The SAS LGTPHCURV9 Macro

Ruifeng Li, Ellen Hertzmark, Mary Louie, Linlin Chen, and Donna Spiegelman

July 3, 2011

Abstract

The %LGTPHCURV9 macro fits restricted cubic splines to unconditional logistic, pooled logistic, conditional logistic, and proportional hazards regression models to examine non-parametrically the (possibly non-linear) relation between an exposure and the odds ratio (OR) or incidence rate ratio (IRR) of the outcome of interest. It allows for controlling for covariates. It also allows stepwise selection among spline variables. The output is the set of p-values from the likelihood ratio tests for non-linearity, a linear relation, and any relation, as well as a graph of the OR, IRR, predicted cumulative incidence or prevalence, or the predicted incidence rate (IR), with or without its confidence band. The confidence band can be shown as the bounds of the confidence band, or as a "cloud" (gray area) around the OR/IRR/RR curve. In addition, the macro can display a smoothed histogram of the distribution of the exposure variable in the data being used.

Keywords: SAS version 9.1, macro, logistic regression, proportional hazards regression, Cox regression, cubic splines, nonlinearity, curves, smoothed histogram, plotting, stepwise selection, incidence rate, odds ratio, rate ratio

Contents

1 Description

%LGTPHCURV9 is a SAS macro that examines the relationship between a continuous exposure (*EXPOSURE*) and a dichotomous outcome (*CASE*) or failure time (*TIME*), controlling for covariates. It also produces publication-quality graphs of this relationship.

2 Invocation and Details

In order to run this macro, your program must know where to look for it. You can tell SAS where to look for macros by using the options

```
options nocenter ps=78 ls=80 replace formdlim='='
mautosource
sasautos=('/usr/local/channing/sasautos',
   '/proj/nhsass/nhsas00/nhstools/sasautos');
```

This will allow you to use all the SAS read macros for the data sets in **/proj/nhsass/nhsas00/nhstools/sasautos** as well as other public SAS macros, such as %PM, %INDIC3, %EXCLUDE, %MPHREG, %CALADJ, and %PCTL in **/usr/local/channing/sasautos**.

In the rest of this section, we will list all the input parameters, some of which are required and some of which are optional, but strongly suggested, and some of which are truly optional.

NOTE: if a parameter has a default value, it is given to the right of the '='.

```
REQUIRED PARAMETERS
===================
  DATA     = < the name of the input dataset,
  EXPOSURE = < variable name of the exposure of interest,
  CASE     = < variable name of the outcome (LOGISTIC) or
               censoring variable (CONDLOG or COX),
               coded 0 (non-case or censored) or 1 (case or not censored) >,
  TIME     = < censored failure time for COX or conditional logistic models >,
               NOTE: if this parameter is not null (blank), LGTPHCURV9 will
                     automatically fit a COX model.

STRONGLY SUGGESTED PARAMETERS
=============================
  OUTPLOT  = PS < type of software in which to make the plot:
               PS=encapsulated postscript (the default),
               other options are JPEG and HTML >,
  PICTNAME = < filename of the file for the plot.
               You should give a mnemonic name for this >.
               default= <name of dataset>.<exposure>.<outplot>
               If you are experimenting with number of knots, or with
               linear vs. spline plots, or with axis specifications,
               and do not give a value for PICTNAME,
               you will end up with only the most recently made graph with
               the given EXPOSURE variable >,
  REFVAL   = MIN < the reference value for the exposure variable,
                 options are data-derived MEAN, MIN, MEDIAN, MODE,
                 or a value you supply >,
  AXORDV   = < range of the vertical axis and
               tick-mark spacing for odds ratio or rate ratio
               plots, (<low> to <high> by <increment>) >,
  AXORDI   = < range of the vertical axis and tick mark spacing
               for incidence rate plots,
               (<low> to <high> by <increment>) >,
  AXORDP   = < range of the vertical axis and tick mark spacing
               for predicted probability plots,
               (<low> to <high> by <increment>) >,
  HLABEL   = < label for the horizontal axis >,
  VLABEL   = < label for the vertical axis, for odds
               ratio or rate ratio plots >,
  VLABELI  = < label for the vertical axis, for
               incidence rate plots >,
  VLABELP  = < label for the vertical axis, for
               predicted probability plots >,

PARAMETERS RELATED TO THE INPUT DATA SET
========================================
  LPCT     = < to delete from the data set used to fit the
               spline model observations with EXPOSURE
               below a percentile you specify >,
  HPCT     = < to delete from the data set used to fit the
               spline model observations with EXPOSURE
               above a percentile you specify >,
  LOWCUT   = < to delete from the data set used to fit the
               spline model observations with EXPOSURE below a
               value that you specify >,
  HICUT    = < to delete from the data set used to fit the
               spline model observations with EXPOSURE above a
               value that you specify >,
  WHERE    = < a sub-setting clause to restrict the data used >,
  EXTRAV   = < extra variables to be kept in the working data
               set, primarily for use in the WHERE parameter.
               for example if the WHERE parameter refers to a
               variable not in the EXPOSURE, ADJ, CASE, BYVAR,
               or TIME, you MUST list this variable here >,

PARAMETERS RELATED TO THE MODEL
===============================
  MODEL    = LOGISTIC < the type of model you want to run
               (LOGISTIC, CONDLOG, or COX) >,
  ADJ      = < the list of covariates >,
  STRATA   = < matching strata for COX or CONDLOG model
               (typically used for conditional logistic regression) >,
  NK       =  < number of automatically placed knots to use >,
               The only values of NK currently supported are
               3, 4, 5, 6, 7, 8, 9, 10, 17, 21, 25, and 50.
               If you do not specify a value for NK or give a list of
               knot locations in KNOT (see next), the macro will set
               NK=21 if automatic stepwise selection (SELECT=3) is
               specified or NK=4 otherwise.
  KNOT     = < values of EXPOSURE at knot points, if you want to
               specify the knots yourself >
               One situation in which you need to use KNOT= is when
               the distribution of the data is such that the
               automatic positioning of the  knot points by their
               percentile values does not yield NK distinct knots.
               This happens in NHS, with NK=4 for EXPOSURE=
               alcohol consumption.
  MODOPT   = < options to be used in the model statement, e.g.
               to control the number of iterations or the
               criteria for convergence >,
  SELECT   = 1 < whether to use all spline variables in model:
               1 = use all
               2 = use spline variables provided by user (e.g.
                   based on a previous selection procedure)
               3 = use automatic stepwise selection >,
               NOTE:  If you use SELECT=3 with automatically
               placed knots and have not specified NK,
               the macro sets NK=21.
  SLS      = .05 < p value for entry into model, if SELECT=3 >,
  SLE      = .05 < p value to stay in the model, if SELECT=3 >,
  USERSPLV = <list of spline variables to be included
              if SELECT=2 (does not include the linear
              variable EXPOSURE)>,

PARAMETERS RELATED TO THE OUTPUT
================================
  MODPRINT = T < whether to print output generated by PROC
               LOGISTIC or PROC PHREG,
               F = suppress, T = print >,
               While some of the examples in this documentation use
               MODPRINT=F, we strongly suggest that users leave
               MODPRINT at the default.
  PRINTCV  = F < whether to print the covariances of the
               coefficients in the .log >,
  TESTREP  = LONG <specify long or short report on test for
               non-linearity > (SHORT to be used only if you really
               don’t need the wordy directions).
  HEADER1  = < Description of the analysis to be used in the
               report of the test for non-linearity >,
  PRINTPOINTS = < a list of values of EXPOSURE for which you wish
                  the macro to print the numeric values of
                  what is plotted in the graph >,
  PLOTPRINT = F  <whether you want the macro to print the whole set
                  of plotting points >,

PARAMETERS RELATED TO THE GRAPH OF THE REGRESSION SPLINE
========================================================
  PLOT     = 2 < type of graphics you want:
               0 = NO PLOT (i.e. just do test for non-linearity)
               1 = PROC PLOT (prints in .saslog)
               2 = PROC GPLOT
               3 = PROC PLOT and PROC GPLOT,
               4 = text file for use with PC software or other
                   graphing programs >,
  PWHICH   = SPLINE < whether to plot results of linear or
               spline model (LINEAR or SPLINE) >,
  GRAPHTIT = < label (title) for the top of the plot.
               If the value is NONE (upper case required),
               the graph will have no title.
               If GRAPHTIT is empty, HEADER1 will be used. >,
  TITLEFONT= swissb < font to use for graph title, if any >,
  TITLEMULT= 1  < multiplier to make font of title larger or smaller>,
  FOOTER   = default < footnote for the bottom of the plot
               DEFAULT lists the first 8 covariates as listed in ADJ.
               The user might prefer to list the conceptual variables
               (e.g. age, time period, BMI).
               To avoid the footnote entirely, use NONE
              (upper case required) >,
  FOOTFONT= swiss < font for footnote, if any>,
  FOOTMULT= 1  < multiplier to change size of footnote>,
  AXLABFONT= swissb < font for axis labels >,
  AXLABMULT= 1  < multiplier to change size of axis labels >,
  AXVALFONT= swiss < font for axis values >,
  AXVALMULT= 1  < multiplier to change size of axis values >,
  PLOTORRR = T < whether to plot the confidence band >,
  CI       = 2 < type of 95% confidence intervals to be displayed
               (1=clouds, 2=dotted lines) >,
  E        = T < whether to plot the OR/RR or the log(OR/RR),
               E=T is to plot OR/RR,
               E=F is to plot log(OR/RR) >,
  AXORDVLOG10 = F  < whether to plot the vertical axis on the log10 scale.
                     This is useful if your OR or RR has a wide range
                     (and you really want to show it, as opposed to
                     trimming the data or using CUTOFF (see above) )
                     NOTE:  Setting AXORDVLOG10=T will result in a WARNING:
                     The ORDER= list on axis statement 2 was ignored because
                     the log of 0 is undefined >,
  VLABELSTYLE = V  <whether to have the vertical axis label run parallel
                   to the axis (V) or to have it print horizontally (H) >,
  ----------------------------------------------------------
  ORDATA   = < name of the dataset that contains odds
               ratio/incidence rate ratio
               and confidence interval from a categorical (or
               indicator analysis, if you want to plot this on the
               same graph as the spline model >,
  OR        = < name of odds ratio/incidence rate ratio
               variable in ORDATA >,
  OR_LOWER = < variable name of lower CI in ORDATA >,
  OR_UPPER = < name of upper CI in ORDATA >,
  X_VALUE  = < name of exposure variable in ORDATA >,
  KLINES   = F < whether to plot reference lines at the
               knot points (T or F).  This is typically used when you
               are developing the graph, but not in the final plot >,
  PLOTDEC  = F < whether to plot the decile cutoffs of EXPOSURE
               on the reference line (OR=1 or RR=1) (T or F) >,
  CUTOFF   = < a value at which to truncate the vertical axis
               in the form <type> <maximum> (e.g. 2 5)
               TYPE determines how the points with values above
               MAXIMUM will be treated.  MAXIMUM should be some
               level higher than all interesting parts of the
               graph. If TYPE is 1, then the macro truncates only the
               95% CI upper limit at MAXIMUM.
               If TYPE is 2, then the macro truncates the
               95% CI upper limit and the spline curve at MAXIMUM.
               Another options for cases where the values of the
               predicted OR/RR get extreme is to use the
               AXORDVLOG10 parameter (see below). >,
  PLOTPROB = F < whether to plot probability estimates
               from logistic models >,
  PLOTINC  = F < whether to plot incidence rates from logistic models >,
  ADJDAT   = < name of a data set with one observation having
               the values at which you wish to plot the incidence
               rates or probabilities (if PLOTINC=T or PLOTPROB=T).
               NOT necessary if you are plotting OR or RR.
               For details of how to create this dataset, see FAQ. >,
  PERLENG  = 2 < study period, or time assigned to each observation
                 when pooled logistic is used for plotting incidence
                 rates>,
  PYUNIT   = 100000 < denominator for incidence rate if PLOTINC=T.
             For example, if the study period is 2 years and you would
             like to calculate the incidence per 100K person-years,
             and suppose the estimated 2-year probability is .002,
             then the estimated incidence rate will be
              (100000*.002/2)=100 per 100K P-Y >,
  NOPST    = T < to suppress plotting of OR/RR spline transformation
               on graphics device if PLOTPROB or PLOTINC is T and PLOT>1 >,

  HORIGIN = 1.5 < location of the horizontal origin in inches
                  Leave this alone unless there is a problem.
                  See Frequently Asked Questions. >,

PARAMETERS RELATED TO DISPLAY OF EXPOSURE DISTRIBUTION
```

```
  =======================================================
    DISPLAYX = T < whether to display the distribution of the
                 exposure variable (EXPOSURE),
                 F=none,
                 T=smoothed histogram>,

    TECHNICAL PARAMETERS RELATED TO DISPLAY OF EXPOSURE DISTRIBUTION
    -----------------------------------------------------------------
    WE STRONGLY SUGGEST THAT USERS LEAVE THESE PARAMETERS AT THEIR
    DEFAULT VALUES, unless the procedure does not converge.
    ---------------
    BWM = 1 < smoothing parameter for frequency
              histogram (DISPLAYX=T only) >,
              Larger numbers usually result in more smoothed
              (less detailed) graphs>,
    DISTMETH=SJPI < method for making smoothed histogram
                 (DISPLAYX=T only),
                 SJPI=Sheather-Jones plug-in
                 SNR=simple normal reference
                 SROT=Silverman's rule of thumb
                 OS=oversmoothed
                 NOTE:  If you get a diagnostic message saying that
                 the Sheather-Jones plug-in does not converge, try
                 one of the other methods or increase BWM >,
    N_GRID   = 500 <number of intervals/grids between minimum and
                 maximum values of EXPOSURE (SMOOTH option only) >,

  MAKING A FILE FOR PC GRAPHICS, PLOT=4 ONLY
  ========================================
    PLOTDATA = <&DATA>.<&EXPOSURE>.txt
               < file name of output for PLOT=4 >,
    FILEMODE = MOD < output mode for PLOTDATA file >.

   For example,  if we use option {\em PLOT}=4 for Example 3 below (instead of {\em PLOT}=2),
    the first 5 lines of the ASCII file will look like:
```

```
BMI            Splan          lower          Upper
16.07          2.927          1.842          4.652
16.11784       2.894          1.830          4.577
16.16568       2.862          1.818          4.503
16.21352       2.829          1.807          4.430
```

