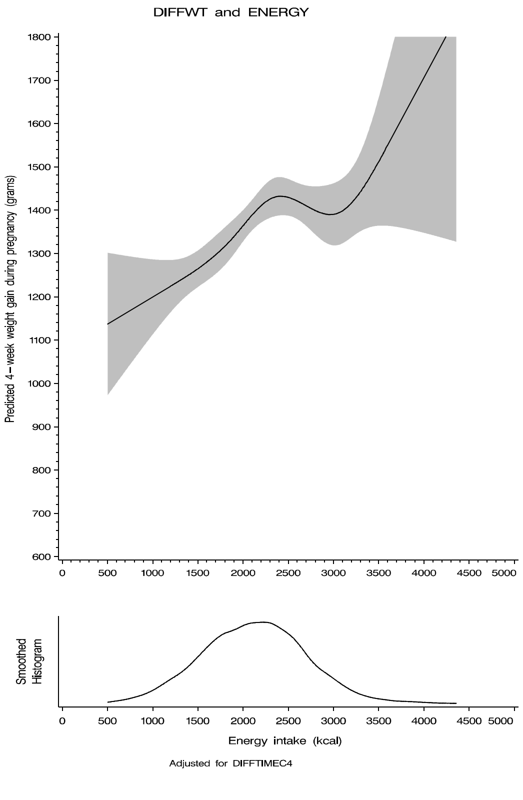
The output is

```
------------------------------------------------------------------------------


/udd/stleh/doctn/glmcurv  Program paper1   12JUN2012   13:45    stleh        2
Percent of range of ENERGY below first knot is 18   .
Percent of range of ENERGY above last knot is 20   .


------------------------------------------------------------------------------


/udd/stleh/doctn/glmcurv  Program paper1   12JUN2012   13:45    stleh        3
    Knots for ENERGY:
    1185 1899 2344 3056 3600


------------------------------------------------------------------------------


/udd/stleh/doctn/glmcurv  Program paper1   12JUN2012   13:45    stleh        4




values of spline variables when ENERGY is 2200 and for extrapoints, if any

  Obs     ENERGY     ENERGY1     ENERGY2     ENERGY3

14564      2200      179.293     4.67590        0

------------------------------------------------------------------------------
```

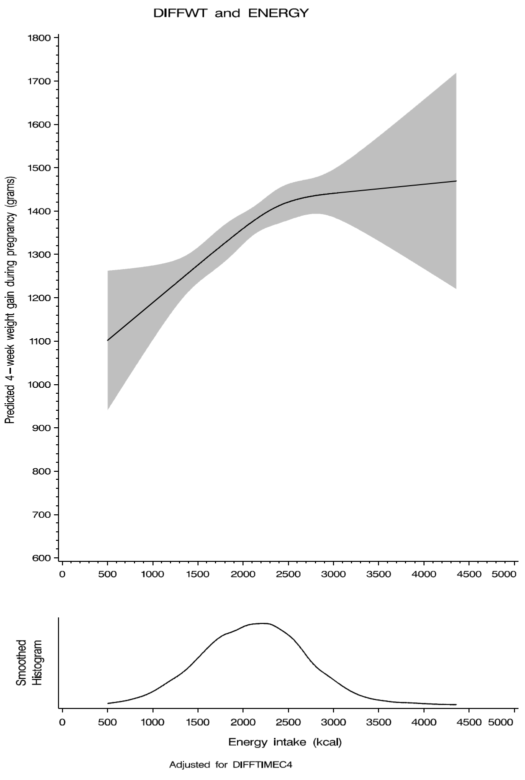
```
/udd/stleh/doctn/glmcurv  Program paper1   12JUN2012   13:45    stleh        5

6761 subjects with 14062 observations
model with adjusters only

Obs   VARNAME       COEFF    STDERR   LOWERCL   UPPERCL        Z       P

 1   Intercept   1360.842   12.5110  1336.321  1385.363   108.77  <.0001
 2   DIFFTIMEC4  245.4710   46.1888  154.9426  335.9994     5.31  <.0001

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   12JUN2012   13:45    stleh        6

6761 subjects with 14062 observations
model with linear exposure

Obs  VARNAME        COEFF    STDERR   LOWERCL   UPPERCL        Z       P

 1   Intercept   1105.471   48.3168  1010.771  1200.170    22.88  <.0001
 2   ENERGY        0.1203    0.0220    0.0772    0.1635     5.47  <.0001
 3   DIFFTIMEC4  242.8707   46.1560  152.4066  333.3349     5.26  <.0001

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   12JUN2012   13:45    stleh        7

6761 subjects with 14062 observations
model with splines, if any

Obs  VARNAME        COEFF    STDERR   LOWERCL   UPPERCL        Z       P

 1   Intercept   1073.936  125.9471  827.0843  1320.788     8.53  <.0001
 2   ENERGY        0.1255    0.0856   -0.0423    0.2933     1.47  0.1426
 3   ENERGY1       0.4241    0.4476   -0.4532    1.3013     0.95  0.3434
 4   ENERGY2      -3.4259    2.1475   -7.6349    0.7831    -1.60  0.1106
 5   ENERGY3       5.9696    3.0865   -0.0798   12.0191     1.93  0.0531
 6   DIFFTIMEC4  241.3327   46.1554  150.8698  331.7955     5.23  <.0001

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   12JUN2012   13:45    stleh        8

6761 subjects with 14062 observations
    Number of observations in the whole data set:  14062

    Dependent variable: DIFFWT
    Exposure: ENERGY
    Range of exposure in data used:  502.4 to 4357.65
```

```
Number of knots: 5
You chose to use all 3 spline variables: ENERGY1 ENERGY2 ENERGY3
Adjusted for:
     DIFFTIMEC4

The DISTRIBUTION (DIST) is N and the LINK function is ID

DIFFWT and ENERGY
Name of graph file:  wtgainpreg.energy.ps
Graph option:  SPLINE

Line Test Name                                          P value
-------------------------------------------------------------
1    Test for curvature (i.e. non-linear relation)  0.0712

2    Test for overall significance of curve       <.0001
3    Test for linear relation                     <.0001
```

Note that we did not use automatic selection.
The graph is



Since the test for nonlinearity is not significant, we would be unlikely to include this graph in a paper.
For comparison, we show the graph obtained from a call to the macro using the default 4 knots.



The LINE 1 p-value for this analysis was .25.

3.6 Example 6. Using BYVAR, AXLABMULT, and other customizations, as well as automatic selection without specifying the number of knots

As an example, we will stratify the dataset used in Example 5 by BMI at entry to the study. The macro call is

```
%glmcurv9(data=final, subject=idno2, reptype=mdep(1), outcome=diffwt,
dist=n,
select=3,
refval=2000,
printpoints=1500 1800 2000 2200 2500,
byvar=bmi1825,
where=bmi1825 ne .,
shortlegend=t,
leglab1=BMI ge 25, leglab2=BMI lt 25,
graphtit=Monthly weight gain in pregnancy by Energy intake and BMI,
footer=Adjusted for length of time interval,
pictname=wtgainpreg.enbmib.ps,
axordv=600 to 1800 by 100,
vlabel=%str(Predicted 4-week weight gain during pregnancy (grams) ),
axordh=0 to 5000 by 500,
hlabel=%str(Energy intake (kcal) ),
axlabmult=1.5,
testrep=short,
cutoff=2 1800,
exposure=energy, adj=difftimec4 , adjdat=adjdat1);
```