Note: The first column is the exposure values; the second column is the estimated values of the linear combination of exposure plus the selected spline variables (If *PLOTORRR*=T, they are odds ratios if the model is logistic or rate ratios if *MODEL*= COX; if *PLOTPROB*=T, they are the predicted probabilities for the average time represented by one observation (e.g. 2 years in most channing studies); if *PLOTINC=T*, they are the predicted probabilities and the incidence rates); the third and fourth are the 95% lower and upper confidence values of the second column. If you use option *CUTOFF*, it is possible to have missing values (if the real values are bigger than the cutoff points).

3 Examples

Examples 1-6 and 8-10 use data from a case-control study of ovulatory infertility in NHS II. The exposures of interest are BMI and hours of vigorous exercise per week. Example 7 uses data from a study of smoking and lung cancer among current smokers in NHS. Example 11 uses data from a case-control study of CHD using blood measurements, and Example 12 shows a stratified proportional hazards model.

3.1 Example 1. Required parameters not given (to demonstrate ERROR messages)

In the following call to %LGTPHCURV9, I neglected to give values for *EXPOSURE* and *CASE*. In addition, since it calls for a CONDLOG model, the *TIME* and *STRATA* parameters are required.

The macro call is:

```
title2 ’example 1--errors in macro call’;
%lgtphcurv9(data=merge0, model=condlog, refval=22, displayX=T,
pictname=pregc21.bmi4apd, hlabel=%quote(BMI (kg / sq m)));
```

The macro printed out the following diagnostic messages in both the .log and the .lst files, then stopped.

```
================================================================================

/udd/stleh/doctn/lgtphcurv Program example1-6 13JAN2011 22:09
example 1--errors in macro call
ERR’OR in MACRO call:  You did not give one or more of the required parameters.

  You did not give a variable name to use as EXPOSURE, as required

  You did not give a variable name to use as CASE, as required

  You did not name a TIME variable,
    as required when you use model=COX or CONDLOG.

  You did not name a STRATA parameter,
    as required when you use model=CONDLOG.

================================================================================
```

3.2 Example 2. Bare bones invocation

This analysis uses no optional parameters (except MODPRINT).

The macro call is

```
title2 'example 2--bare bones';
%lgtphcurv9(data=merge0, exposure=BMI, case=case,
  adj= age2 age3 age4 period2 period3);
```

The output is

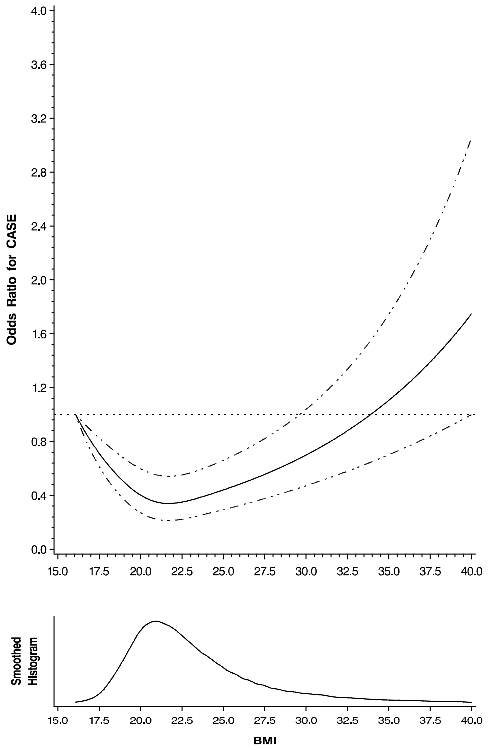
```
================================================================================


example 2--bare bones
Percent of range of BMI below the first knot is 11  .
Percent of range of BMI above the last knot  is 38  .

================================================================================


example 2--bare bones
    Knots for BMI:
    18.64 21.14 23.52 31.02

================================================================================

example 2--bare bones


values of spline variables when BMI is 16.07000000

  Obs     BMI     BMI1    BMI2

27103    16.07      0       0

================================================================================

example 2--bare bones

   Analysis of Maximum Likelihood Estimates & Odds Ratio (with adjusters only)

VARIABLE    ESTIMATE     STDERR   PROBCHISQ   ODDSRATIO   LOWERCL   UPPERCL

Intercept    -3.3729     0.0659    <.0001      0.03429    0.03013   0.03902
AGE2         -0.2041     0.0813    0.0120      0.81539    0.69532   0.95620
AGE3         -0.2336     0.1075    0.0297      0.79164    0.64129   0.97724
AGE4          0.5820     0.1774    0.0010      1.78963    1.26410   2.53363
PERIOD2       0.1406     0.0798    0.0782      1.15096    0.98425   1.34591
PERIOD3      -0.0415     0.0988    0.6744      0.95932    0.79036   1.16441

================================================================================
example 2--bare bones

Association of Predicted Probabilities and Observed Responses

measure               value       measure      value

Percent Concordant     42.7       Somers' D    0.100
Percent Discordant     32.7       Gamma        0.133
Percent Tied           24.6       Tau-a        0.006
Pairs                 21907512    c            0.550

================================================================================

example 2--bare bones

Analysis of Maximum Likelihood Estimates & Odds Ratios (linear model with adjust

VARIABLE    ESTIMATE     STDERR   PROBCHISQ   ODDSRATIO   LOWERCL   UPPERCL

Intercept    -4.6925     0.1989    <.0001      0.00916    0.00621   0.01353
AGE2         -0.2243     0.0814    0.0059      0.79904    0.68118   0.93729
AGE3         -0.2798     0.1078    0.0094      0.75592    0.61198   0.93372
AGE4          0.5224     0.1779    0.0033      1.68612    1.18987   2.38934
PERIOD2       0.1280     0.0799    0.1094      1.13654    0.97171   1.32933
PERIOD3      -0.0780     0.0991    0.4311      0.92496    0.76170   1.12322
BMI           0.0573    0.00800    <.0001      1.05899    1.04251   1.07573

================================================================================

example 2--bare bones

Association of Predicted Probabilities and Observed Responses

measure               value       measure      value

Percent Concordant     53.2       Somers' D    0.139
Percent Discordant     39.3       Gamma        0.150
Percent Tied            7.6       Tau-a        0.008
Pairs                 21907512    c            0.570

================================================================================

example 2--bare bones

Analysis of Maximum Likelihood Estimates & Odds Ratio (spline model with adjuste

VARIABLE    ESTIMATE     STDERR   PROBCHISQ   ODDSRATIO   LOWERCL   UPPERCL

Intercept     1.2414     1.0421    0.2335      3.46050    0.44885   26.6793
AGE2         -0.2091     0.0816    0.0104      0.81127    0.69139    0.9519
```

```
AGE3         -0.2545     0.1081    0.0185      0.77534    0.62733    0.9583
AGE4          0.5610     0.1785    0.0017      1.75242    1.23497    2.4867
PERIOD2       0.1342     0.0801    0.0936      1.14365    0.97758    1.3379
PERIOD3      -0.0748     0.0993    0.4514      0.92797    0.76389    1.1273
BMI          -0.2376     0.0527    <.0001      0.78849    0.71107    0.8743
BMI1          1.4023     0.3517    <.0001      4.06441    2.03987    8.0983
BMI2         -2.8876     0.8168    0.0004      0.05571    0.01124    0.2762

================================================================================

example 2--bare bones

Association of Predicted Probabilities and Observed Responses

measure               value       measure      value

Percent Concordant     56.2       Somers’ D    0.195
Percent Discordant     36.8       Gamma        0.209
Percent Tied            7.0       Tau-a        0.012
Pairs                 21907512    c            0.597

================================================================================


example 2--bare bones

    CASE and BMI
    PROC LOGISTIC
    Data set:  MERGE0, with 27102 observations
    Outcome variable name:  CASE, with 834 events and 26268 non-events
    Exposure of interest: BMI
    Exposure variable name: BMI
    Range of exposure in data used:  16.07  to 39.99
    Adjusted for:
          age2  age3  age4  period2  period3

    Reference value is  MIN:  16.07000000
    Number of knots: 4
    You chose to use all 2 spline variables: BMI1 BMI2

    Name of graph file:  merge0.BMI.PS

    Model w/o exposure of interest, -2 Log Likelihood: 7421.1723178
                      Linear Model, -2 Log Likelihood: 7374.2227684
                      Spline Model, -2 Log Likelihood: 7332.4752231


    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.
                       If the P value is missing, the
                       automatic selection procedure did
                       not select any spline variables.
                       The relationship between the expo-
                       sure and the outcome, if any, is
                       linear.  SEE LINE 3.              <.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 for the relation of
                       the EXPOSURE to the CASE or TIME. <.0001
    -------------------------------------------------------------
    3    Test for      If LINE 1 indicated a possible
         linear        linear relation between the
         relation      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).                       <.0001

================================================================================
```