The output is

```
------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh        2
WARNING in macro call: since no link is provided,
      the default canonical link function for N , ID , will be used.

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh        3
Percent of range of ENERGY below the first knot is 9  .
Percent of range of ENERGY above the last knot  is 22  .

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh        4
    Knots for ENERGY:
    839.91667 1131.65 1354.28 1508.183 1626.15417 1724.1625 1808.9324 1892.28

1977.0625 2050.9125 2126.5575 2202.06875 2272.875 2343.65813 2418.20833 2499.7
625
    2581.4975 2682.55 2823.93333 3010.52625 3520.9225

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh        5



values of spline variables when ENERGY is 2000 and for extrapoints, if any


              E       E       E       E       E       E       E       E
        E     N       N       N       N       N       N       N       N
        N     E       E       E       E       E       E       E       E
        E     R       R       R       R       R       R       R       R
  O     R     G       G       G       G       G       G       G       G
  b     G     Y       Y       Y       Y       Y       Y       Y       Y
  s     Y     1       2       3       4       5       6       7       8

12726 1500  40.013   6.953   0.430   0.000  0.0000  0.0000  0.0000  0.0000
12727 1800 123.121  41.535  12.319   3.457  0.7310  0.0607  0.0000  0.0000
12728 2000 217.206  91.094  37.457  16.551  7.2691  2.9199  0.9704  0.1739
12729 2200 350.027 169.647  84.156  46.066 26.2900 14.9893  8.3207  4.0539
12730 2500 636.494 356.448 209.238 135.737 92.8344 64.9706 45.9162 31.2259
12731 2000 217.206  91.094  37.457  16.551  7.2691  2.9199  0.9704  0.1739
```