First, the macro gives the fractions of the range of BMI that are outside the outer knots. Since the form of the relationship between the *EXPOSURE* and the *CASE* is constrained outside the knots, it is desirable that the fraction of the graph outside the knots not be too large. Furthermore, since the outermost knots are at 'worst' at the 5th and the 95th percentiles of *EXPOSURE*, a large fraction implies that a large area of the graph will represent a very small part of the data. There are two main methods to reduce the fraction of the range outside the knots:
1. Use *KNOT=* to add one or more knots (e.g. give the same set of knots chosen automatically plus one at the appropriate end;
2. Trim the data (using *HPCT, LPCT, HICUT, LOWCUT*).

Since *HEADER1* was not specified, the output begins by listing the *CASE* and the *EXPOSURE*. The graph file has the default name.

The graph is



Adjusted for age2 age3 age4 period2 period3

Note that since we did not give a value for *REFVAL*, it is 'centered' at the minimum value, 16.07.

Since *HLABEL* was not specified, the label for the horizontal axis is just the *EXPOSURE* variable name. Similarly, the *CASE* variable name is shown in the label for the vertical axis.

3.3 Example 3. 5 knot spline with given knots and many optional parameters

Below is a call to %LGTPHCURV9 using a logistic model. In addition, the knots are given using the KNOT parameter to maintain uniformity over a set of analyses. The knot values given are 18.64 21.14 23.52 31.02 36.5 picked by the macro in Example 2, plus one to decrease the fraction of the data outside the knots.

```
%lgtphcurv9(data=merge0, model=logistic, refval=22, exposure=BMI, case=case,
          /* cutoff=2 5,*/ pictname=example3b.ps, outplot=ps, klines=T,
          hlabel=%quote(BMI (kg/sq m)), plotdec=T,
          knot=18.64  21.14  23.52  31.02  36.5,
          vlabel=Odds Ratio for Ovulatory Infertility,
          vlabelstyle=h,
          axlabmult=1.2, axvalmult=1.1,
          ordata=ordata, or=odr, or_lower=lower, or_upper=upper, x_value=mean,
          adj= age2 age3 age4 period2 period3, select=1, plot=2,
          graphtit=Ov Inf vs BMI 4 knot spline adj for age and time pd,
          axordv=0 to 5 by .5);
```

Since I wanted to plot the results from an indicator model (using *ORDATA*), I had to have a data set with the necessary variables. In this case, that data set was named `ORDATA`. It was made by running the logistic model with the indicators (PROC LOGISTIC, with the `/ rl` option in the model statement. I could also have used the ODS dataset ORS. I also had to determine some 'central point' for each of the categories represented by the indicators. I computed both the mean and the median. Then I made the data set `ORDATA`. Each observation has the odds ratio, its lower and upper 95% confidence limits, and the mean and median of the exposure in the group for which this was the indicator. In calling %LGTPHCURV9, I decided to use the mean as the x_value for plotting.

As before, the macro noted how much of the range of the exposure was outside the outer knot points. It then listed the knots in both the .log and the .lst files. Then it printed the values of the spline variables when the *EXPOSURE* was set equal to the *REFVAL*. Finally, it printed the three models (adjusters (covariates) only, adjusters (covariates) plus linear *EXPOSURE*, adjusters (covariates) plus linear and spline *EXPOSURE*

The output in the .lst file looks as follows:

```
================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50
Percent of range of BMI below first knot is 11  .
Percent of range of BMI above last knot is 15  .

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50
    Knots for BMI:
    18.64 21.14 23.52 31.02 36.5

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50



values of spline variables when BMI is 22

  Obs    BMI      BMI1           BMI2    BMI3

27604     22    0.11892    .001994033      0

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50


   Analysis of Maximum Likelihood Estimates & Odds Ratio (with adjusters only)

VARIABLE    ESTIMATE     STDERR   PROBCHISQ   ODDSRATIO   LOWERCL   UPPERCL

Intercept    -4.7669      0.1943    <.0001       0.00851    0.00581   0.01245
BMI           0.0559    0.00795    <.0001       1.05749    1.04113   1.07411
```

```
================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50


Association of Predicted Probabilities and Observed Responses

measure               value       measure      value

Percent Concordant     48.7       Somers' D    0.081
Percent Discordant     40.6       Gamma        0.091
Percent Tied           10.8       Tau-a        0.005
Pairs                 21907512    c            0.541

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50


Analysis of Maximum Likelihood Estimates & Odds Ratios (linear model with adjust

VARIABLE    ESTIMATE     STDERR   PROBCHISQ   ODDSRATIO   LOWERCL   UPPERCL

Intercept    -4.6925     0.1989    <.0001      0.00916    0.00621   0.01353
AGE2         -0.2243     0.0814    0.0059      0.79904    0.68118   0.93729
AGE3         -0.2798     0.1078    0.0094      0.75592    0.61198   0.93372
AGE4          0.5224     0.1779    0.0033      1.68612    1.18987   2.38934
PERIOD2       0.1280     0.0799    0.1094      1.13654    0.97171   1.32933
PERIOD3      -0.0780     0.0991    0.4311      0.92496    0.76170   1.12322
BMI           0.0573    0.00800    <.0001      1.05899    1.04251   1.07573

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50


Association of Predicted Probabilities and Observed Responses

measure               value       measure      value

Percent Concordant     53.2       Somers' D    0.139
Percent Discordant     39.3       Gamma        0.150
Percent Tied            7.6       Tau-a        0.008
Pairs                 21907512    c            0.570

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50
```

```
Analysis of Maximum Likelihood Estimates & Odds Ratio (spline model with adjuste

VARIABLE    ESTIMATE     STDERR   PROBCHISQ   ODDSRATIO   LOWERCL   UPPERCL

Intercept     0.8925     1.0953    0.4152       2.4412    0.28529    20.890
AGE2         -0.2090     0.0816    0.0104       0.8114    0.69147     0.952
AGE3         -0.2537     0.1081    0.0189       0.7759    0.62779     0.959
AGE4          0.5603     0.1786    0.0017       1.7513    1.23415     2.485
PERIOD2       0.1334     0.0801    0.0956       1.1427    0.97679     1.337
PERIOD3      -0.0744     0.0993    0.4536       0.9283    0.76417     1.128
BMI          -0.2191     0.0556    <.0001       0.8032    0.72024     0.896
BMI1          2.4029     0.8730    0.0059      11.0557    1.99744    61.192
BMI2         -4.4886     2.2016    0.0415       0.0112    0.00015     0.841
BMI3          1.8342     1.5581    0.2391       6.2601    0.29530   132.708

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50


Association of Predicted Probabilities and Observed Responses

measure               value       measure      value

Percent Concordant     56.2       Somers' D    0.194
Percent Discordant     36.8       Gamma        0.209
Percent Tied            7.0       Tau-a        0.012
Pairs                 21907512    c            0.597

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:50


Association of Predicted Probabilities and Observed Responses
    Ov Inf vs BMI 4 knot spline adj for age and time pd
    PROC LOGISTIC
    Data set:  MERGE0, with 27102 observations
    Outcome variable name:  CASE, with 834 events and 26268 non-events
    Exposure of interest: BMI (kg/sq m)
    Exposure variable name: BMI
    Range of exposure in data used:  16.07  to 39.99
    Adjusted for:
          age2  age3  age4  period2  period3

    Reference value is  USER VALUE:  22
    Number of knots: 5
    You chose to use all 3 spline variables: BMI1 BMI2 BMI3
```

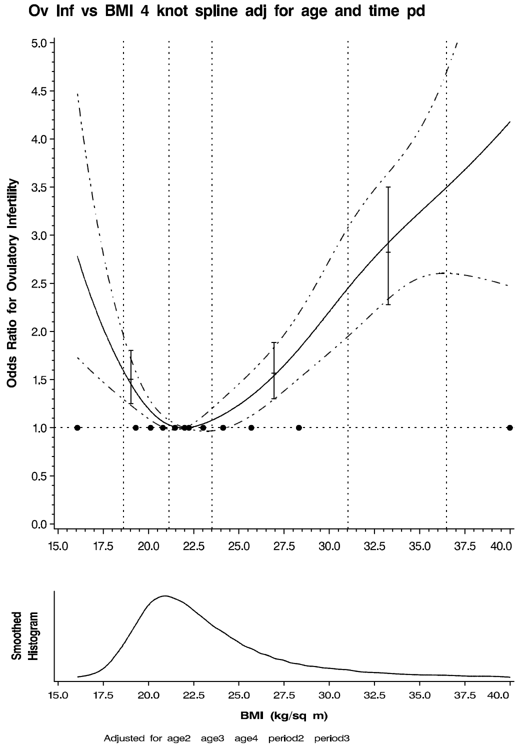
```
Name of graph file:  example3.ps

Model w/o exposure of interest, -2 Log Likelihood: 7448.592084
                  Linear Model, -2 Log Likelihood: 7374.2227684
                  Spline Model, -2 Log Likelihood: 7331.2779011


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.
                   If the P value is missing, the
                   automatic selection procedure did
                   not select any spline variables.
                   The relationship between the expo-
                   sure and the outcome, if any, is
                   linear.  SEE LINE 3.              <.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 for the relation of
                   the EXPOSURE to the CASE or TIME. <.0001
-------------------------------------------------------------
3    Test for      If LINE 1 indicated a possible
     linear        linear relation between the
     relation      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).                       <.0001
```

Three logistic regression models were fit and abbreviated output was printed:

```
1.  The ordinary logistic regression model without
    the exposure of interest;
2.  The ordinary logistic regression model with the exposure of interest
    entered as a single linear term;
3.  The NK-knot spline logistic regression model (with the
exposure of interest expressed as EXPOSURE EXPOSURE1 EXPOSURE2 EXPOSURE3
--the number of spline variables being 2 less than the number of knots).
```

Finally, the macro printed out a summary of the whole procedure. It told us the SAS procedure used, the name of the data set used and the number of observations in the data set, the name of the exposure, the name of the outcome variable, the header for the graph, the reference value, the range of the exposure in the data, the list of adjusting variables (covariates), the number of knots, which spline variables were used, and finally the -2 log likelihoods for the 3 models and the results of the corresponding likelihood ratio tests. In this case, all p values are highly significant.

The test for the overall significance of the curve (Line 2) gives the p-value for the comparison of the spline model with the model having only the adjusters (covariates) (model without the exposure of interest). The test for linear association (Line 3) compares the linear model with the model having only the covariates. The test for non-linearity (Line 1) compares the spline model to the linear model. If the p-value is large (e.g. over .05), you should probably graph the linear relation (i.e. use *PWHICH*=LINEAR), though there is some room for judgment here.
The graph is



Note that the graph has a title (*HEADER1*), that the axes have nice labels (*HLABEL* and *VLABEL*), that the deciles of the data (0, 10, ... 100 percentile levels) are graphed on the OR=1 reference line, that the knot lines are shown, that a smooth histogram is plotted, and that the odds ratios from a model with indicators for BMI levels are also graphed. The footer at the bottom lists the adjusting variables. After running the macro with the default values for the *AXLABMULT* and *AXVALMULT*, I decided to make the type a little bigger using *AXLABMULT*=1.2 and *AXVALMULT*=1.1.

NOTE: In order for the odds ratios and 95% confidence intervals from the indicator model to coincide (roughly) with the confidence intervals for the spline curve, the spline curve must be centered near the "center" (mean or median) of the reference group in the indicator model. The bars from the indicator model are located at the horizontal values given by the x_value of their categories. In our example, ORDATA contains variables named MEAN and MEDIAN. If I had used the variable named MEDIAN, the bars would have been in slightly different locations.

NOTE: Since we used the *CUTOFF* option, both the spline curve and its confidence interval have been truncated at 5.

IMPORTANT: If you are showing the confidence band as a cloud (*CI*=1) and the curve for the upper confidence limit goes above the upper end of the vertical axis, you must use *CUTOFF*=2 *upper limit.* If you do not, the cloud will end abruptly when the upper confidence limit reaches the upper limit. (See Example 4 and Frequently Asked Questions).

3.4 Example 4. 3 Knot spline

The macro call is

```
%lgtphcurv9(data=merge0, refval=22, exposure=BMI, case=case,
          pictname=example4.ps, ci=1,
          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,
          adjdat=adjref, modprint=f);
```

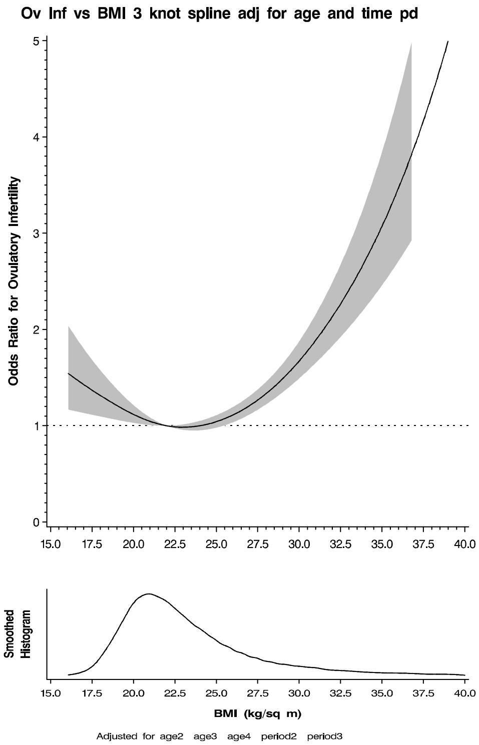
The output is

```
================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:52
example 4--3 knot spline
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/examples.lgtphcurv Program example1-6 06JUL2010 13:52
example 4--3 knot spline
    Knots for BMI:
    18.64 22.24 31.02

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:52
example 4--3 knot spline


values of spline variables when BMI is 22

  Obs    BMI      BMI1
```

```
27604     22    0.24750

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:52
example 4--3 knot spline


    Ov Inf vs BMI 3 knot spline adj for age and time pd
    PROC LOGISTIC
    Data set:  MERGE0, with 27102 observations
    Outcome variable name:  CASE, with 834 events and 26268 non-events
    Exposure of interest: BMI (kg/sq m)
    Exposure variable name: BMI
    Range of exposure in data used:  16.07  to 39.99
    Adjusted for:
          age2  age3  age4  period2  period3

    Reference value is  USER VALUE:  22
    Number of knots: 3
    You chose to use all 1 spline variables: BMI1

    Name of graph file:  example4.ps

    Model w/o exposure of interest, -2 Log Likelihood: 7448.592084
                      Linear Model, -2 Log Likelihood: 7374.2227684
                      Spline Model, -2 Log Likelihood: 7342.7141655


    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.
                       If the P value is missing, the
                       automatic selection procedure did
                       not select any spline variables.
                       The relationship between the expo-
                       sure and the outcome, if any, is
                       linear.  SEE LINE 3.              <.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 for the relation of
                   the EXPOSURE to the CASE or TIME. <.0001
-------------------------------------------------------------
3    Test for      If LINE 1 indicated a possible
     linear        linear relation between the
     relation      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).                        <.0001
```

The graph is



The graph with 3 knots shows less excess risk at lower values of BMI.
Note that the confidence cloud ends abruptly before the end of the curve. This occurred because the upper confidence limit hit the upper limit of the vertical axes, and I did not use *CUTOFF* in the macro call..

3.5 Example 5. select=3 (automatic stepwise selection)

This is similar to Examples 3 and 4, but this time we use automatic stepwise selection (*SELECT*=3) to choose from 17 knots. Now that we are experienced in reading the macro output, we use *TESTREP*=SHORT.

The macro call is

```
title2 'example 5--automatic selection from 17 knot spline';
%lgtphcurv9(data=merge0, model=logistic, refval=22, exposure=BMI, case=case,
          pictname=example5.ps, klines=F, displayx=F, nk=17, select=3,
          hlabel=%quote(BMI (kg/sq m)),
          vlabel=Odds Ratio for Ovulatory Infertility,
          ordata=ordata, or=odr, or_lower=lower, or_upper=upper, x_value=mean,
          adj= age2 age3 age4 period2 period3, plot=2,
          graphtit=Ov Inf vs BMI  adj for age and time pd,
          footer=Adjusted for age and time period,
          modprint=f, testrep=short, adjdat=adjref);
```

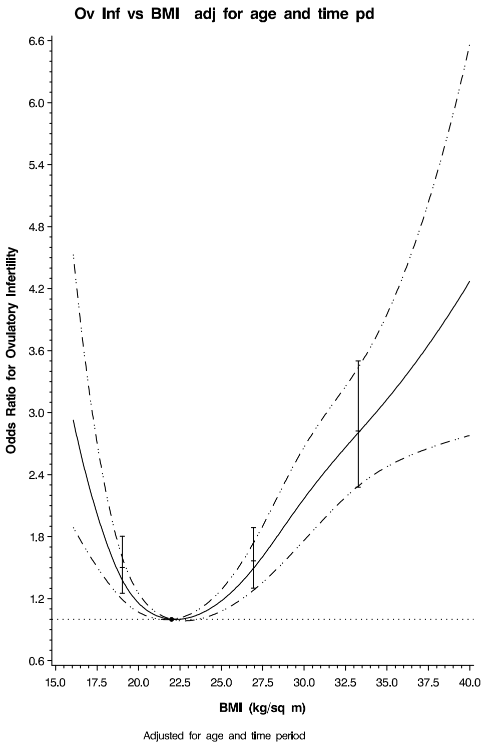
The partial output is

```
================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
example 5--automatic selection from 17 knot spline
Percent of range of BMI below the first knot is 8  .
Percent of range of BMI above the last knot  is 23  .

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
example 5--automatic selection from 17 knot spline
    Knots for BMI:
    18.02 19.11 19.69 20.12 20.53 20.98 21.31 21.79
    22.24 22.67 23.24 23.91 24.66 25.69 26.95 29.23
    34.39

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
example 5--automatic selection from 17 knot spline


values of spline variables when BMI is 22

  Obs BMI   BMI1    BMI2     BMI3     BMI4     BMI5         BMI6       BMI7

27604  22 0.23526 0.090073 0.045998 0.024796 0.011854 .003960073 .001225886

  Obs       BMI8   BMI9   BMI10   BMI11   BMI12   BMI13   BMI14   BMI15

27604 .000034559     0      0       0       0       0       0       0

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
```

```
example 5--automatic selection from 17 knot spline


Step  0 :  variable BMI1  added

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
example 5--automatic selection from 17 knot spline


Step 1 :  no variable dropped

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
example 5--automatic selection from 17 knot spline


Step  2 :  variable BMI2  added

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
example 5--automatic selection from 17 knot spline


Step  3 :   no variable added

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
example 5--automatic selection from 17 knot spline


stepwise procedure cannot add or delete more variables.

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:53
example 5--automatic selection from 17 knot spline


    Ov Inf vs BMI  adj for age and time pd
    PROC LOGISTIC
    Data set:  MERGE0, with 27102 observations
    Outcome variable name:  CASE, with 834 events and 26268 non-events
    Exposure of interest: BMI (kg/sq m)
    Exposure variable name: BMI
    Range of exposure in data used:  16.07  to 39.99
```

```
Adjusted for:
      age2  age3  age4  period2  period3

Reference value is  USER VALUE:  22
Number of knots: 17
You chose to select spline variables automatically, with sls=.05 and sle=.05
The following spline variables were selected:
     BMI1 BMI2

Name of graph file:  example5.ps

Model w/o exposure of interest, -2 Log Likelihood: 7448.592084
                  Linear Model, -2 Log Likelihood: 7374.2227684
                  Spline Model, -2 Log Likelihood: 7332.0838498


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                      <.0001
```

The graph is



So using the stepwise selection based on the default selection criteria, there are 2 spline variables selected and they are the first and second.