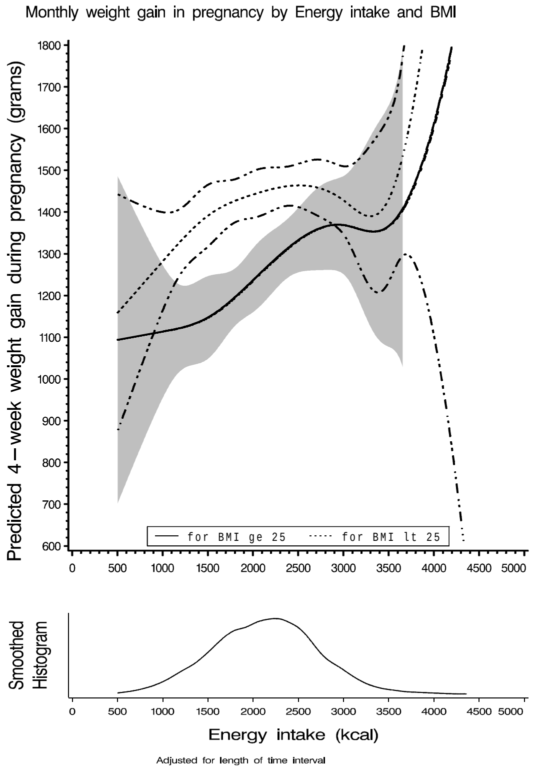
```
                 E       E       E       E       E        E        E     E E E
         E       N       N       N       N       N        N        N     N N N
         N       E       E       E       E       E        E        E     E E E
         E       R       R       R       R       R        R        R     R R R
         R       G       G       G       G       G        G        G     G G G
  O      G       Y       Y       Y       Y       Y        Y        Y     Y Y Y
  b      Y       1       1       1       1       1        1        1     1 1 1
  s      9       0       1       2       3       4        5        6     7 8 9

12726  0.0000  0.0000 0.00000 0.00000 0.00000 0.00000 0.000000         0 0 0 0
12727  0.0000  0.0000 0.00000 0.00000 0.00000 0.00000 0.000000         0 0 0 0
12728  0.0017  0.0000 0.00000 0.00000 0.00000 0.00000 0.000000         0 0 0 0
12729  1.5415  0.4610 0.05511 0.00000 0.00000 0.00000 0.000000         0 0 0 0
12730 19.8955 12.6008 7.24562 3.67919 1.63004 0.53166 0.076126 1.8638E-9 0 0 0
12731  0.0017  0.0000 0.00000 0.00000 0.00000 0.00000 0.000000         0 0 0 0

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh        6
For the stratum bmi1825=0 (BMI ge 25)
2298 subjects with 4675 observations
model with adjusters only

Obs   VARNAME       COEFF    STDERR   LOWERCL   UPPERCL        Z       P

 1   Intercept   1254.305   23.0996  1209.030  1299.579    54.30  <.0001
 2   DIFFTIMEC4  294.0854   84.3627  128.7375  459.4332     3.49  0.0005

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh        7
For the stratum bmi1825=0 (BMI ge 25)
2298 subjects with 4675 observations
model with linear exposure

Obs  VARNAME        COEFF    STDERR   LOWERCL   UPPERCL        Z       P

 1   Intercept   951.8493   85.6149  784.0472  1119.651    11.12  <.0001
 2   ENERGY        0.1436    0.0398    0.0657    0.2216     3.61  0.0003
 3   DIFFTIMEC4  288.7176   84.5548  122.9931  454.4420     3.41  0.0006

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh        8
For the stratum bmi1825=0 (BMI ge 25)
2298 subjects with 4675 observations
automatic gee selection
Step  0 :   no variable added

------------------------------------------------------------------------------
```

```
/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh        9
For the stratum bmi1825=0 (BMI ge 25)
2298 subjects with 4675 observations
automatic gee selection
stepwise procedure cannot add or drop more variables.

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       10
For the stratum bmi1825=0 (BMI ge 25)
2298 subjects with 4675 observations
    No spline variables are selected by the current criteria.
    You can either change the parameter values for sls, sle or nk, or
    bear in mind that the only valid test is the linear test.
    The graph output will be the linear graph.

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       11
For the stratum bmi1825=0 (BMI ge 25)
2298 subjects with 4675 observations
    Number of observations in the whole data set:  12224
    number of observations in the stratum BMI1825=1: .

    Dependent variable: DIFFWT
    Exposure: ENERGY
    Range of exposure in data used:  502.4 to 4357.65
    Number of knots: 21
    No spline variable selected by the current criteria
    Adjusted for:
         DIFFTIMEC4

    The DISTRIBUTION (DIST) is N and the LINK function is ID

    Monthly weight gain in pregnancy by Energy intake and BMI
    Name of graph file:  wtgainpreg.enbmib.ps
    Graph option:  LINEAR

    No spline variables were selected.  There is no spline model.
    The p-values for lines 1 and 2 will therefore be missing.


    Line Test Name                                      P value
    -------------------------------------------------------------
    1    Test for curvature (i.e. non-linear relation)  .

    2    Test for overall significance of curve       .
    3    Test for linear relation                     0.0003
```

```
-------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       12
For the stratum bmi1825=0 (BMI ge 25)
2298 subjects with 4675 observations
Predicted values of DIFFWT and 95% confidence bounds at designated values of E
for bmi1825 = 0

                               lower 95%     upper 95%
                              confidence    confidence
                               bound for     bound for
                 predicted     predicted     predicted
Obs    ENERGY      value         value         value

502     1500      1167.28       1103.76       1230.80
503     1800      1210.36       1160.46       1260.26
504     2000      1239.09       1193.61       1284.56
505     2200      1267.81       1221.64       1313.98
506     2500      1310.90       1254.88       1366.91
507     2000      1239.09       1193.61       1284.56

-------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       13
For the stratum bmi1825=1 (BMI lt 25)
3618 subjects with 7549 observations
model with adjusters only

Obs   VARNAME       COEFF    STDERR   LOWERCL   UPPERCL        Z       P

 1   Intercept   1430.817   16.1385  1399.186  1462.448    88.66  <.0001
 2   DIFFTIMEC4  258.7568   59.9751  141.2078  376.3058     4.31  <.0001

-------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       14
For the stratum bmi1825=1 (BMI lt 25)
3618 subjects with 7549 observations
model with linear exposure

Obs  VARNAME        COEFF    STDERR   LOWERCL   UPPERCL        Z       P

 1   Intercept   1292.287   63.2840  1168.252  1416.321    20.42  <.0001
 2   ENERGY        0.0645    0.0285    0.0085    0.1204     2.26  0.0239
 3   DIFFTIMEC4  258.1707   59.9108  140.7477  375.5937     4.31  <.0001

-------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       15
For the stratum bmi1825=1 (BMI lt 25)
```

```
3618 subjects with 7549 observations
automatic gee selection
Step  0 :   no variable added

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       16
For the stratum bmi1825=1 (BMI lt 25)
3618 subjects with 7549 observations
automatic gee selection
stepwise procedure cannot add or drop more variables.

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       17
For the stratum bmi1825=1 (BMI lt 25)
3618 subjects with 7549 observations
    No spline variables are selected by the current criteria.
    You can either change the parameter values for sls, sle or nk, or
    bear in mind that the only valid test is the linear test.
    The graph output will be the linear graph.

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       18
For the stratum bmi1825=1 (BMI lt 25)
3618 subjects with 7549 observations
    Number of observations in the whole data set:  12224
    number of observations in the stratum BMI1825=2: .

    Dependent variable: DIFFWT
    Exposure: ENERGY
    Range of exposure in data used:  502.4 to 4357.65
    Number of knots: 21
    No spline variable selected by the current criteria
    Adjusted for:
         DIFFTIMEC4

    The DISTRIBUTION (DIST) is N and the LINK function is ID

    Monthly weight gain in pregnancy by Energy intake and BMI
    Name of graph file:  wtgainpreg.enbmib.ps
    Graph option:  LINEAR

    No spline variables were selected.  There is no spline model.
    The p-values for lines 1 and 2 will therefore be missing.


    Line Test Name                                       P value
    -------------------------------------------------------------
    1    Test for curvature (i.e. non-linear relation)  .

    2    Test for overall significance of curve        .
    3    Test for linear relation                      0.0243

------------------------------------------------------------------------------

/udd/stleh/doctn/glmcurv  Program paper1   07OCT2014   18:02    stleh       19
For the stratum bmi1825=1 (BMI lt 25)
3618 subjects with 7549 observations
Predicted values of DIFFWT and 95% confidence bounds at designated values of E
for bmi1825 = 1

                                  lower 95%     upper 95%
                                 confidence    confidence
                                  bound for     bound for
                      predicted   predicted     predicted
Obs     ENERGY          value       value         value

502      1500         1388.98      1341.03       1436.94
503      1800         1408.32      1371.29       1445.35
504      2000         1421.22      1388.58       1453.85
505      2200         1434.11      1402.34       1465.88
506      2500         1453.45      1416.10       1490.80
507      2000         1421.22      1388.58       1453.85
```