NOTE: The meanings/values of "BMI1" and "BMI2" differ, depending on the number of knots

used and their placement. When *NK*=3 (Example 4), the value of `BMI1` when `BMI`=22 was .24750. But with the 17 knots, it is .23526, while with 5 knots (Example 3) it was .11892.
MORAL: The spline variables are just a way to make pretty curves. Neither they nor their coefficients in the model should be interpreted directly.

The SHORT report on the likelihood ratio tests gives only the p-values.

3.6 Example 6. select=2 with usersplv

Based on example 5, there are 2 spline variables selected into the model, so if you would like to modify the graph (e.g. fix up the axes or their labels), you don't need to re-run the stepwise selection, which may take a lot of time. Instead, just give the spline variables you want to use. You still need to set *NK* to 17 to calculate correctly the spline variable values (The reason you still need to set *NK*=17 is that the computations of the spline functions may rely on more knots than those related to the selected spline variables, so if some of those knots are omitted, the spline functions with the same names as those selected will not be the same as those selected originally. See the NOTE above, as well as Section 4, Computational Methods).

The macro call is

```
title2 'example 6--17 knot spline, but use the spline variables selected in example 5';
%lgtphcurv9(data=merge0, model=logistic, refval=22, exposure=BMI, case=case,
          pictname=example6.ps, klines=F, displayx=F, nk=17,
          select=2, usersplv=BMI1 BMI2,
          hlabel=%quote(BMI (kg/sq m)),
          vlabel=Odds Ratio for Ovulatory Infertility,
          ordata=ordata, or=odr, or_lower=lower, or_upper=upper, x_value=mean,
          adj= age2 age3 age4 period2 period3, plot=2,
          graphtit=Ov Inf vs BMI 3 selected from 17 knot spline,
          footer=NONE,
          modprint=f, testrep=short, adjdat=adjref)
```

The output is almost the same as Example 5 except it will point out that you chose to use these spline variables instead of saying that you chose to select spline variables automatically.

```
================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:55
example 6--17 knot spline, but use the spline variables selected in example 5
Percent of range of BMI below the first knot is 8  .
Percent of range of BMI above the last knot  is 23  .

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:55
example 6--17 knot spline, but use the spline variables selected in example 5
    Knots for BMI:
    18.02 19.11 19.69 20.12 20.53 20.98 21.31 21.79
    22.24 22.67 23.24 23.91 24.66 25.69 26.95 29.23
    34.39
```

```
================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:55
example 6--17 knot spline, but use the spline variables selected in example 5


values of spline variables when BMI is 22

  Obs BMI   BMI1    BMI2     BMI3     BMI4     BMI5         BMI6       BMI7

27604  22 0.23526 0.090073 0.045998 0.024796 0.011854 .003960073 .001225886

  Obs       BMI8   BMI9   BMI10   BMI11   BMI12   BMI13   BMI14   BMI15

27604 .000034559     0      0       0       0       0       0       0

================================================================================

/udd/stleh/doctn/examples.lgtphcurv Program example1-6 06JUL2010 13:55
example 6--17 knot spline, but use the spline variables selected in example 5


    Ov Inf vs BMI 3 selected from 17 knot spline
    PROC LOGISTIC
    Data set:  MERGE0, with 27102 observations
    Outcome variable name:  CASE, with 834 events and 26268 non-events
    Exposure of interest: BMI (kg/sq m)
    Exposure variable name: BMI
    Range of exposure in data used:  16.07  to 39.99
    Adjusted for:
          age2  age3  age4  period2  period3

    Reference value is  USER VALUE:  22
    Number of knots: 17
    You  chose to use these spline variables: BMI1 BMI2

    Name of graph file:  example6.ps

    Model w/o exposure of interest, -2 Log Likelihood: 7448.592084
                      Linear Model, -2 Log Likelihood: 7374.2227684
                      Spline Model, -2 Log Likelihood: 7332.0838498


    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                      <.0001
```

The graph is exactly the same as Example 5.

3.7 Example 7. plotinc=T; use PRINTPOINTS

This is a pooled logistic example from NHS. The exposure is AGE and the binary outcome is lung cancer. The covariates are cigarettes/day, duration of smoking, age at the start of smoking and the follow-up period.

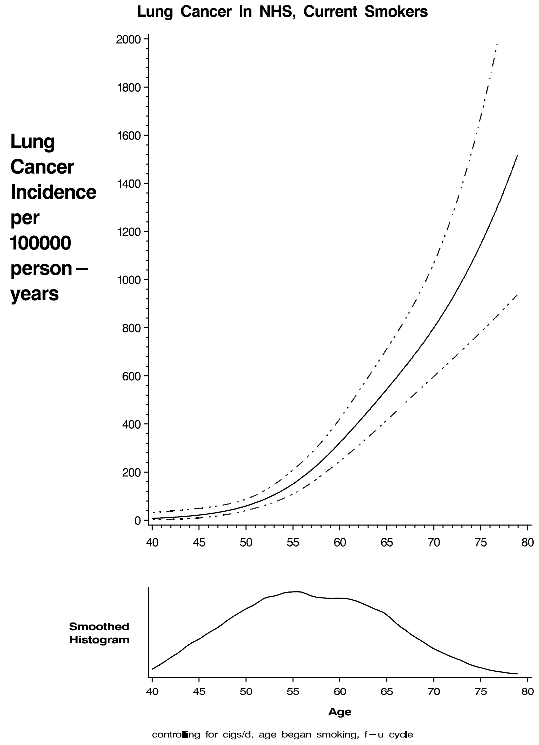
```
title2 'Example 7--plotting incidence rate';
%lgtphcurv9(data=nhscurr, pictname=example7.ps,  plotinc=t,
exposure=age,  case=lung19,
header1=%quote(Lung Cancer in NHS, Current Smokers),
footer=%quote(controlling for cigs/d, age began smoking, f-u cycle),
nk=4, klines=f,
testrep=short,
adj=&cig_ &agestr_ &period_, adjdat=adjcurr,
axordh=40 to 80 by 5,  refval=40,
axordi=0 to 2000 by 200,
hlabel=Age,
horigin=2,
vlabelstyle=H,
printpoints=40 50 60 70,
vlabeli=Lung Cancer Incidence per 100000 person- years);
```

The final results are

```
================================================================================

/udd/stleh/doctn/lgtphcurv  Program example7   17JAN2011   20:22    stleh
Example 7--plotting incidence rate

    Lung Cancer in NHS, Current Smokers
    PROC LOGISTIC
    Data set:  NHSCURR, with 101738 observations
    Outcome variable name:  LUNG19, with 509 events and 101229 non-events
    Exposure of interest: Age
    Exposure variable name: AGE
    Range of exposure in data used:  40  to 78.916666508
    Adjusted for:
          cig1  cig3  cig4  agestr1  agestr3
          period1  period2  period3  period5  period6
          period7

    Reference value is  USER VALUE:  40
    Number of knots: 4
    You chose to use all 2 spline variables: AGE1 AGE2

    Name of graph file:  example7.ps
    Model w/o exposure of interest, -2 Log Likelihood: 6245.3714931
                      Linear Model, -2 Log Likelihood: 5951.4649953
                      Spline Model, -2 Log Likelihood: 5936.4018654


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

    1    Test for curvature (i.e. non-linear relation) 0.0005
    2    Test for overall significance of curve       <.0001
    3    Test for linear relation                     <.0001

================================================================================

/udd/stleh/doctn/lgtphcurv  Program example7   17JAN2011   20:22    stleh
values for points requested by user as PRINTPOINTS

AGE     PROB     LOWERPRB   UPPERPRB   AGE     INC     LOWERINC   UPPERINC

 40   0.000154   0.000036   0.000650    40    7.679      1.812      32.52
 50   0.001189   0.000806   0.001755    50   59.474     40.307      87.74
 60   0.006449   0.004924   0.008442    60  322.437    246.185     422.11
 70   0.015978   0.011943   0.021347    70  798.904    597.156    1067.34
```

Note that after the report on the tests for nonlinearity and overall significance, the macro printed the incidence (since we set PLOTINC=T) and confidence intervals for the ages requested in PRINTPOINTS.

The graph is



3.8 Example 8. A variable without curvature

In the following 3 we revert to the ovulatory infertility study, but the exposure is hours of vigorous exercise per week, from a subset of the data set used in Examples 1-6. The macro reads the data directly frrom a permanent dataset (in `libname ellen`). The study is restricted to nulliparous nonsmokers using the *WHERE* and *EXTRAV* parameters.