The graph is



Since the macro did not select any spline variables in either stratum of BMI, it plotted the linear relationships.

4 Computational Methods

4.1 Automatic knot placement, given a desired number of knot points

If you specify a number of knots (*NK*), the macro will automatically determine the appropriate percentiles of the data and place the knots there. If you request automatic spline variable selection (*SELECT*=3) and have not given *NK*, the macro will set *NK*=21. If you do not give *NK* or *KNOT* and do not use automatic spline variable selection, the macro will set *NK*=4. As always, this can be overridden by providing a list of knot locations.

```
NK  Knot locations as percentiles of EXPOSURE
--  ----------------------------
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 18.3333 34.1667 50 65.8333 81.6667 97.5
8   1 15 29 43 57 71 85 99
9   2 14 26 38 50 62 74 86 98
10   2 12.6667 23.3333 34 44.6667 55.3333 66
     76.6667 87.3333 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
```

4.2 Computation of the spline variables:

Let t_j be the jth knot point.

Let $kd = (t_{nk} - t_1)^{2/3}$, 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 jth spline variable (j runs from 1 to NK-2) is given by

$$x_j = max((x - t_j)/kd, 0)^3$$
$$+(t_{nk-1} - t_j) * max((x - t_{nk})/kd, 0)^3$$
$$-(t_{nk} - t_j) * max((x - t_{nk-1})/kd, 0)^3)/(t_{nk} - t_{nk-1})$$

For $x < t_j$ the value of the jth 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.

Note that the value of x_j depends on the values of the first, *nk*th, and *nk-1*st knots. That is why the value of the spline variable depends on the locations of knots other than the *j*th knot.