The macro call is

```
title2 'example 8';
%lgtphcurv9( data=ellen.merge,  exposure=VH, case=case, pictname=example8.ps,
           hlabel=Hours of Vigorous Exercise,
           vlabel=Odds Ratio for Ovulatory Infertility,
           klines=F,
           adj= age2 age3 age4 period2 period3,
           graphtit=Ov Inf vs Vig Ex 4 knot spline adj for age and time pd,
           where=par0 eq 1 and nvsmk eq 1, extrav=par0 nvsmk,
           testrep=short, modprint=f,
           axvalmult=1.1, axlabmult=1.5,
           footer=nulliparous non-smokers, footmult=1.5)
```

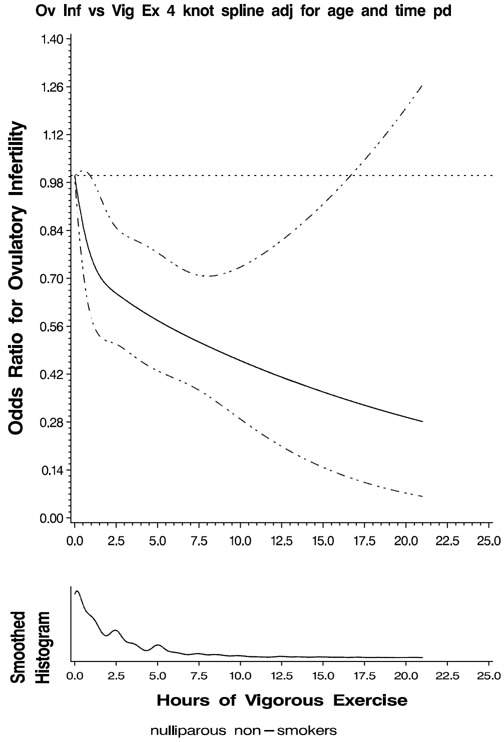
The partial output is

```
================================================================================

/udd/stleh/doctn/lgtphcurv  Program example8-10   17JAN2011   20:29    stleh
example 8
```

```
Percent of range of VH below the first knot is 0  .
Percent of range of VH above the last knot  is 60  .

================================================================================

/udd/stleh/doctn/lgtphcurv  Program example8-10   17JAN2011   20:29    stleh
example 8
    Knots for VH:
    0 0.67 2.5 8.37

================================================================================

/udd/stleh/doctn/lgtphcurv  Program example8-10   17JAN2011   20:29    stleh
example 8


values of spline variables when VH is 0.00000000

 Obs    VH    VH1    VH2

5640     0     0      0

================================================================================

/udd/stleh/doctn/lgtphcurv  Program example8-10   17JAN2011   20:29    stleh
example 8

    Ov Inf vs Vig Ex 4 knot spline adj for age and time pd
    PROC LOGISTIC
    Data set:  ELLEN.MERGE, with 5639 observations
    Outcome variable name:  CASE, with 385 events and 5254 non-events
    Exposure of interest: Hours of Vigorous Exercise
    Exposure variable name: VH
    Range of exposure in data used:  0  to 21
    Adjusted for:
          age2  age3  age4  period2  period3

    Reference value is  MIN:  0.00000000
    Number of knots: 4
    You chose to use all 2 spline variables: VH1 VH2

    Name of graph file:  example8.ps

    Model w/o exposure of interest, -2 Log Likelihood: 2801.1766207
                      Linear Model, -2 Log Likelihood: 2785.5279575
                      Spline Model, -2 Log Likelihood: 2782.3982229


    Line Test Name                                       P value
    -------------------------------------------------------------
    1    Test for curvature (i.e. non-linear relation) 0.2091
    2    Test for overall significance of curve        0.0003
    3    Test for linear relation                      0.0001

================================================================================
```

Since the test for non-linearity has p-value .21, we will use pwhich=linear in the next graph. This time, the macro used all the spline variables, producing the curvature seen below. We will also switch to *NK*=3, because once you have decided to plot the linear graph, the number of knots is irrelevant (so you might as well use a small number).

The graph is



Note: at the beginning of the macro output, the output notes that 60% of the range of `VH` was above the last knot. You can see from the graph that the bulk of the distribution is below 10. We will take care of this problem in Example 10.

3.9 Example 9. pwhich=linear, logarithmic vertical axis, horizontal label

The macro call is

```
title2 'example 9';
%lgtphcurv9( data=ellen.merge, model=logistic, pictname=example9.ps,
           hlabel=Hours of Vigorous Exercise, nk=3,
           vlabel=Odds Ratio for Ovula- tory Infer- tility,
           exposure=VH, case=case,
           adj= age2 age3 age4 period2 period3,
           printcv=T,  pwhich=LINEAR,
           displayx=T,
           graphtit=Ov Inf vs Vig Ex  linear relation  adj for age and time pd,
           adjdat=adjref,
           axordvlog10=T,  vlabelstyle=h,
   klines=F,
           where=par0 eq 1 and nvsmk eq 1, extrav=par0 nvsmk,
           testrep=short, modprint=f,
           footer=nulliparous non-smokers)
```

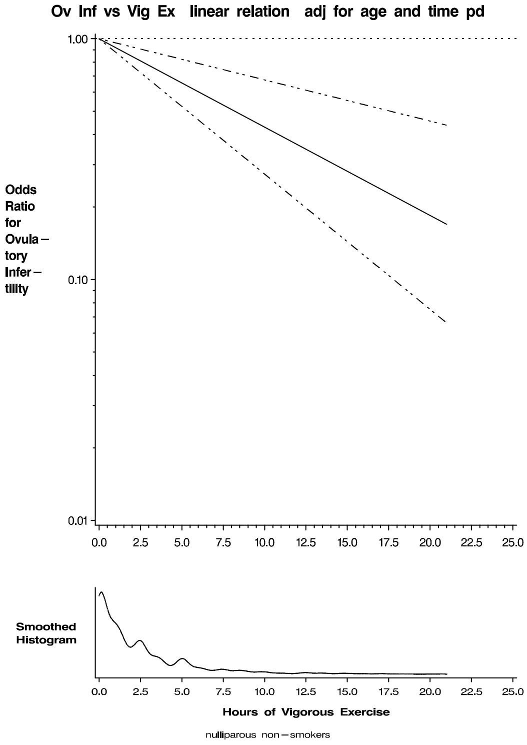
Note that the label for the vertical axis (*AXLABV*) has hyphenated words with spaces after the hyphens. The macro will put each "word" (i.e. string separated by spaces) on a separate line. It is not desirable for the axis label to take up too much of the space alloted to the graph.

The partial output is

```
================================================================================


example 9
Percent of range of VH below the first knot is 0  .
Percent of range of VH above the last knot  is 60  .

================================================================================


example 9
    Knots for VH:
    0 1.25 8.37

================================================================================

example 9


values of spline variables when VH is 0.00000000

 Obs    VH    VH1

5640     0     0

================================================================================


example 9

    Ov Inf vs Vig Ex  linear relation  adj for age and time pd
    PROC LOGISTIC
    Data set:  ELLEN.MERGE, with 5639 observations
    Outcome variable name:  CASE, with 385 events and 5254 non-events
    Exposure of interest: Hours of Vigorous Exercise
    Exposure variable name: VH
    Range of exposure in data used:  0  to 21
    Adjusted for:
          age2  age3  age4  period2  period3

    Reference value is  MIN:  0.00000000
    Number of knots: 3
    You chose to use all 1 spline variables: VH1

    Name of graph file:  example9.ps

    Model w/o exposure of interest, -2 Log Likelihood: 2801.1766207
                      Linear Model, -2 Log Likelihood: 2785.5279575
                      Spline Model, -2 Log Likelihood: 2783.0527101


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

    1    Test for curvature (i.e. non-linear relation) 0.1157
    2    Test for overall significance of curve        0.0001
    3    Test for linear relation                      0.0001

================================================================================

example 9


The variance-covariance matrix among all spline variables is:

       VH

VH 0.0005

================================================================================
```

Note that the p value for curvature is different from that in Example 8, because we changed the number of knots.

The graph is



3.10 Example 10. Restricting the range of the exposure

There was some concern that the downward direction of the graph was being overly influenced by the (relatively rare) very high values for the exposure. We therefore decided to restrict the data to `VH le 10`. In general, it is undesirable for a large part of the graph to be generated by very small fraction of the data. In addition, we are supplying a large number of knot values and using automatic selection.

The macro call is

```
title3 'example 10';
%lgtphcurv9( data=ellen.merge, model=LOGISTIC, pictname=example10.ps,
          hlabel=Hours of Vigorous Exercise, select=3, hicut=10,
          knot=0 .3 .5 .7 .9 1.1 1.3 1.5 1.7 1.9 2.1 2.3 2.5 3 3.5
          4 4.5 5 5.5 6 7 8 9,
          vlabel=Odds Ratio for Ovulatory Infertility,
          exposure=VH, case=case,
          adj=age2 age3 age4 period2 period3,
          n_grid=500,  displayx=T,
          e=T, graphtit=Ov Inf vs Vig Ex--automatic selection ,
          plot=2, outplot=PS,
          klines=F, axordv=0 to 1.1 by .1, modprint=F,
          where=par0 eq 1 and nvsmk eq 1, extrav=par0 nvsmk,
          testrep=short,
          footer=nulliparous non-smokers, footmult=1.5);
```

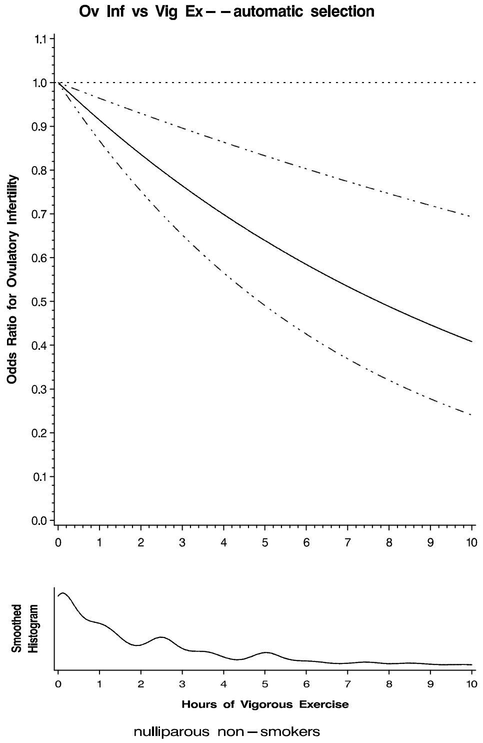
The partial output is

```
================================================================================



example 10
Percent of range of VH below first knot is 0  .
Percent of range of VH above last knot is 10  .

================================================================================



example 10
    Knots for VH:
    0 .3 .5 .7 .9 1.1 1.3 1.5
    1.7 1.9 2.1 2.3 2.5 3 3.5 4
    4.5 5 5.5 6 7 8 9

================================================================================

example 10

values of spline variables when VH is 0.00000000

                                    V  V  V  V  V  V  V  V  V  V  V  V
  O      V  V  V  V  V  V  V  V  V  H  H  H  H  H  H  H  H  H  H  H  H
  b   V  H  H  H  H  H  H  H  H  H  1  1  1  1  1  1  1  1  1  1  2  2
  s   H  1  2  3  4  5  6  7  8  9  0  1  2  3  4  5  6  7  8  9  0  1

5499  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0

================================================================================



example 10

Step  0 :   no variable added

================================================================================



example 10

stepwise procedure cannot add or delete more variables.

================================================================================
example 10

NOTE: No spline variables are selected by the current criteria.
      You can either change the paramter values for sls, sle or nk, or
      bear in mind the only valid test is the linear test.
      The graph output will be the linear graph.

================================================================================
```

```
example 10
    Ov Inf vs Vig Ex--automatic selection
    PROC LOGISTIC
    Data set:  ELLEN.MERGE, with 5498 observations
    Outcome variable name:  CASE, with 380 events and 5118 non-events
    Exposure of interest: Hours of Vigorous Exercise
    Exposure variable name: VH
    Range of exposure in data used:  0  to 10
    Adjusted for:
          age2  age3  age4  period2  period3

    Reference value is  MIN:  0.00000000
    Number of knots: 23
    You chose to select spline variables automatically, with sls=.05 and sle=.05
    No spline variable is selected by the current criteria

    Name of graph file:  example10.ps

    Model w/o exposure of interest, -2 Log Likelihood: 2756.3413559
                      Linear Model, -2 Log Likelihood: 2744.3169145
    There is no spline model available (no spline var.)


    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                      0.0005
```

Note that the number of observations in the output from using *HICUT*=10 is lower than the number in the original set of nulliparous nonsmokers. You can also use *HPCT* or *LPCT* to delete observations with *EXPOSURE* beyond speficied upper or lower percentiles.

The graph is



Ov Inf vs Vig Ex--automatic selection

nulliparous non-smokers

3.11 Example 11. Conditional logistic regression (MODEL=condlog)

This option normally is used if you have a matched case-control study, that is, if you would like to use a conditional logistic model for your data. The code below calls %LGTPHCURV9 to do a conditional logistic regression in a matched case-control study of LDL cholesterol and CHD, controlling for age, hypertension, BMI, HDL cholesterol, and aspirin use.

The macro call is

```
title2 'example 11a--conditional logistic regression';
%lgtphcurv9(data=lipids, model=condlog, strata=matchid,  time=censor,
case=cc, exposure=ldl_c,
header1=LDL cholesterol and CHD, graphtit=NONE,
adj=hdlg2 hdlg3 hdlg4 hdlg5 agegp4 agegp6 hbp90f asp90 bmi2530 bmi30hi ,
pictname=condlogex.ps,
pwhich=spline, axordh=50 to 300 by 50, axordv=0 to 6 by 1, klines=T,
refval=100, vlabel=Relative Risk of CHD, hlabel=LDL Cholesterol,
footer=%quote(Adjusted for HDL chol, age, hypertension, aspirin use, and BMI),
testrep=short, modprint=f,
plot=2);
```

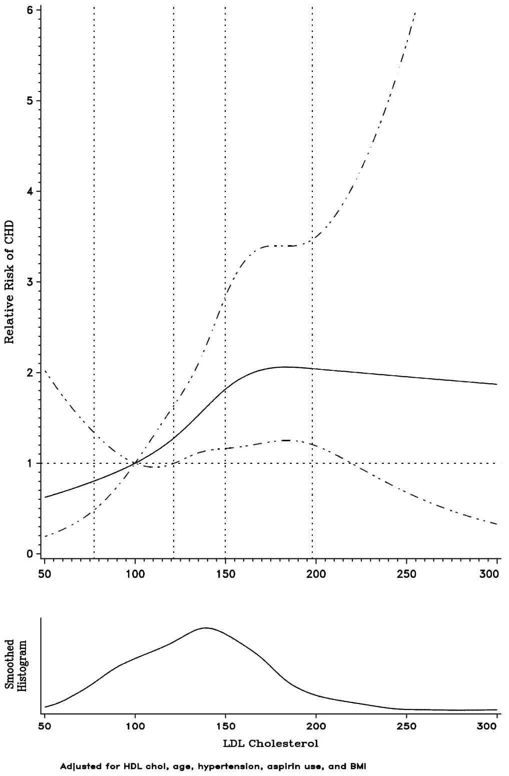
Note that in this macro call, we specified a *FOOTER* listing the conceptual variables.

The output is

```
==========================================================================
/udd/stleh/doctn/lgtphcurv  Program example11ab   17JAN2011   20:37    stleh
example 11a--conditional logistic regression
Percent of range of LDL_C below the first knot is 13  .
Percent of range of LDL_C above the last knot  is 42  .

================================================================================

/udd/stleh/doctn/lgtphcurv  Program example11ab   17JAN2011   20:37    stleh
example 11a--conditional logistic regression
    Knots for LDL_C:
    77.3 121.2 150 197.8

================================================================================

/udd/stleh/doctn/lgtphcurv  Program example11ab   17JAN2011   20:37    stleh
example 11a--conditional logistic regression


values of spline variables when LDL_C is 100

 Obs    LDL_C     LDL_C1    LDL_C2

1172     100     0.80557       0

================================================================================

/udd/stleh/doctn/lgtphcurv  Program example11ab   17JAN2011   20:37    stleh
example 11a--conditional logistic regression

    LDL cholesterol and CHD
    PROC PHREG
    Conditioned on matchid
    Data set:  LIPIDS, with 670 observations
    Time variable name:  CENSOR
    Censoring variable name:  CC with 222 events and 448 censored
    Exposure of interest: LDL Cholesterol
    Exposure variable name: LDL_C
    Range of exposure in data used:  43.9  to 307.7
    Adjusted for:
          hdlg2  hdlg3  hdlg4  hdlg5  agegp4
          agegp6  hbp90f  asp90  bmi2530  bmi30hi

    Reference value is  USER VALUE:  100
    Number of knots: 4
    You chose to use all 2 spline variables: LDL_C1 LDL_C2

    Name of graph file:  condlogex.ps

    Model w/o exposure of interest, -2 Log Likelihood: 398.9986202
                      Linear Model, -2 Log Likelihood: 390.34610796
```

```
                          Spline Model, -2 Log Likelihood: 388.74846034


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


    1    Test for curvature (i.e. non-linear relation) 0.4499
    2    Test for overall significance of curve        0.0166
    3    Test for linear relation                      0.0033

================================================================================
```

The graph is



Note that, although there seems to be a ceiling effect, the test for curvature is non-significant.

Because of the large fraction of the range of the data above the last knot and given the lack of evidence for curvature, we decided to try a second model with data trimmed at the 1st and 98th percentiles, with a *PWHICH*=linear.

The macro call is:

```
options ps=40;
title2 'example 11b--conditional logistic regression, trimmed';
%lgtphcurv9(data=lipids, model=condlog, strata=matchid,  time=censor,
case=cc, exposure=ldl_c,
adj=hdlg2 hdlg3 hdlg4 hdlg5 agegp4 agegp6 hbp90f asp90 bmi2530 bmi30hi ,
pictname=example11b.ps,
header1=LDL cholesterol and CHD,
hpct=98, lpct=1,
pwhich=linear, axordh=50 to 250 by 50, axordv=0 to 6 by 1, klines=T,
testrep=short, modprint=f,
footer=%quote(Adjusted for HDL chol, age, hypertension, aspirin use, and BMI),
refval=100, vlabel=Relative Risk of CHD, hlabel=LDL Cholesterol,
plot=3);
options ps=78;
```

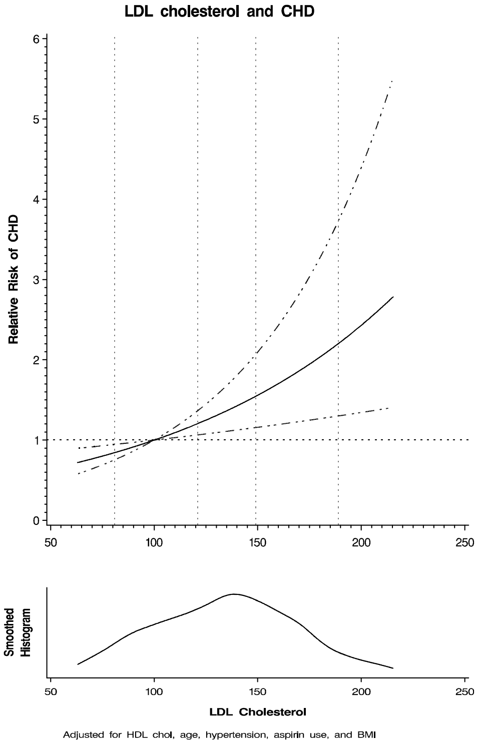
Trimming the data made very little difference to the location of the knots, but the fraction of the data outside the outer knots is greatly reduced.

The output is:

```
================================================================================


example 11b--conditional logistic regression, trimmed
Percent of range of LDL_C below the first knot is 12  .
Percent of range of LDL_C above the last knot  is 17  .

================================================================================


example 11b--conditional logistic regression, trimmed
    Knots for LDL_C:
    81 121.1 149.1 189

================================================================================

example 11b--conditional logistic regression, trimmed


values of spline variables when LDL_C is 100

 Obs    LDL_C     LDL_C1    LDL_C2

1154     100     0.58805       0

================================================================================


example 11b--conditional logistic regression, trimmed

    LDL cholesterol and CHD
    PROC PHREG
    Conditioned on matchid
    Data set:  LIPIDS, with 652 observations
    Time variable name:  CENSOR
    Censoring variable name:  CC with 215 events and 437 censored
    Exposure of interest: LDL Cholesterol
    Exposure variable name: LDL_C
    Range of exposure in data used:  63  to 215.3
    Adjusted for:
          hdlg2  hdlg3  hdlg4  hdlg5  agegp4
          agegp6  hbp90f  asp90  bmi2530  bmi30hi

    Reference value is  USER VALUE:  100
    Number of knots: 4
    You chose to use all 2 spline variables: LDL_C1 LDL_C2

    Name of graph file:  example11b.ps

    Model w/o exposure of interest, -2 Log Likelihood: 378.48496646
                      Linear Model, -2 Log Likelihood: 369.55368676
                      Spline Model, -2 Log Likelihood: 368.56755254


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

    1    Test for curvature (i.e. non-linear relation) 0.6108
    2    Test for overall significance of curve        0.0193
    3    Test for linear relation                      0.0028

================================================================================
```

Line 1 shows that we were justified in using *PWHICH*=linear, even on the trimmed data set.