4.3 Default bandwidth for smoothing:

The default bandwidth is data-specific. Let N be the size of the data set, and STD be the standard deviation of the exposure variable (X) in the data set.

$$bandwidth = (STD/1.349) * (4/3N)^{0.2}$$

Although the user can set the bandwidth (using *BWM*), it is usually fine to let the macro do it automatically. See Frequently Asked Questions below.

4.4 Outline of stepwise model selection:

If you request automatic spline variable selection (em SELECT=3) and have not specified *NK*, the macro will set *NK*=21. As always, this can be overridden by providing a list of knot locations.

The default *SLE* and *SLS* for automatic selection are .05, but the user may specify other values.

In the discussion below, all mentions of 'likelihood' should be interpreted to mean 'partial likelihood' when Cox models are used.

Each step starts with a 'base' model. For the first step, the 'base' model includes the linear term and all the adjusters. For subsequent steps, the 'base' model includes the above plus whatever spline variables are in the model by the end of the step.

For a forward step, each of the spline variables not in the base model is added (singly) to the base model and a likelihood is computed. If, for the spline variable giving the biggest likelihood (i.e. the biggest change from the base model), the likelihood ratio test (LRT) gives a p-value meeting the criterion for entry into the model (SLE), that spline variable is added to the model. Otherwise, no variable is added to the model.

For a backward step, each of the spline variables in the base model is deleted (singly) from the base model, and a likelihood is computed. If, for the spline variable giving the biggest likelihood (i.e. the closest to the base model), the LRT gives a p-value greater than the criterion for staying in the model (SLS), that spline variable is dropped from the model. Otherwise, no variable is dropped. If a variable is dropped, the macro uses this new base model and tests the remaining spline variables to see whether they can be dropped.

Forward and backward steps alternate until two (2) steps in a row do not change the model, or until the maximum number of steps is attained (default=10).

5 Including the graph in a MS-WORD document

Below are the steps for importing an encapsulated postscript file into a MS-WORD document.

```
0.  If you have a postscript graph, you may have to change the graph file's
    extension to 'eps' on the unix system.
```

```
1.  E-mail the file to yourself as an attachment, and download to your PC.
2.  Open your WORD document.
3.  The sequence of keys (at least in Windows XP and its version of WORD) is
        insert
        picture
        from file
        <locate file>
        convert file (this is a window that WORD gives you)
               encapsulated postscript
```

NOTE: Conversion from encapsulated postscript may not be installed on your computer, but it is available for Windows 95 and beyond. NOTE: When I did the above procedure the picture on my Windows screen was fuzzy. When printed, it was crisp.

If you are really having trouble, consider using one of the other formats (HTML, JPEG, CGM).

6 How should I describe this in my Methods section?

The wording below has been approved by Prof. Donna Spiegelman.

We examined the possibly non-linear relation between *insert the name of the exposure here* and *insert the name of the outcome here, such as the RR of* —- non-parametrically with restricted cubic splines [REF Durrleman and Simon]. Tests for non-linearity used the likelihood ratio test, comparing the model with only the linear term to the model with the linear and the cubic spline terms.

7 WARNINGS

Log-binomial models do not always converge, especially if they have many covariates or continuous variables. You may have to use the Poisson variance (Spiegelman and Hertzmark, AJE).

`PROC GENMOD` does not like missing values for *SUBJECT*, *WITHINVAR*. If you want to use observations with missing values for these variables, you need to use a 'fake' value for them.

If you are using a weighted model, observations with missing values for the weight variable will be deleted from the analysis.

8 Frequently Asked Questions

8.1 Q: Why does a variable named MVARBL keep showing up in my output?

A: , included for the convenience of the programmer MVARBL is the default value of *BYVAR*. It is 1 for all observations in the dataset.

8.2 Q: If the p-value for nonlinearity is greater than .05 but the p-value for the overall significance of the curve is less than .05, why can't I say that the relationship is nonlinear?

A: The overall significance of the curve is coming entirely from the linear relationship (line 3).

8.3 Q: Why does the confidence cloud stop abruptly in the middle of the graph?

A: This happens when the upper (or lower) confidence limit is out of the range of the vertical axis. To solve this for the lower bound, change the lower value in *AXORDV*. To solve this for the upper bound, you may change the upper bound or use the *CUTOFF* parameter.

8.4 Q: Why is the confidence band so wide?

A: A common reason for this is that parts of the range of the exposure have very few observations. This is most likely to occur at the extremes of the data. It shows up as long flat tails in the smoothed histogram. You can also look at the knot positions. For 3 and 4 knot splines, the outer knots are at the 5th and 95th percentile points. If these are far from the lowest and highest values in the data you are using, you may need to trim the data, either by values of the exposure (*HICUT*, *LOWCUT*) or by percentiles of the exposure distribution (*HPCT*, *LPCT*). If you cannot do that, then use the cutoff parameter.

8.5 Q: Why are the values on the horizontal axis printing out vertically?

A: You probably asked for too many major tick marks.
AXORDH should be written so that about 8 to 12 numbers will print out, such as

```
axordh=0 to 100 by 10
```

rather than

```
axordh=0 to 100 by 5
```

This could also happen if you let the macro determine the tick marks and they are not 'round.' In this case, you should see what the graph looks like and determine the horizontal axis ticking yourself.

8.6 Q: I want to plot the smoothed histogram, but the SAS .log says that the Sheather-Jones plug-in did not converge

A: Sometimes the Sheather-Jones plug-in does not work. You can try increasing the smoothing parameter (*BWM*) or using *DISTMETH*=OS.

8.7 Q: Why are the coefficients of the spline variables so large in absolute value?

A: Because the spline variables are often highly correlated, it is not unusual for the coefficients to alternate between very negative and very positive values.

8.8 Q: I got an error saying the x-origin did not leave enough space for the text.

Here is an example of the ERROR message:

```
ERROR: The specified x-origin for the left vertical axis labeled LOWER did not
       leave enough space for the text. You need to specify ORIGIN=( 2.1 INCH
       ). The graph was not produced.
```

A: This can happen when you use *VLABELSTYLE*=H, if some of the words are too long. Try hyphenating the longest words OR change the *HORIGIN* as suggested by the ERROR message. This latter will make your actual graphics area smaller to accommodate your axis label.

8.9 Q: How do I make ADJDAT?

A: If you are plotting probabilities or incidence rates, you need to have values of the covariates at which to plot them. The choice of covariate values will influence the absolute value of the probabilities or incidence rates, but not the shape of the curve. It is often convenient to use the reference levels of all the sets of indicators (or alternatively, the middle indicator), and the medians or some conventional value for the continuous variables (other than the exposure, which should not have a value in this dataset). One way to do this, especially if you have a lot of sets of indicators is as follows:

```
data adjdat;
array nums ....... ; /* the list of all the adjusters.  you can just copy it from the ADJ pa
do over nums;  nums=0;  end ;  /*  effectively sets all sets of indicators to their referen
/* special coding for continuous variables */
bmi76=25;  /* coding to a conventional cutoff */
run;
```

9 References

Durrleman, Sylvain, and Simon, Richard: Flexible regression models with cubic splines. Statistics in Medicine 8: 551-561, 1989.

Govindarajulu, U.S., Malloy, E.J., Ganguli, B., Spiegelman, D., Eisen, E.A.: The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. Intl J Biostat 5(1): Article 2, 2009.

Spiegelman, D, Hertzmark, E: Easy SAS calculations for risk or prevalence ratios and differences. AJE 162(3): 199-200, 2005.

Spiegelman, D, Hertzmark, E: The authors reply to Neogi & Zhang, Tian & Liu, and Petersen & Deddens re: 'Easy SAS calculations for risk or prevalence ratios and differences'. AJE 163(12): 1159-1161, 2006.

10 Credits

Written by Ellen Hertzmark, Ruifeng Li, Biling Hong, and Donna Spiegelman for the Channing Laboratory. Questions can be directed to Biling Hong,
`stbho@channing.harvard.edu`, (617) 432-7336.