The graph is:



3.12 Example 12. MODEL=COX

This example uses data from a study of fibroids (`uflap`) in NHS II. The main exposure is BMI (`bmicp`). The time variable is `sdytime`. The covariates are time started antihypertensive medications (`tv2da`), race/ethnicity (`raceth`), age at menarche (`meng`), age at first OC use (`foc`), age at first birth (`afbi`), marital status (`marryg`), time since last birth (`tslbf`), infertility (`fail`), irregular menses (`irreg`).

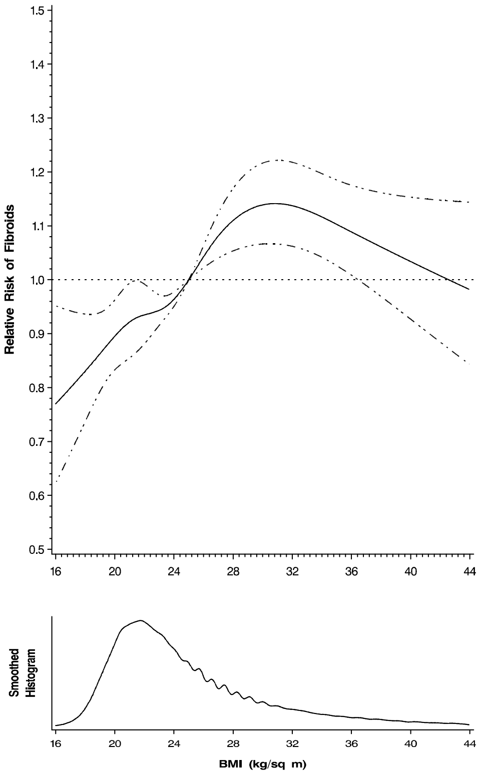
The macro call using 5 knots automatically placed is

```
%lgtphcurv9(data=fibroid1, time=sdytime, strata=agemo period, model=cox,
exposure=bmicp, where=bmimp eq 0, extrav=bmimp, refval=25,
case=uflap,
hicut=44, lowcut=16, nk=5,
adj=tv2da2 tv2da3 tv2da4 tv2da5 tv2da6 tv2da7 tv2da8 tv2da9
afbi1 afbim
fail1 failm
foc0 foc1 foc3 foc4 focm
irreg1 irreg2 irregm
marryg1
meng1 meng2 meng4 meng5 meng6 meng7 mengm
raceth2 raceth3 raceth4 racethm
tslbf0 tslbf2 tslbf3 tslbfm,
plot=2, testrep=short, modprint=f,
hlabel=%quote(BMI (kg/sq m)), axordh=16 to 44 by 4,
vlabel=Relative Risk of Fibroids, axordv=.5 to 1.5 by .1,
pictname=example12a.ps,
header1=BMI and Fibroids, graphtit=NONE,  footer=NONE);
```

Since the model outputs are long and not very informative, we give partial output here.

```
================================================================================

/udd/stleh/doctn/examples.lgtphcurv  Program example12   06JUL2010   14:34    st
Percent of range of BMICP below the first knot is 11  .
Percent of range of BMICP above the last knot  is 31  .

================================================================================

/udd/stleh/doctn/examples.lgtphcurv  Program example12   06JUL2010   14:34    st
    Knots for BMICP:
    19.05 21.31 23.4 26.54 35.36

================================================================================

/udd/stleh/doctn/examples.lgtphcurv  Program example12   06JUL2010   14:34    st



values of spline variables when BMICP is 25

   Obs    BMICP     BMICP1     BMICP2     BMICP3

414109      25     0.79185    0.18887    0.015398

================================================================================

/udd/stleh/doctn/examples.lgtphcurv  Program example12   06JUL2010   14:34    st



    BMI and Fibroids
    PROC PHREG
    Conditioned on agemo period
    Data set:  FIBROID1, with 413607 observations
    Time variable name:  SDYTIME
    Censoring variable name:  UFLAP with 7103 events and 406504 censored
    Exposure of interest: BMI (kg/sq m)
    Exposure variable name: BMICP
    Range of exposure in data used:  16  to 43.94
    Adjusted for:
          tv2da2  tv2da3  tv2da4  tv2da5  tv2da6
          tv2da7  tv2da8  tv2da9  afbi1  afbim
          fail1  failm  foc0  foc1  foc3
          foc4  focm  irreg1  irreg2  irregm
          marryg1  meng1  meng2  meng4  meng5
          meng6  meng7  mengm  raceth2  raceth3
          raceth4  racethm  tslbf0  tslbf2  tslbf3
          tslbfm

    Reference value is  USER VALUE:  25
    Number of knots: 5
    You chose to use all 3 spline variables: BMICP1 BMICP2 BMICP3

    Name of graph file:  example12a.ps

    Model w/o exposure of interest, -2 Log Likelihood: 84047.804272
                      Linear Model, -2 Log Likelihood: 82877.256564
                      Spline Model, -2 Log Likelihood: 82857.157407


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

    1    Test for curvature (i.e. non-linear relation) 0.0002
    2    Test for overall significance of curve       <.0001
    3    Test for linear relation                     <.0001
```

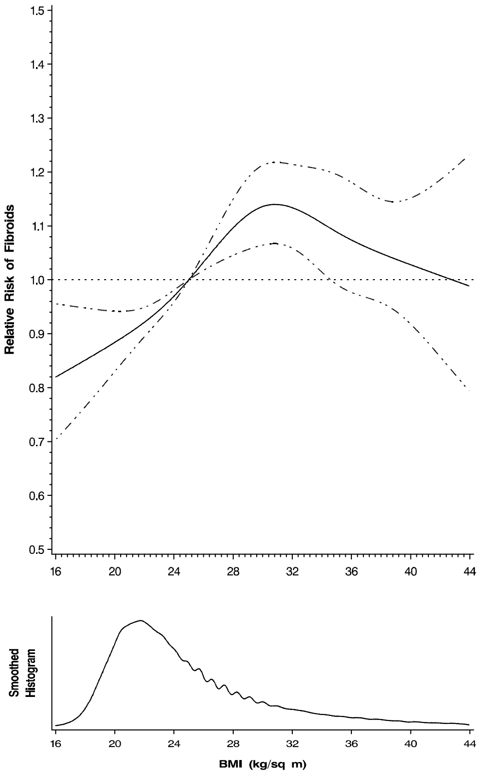
The graph is



Since the knots leave a large space in the middle, and the last knot point is far from the maximum of the exposure range, we tried giving the macro 5 knots, placed at 19, 24, 30, 35, and 40.

The p-value for non-linearity was .0003.

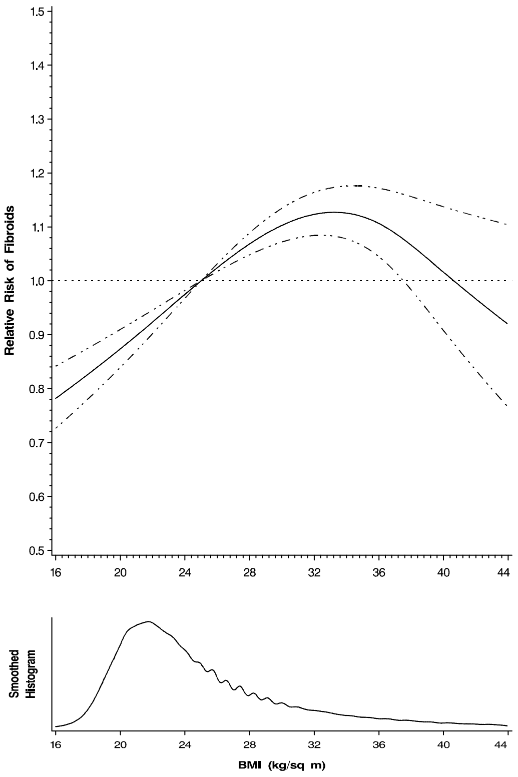
The graph is



Finally, we gave the macro 14 knots (18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 30, 33, 36, 40) and used automatic selection.

3 spline variables (1, 2, and 6), and the p value for non-linearity was .0003.

The graph is



Note that all the graphs look much the same, regardless of the placement of the knots used.

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 Warnings

It is not currently possible to plot the predicted probability and the OR or IRR on the same graph.

6 Frequently Asked Questions

6.1 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.

6.2 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.

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

A: This happens when the upper limit of the confidence "cloud" goes above the upper limit of the graph. To fix this, use *CUTOFF*.

6.4 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.

6.5 Q: How do I put more than one spline curve on a graph?

A: At the moment, the macro does not accommodate 'by' variables. To put more than one curve on a single graph, you need to run %LGTPHCURV9 to get each curve. If the curves are for subsets of a larger dataset, you can do this using the *WHERE* parameter. Don't forget to use *EXTRAV* if a variable occurs in the *WHERE* parameter but nowhere else. In each run of %LGTPHCURV9, you set *PLOT*=4 and *PLOTDATA* to something that will remind you of which group you are dealing with (e.g. smokca.men and smokca.women). Once you have the plotting information, you can plot both in excel, or read them into SAS and use SASGRAPH.

6.6 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.

6.7 Q: Why did SAS print a WARNING saying it ignored the ORDER= list on the axis statement?

A: This happens when you use AXORDVLOG10=T (see Invocation and Details above). It is harmless.

6.8 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.

6.9 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 paran
do over nums;  nums=0;  end ;  /*  effectively sets all sets of indicators to their reference l
/* special coding for continuous variables */
bmi76=25;  /* coding to a conventional cutoff */
run;
```

```
\section{Including the graph in a MS-WORD document}

Below are the steps for importing an encapsulated postscript file into a MS-WORD
document.
\begin{verbatim}
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).

7 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.

8 Credits

This macro is based on a restricted cubic spline macro originally written by Frank Harrell. Any questions should be addressed to Ruifeng Li via email strui@channing.harvard.edu or via phone 617-432-6321.

9 References

Some references are

Smith, Patricia L.: Splines as a useful and convenient
statistical tool. The American Statistician 33(2): 57-, 1979.

Harrell, Frank E, Jr., Lee, Kerry L., Pollock, Barbara G.:
Regression models in clinical studies: determining relationships

between predictors and response. JNCI 80: 1198-1202, 1988.

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.

A reference for the binned kernel density estimator (smoothed histogram) is

Wand, M.P., and Jones, M.C.: "Kernel Smoothing" (Appendix D). London: Chapman and Hall, 1995.

A reference discussing the presentation of the confidence bands is

Greenland, S., Michels, K.B., Robins, J.M., Poole, C., Willett, W.C.: Presenting statistical uncertaintly in trends and dose-response relations. Am J Epidemiol 149: 1077-1086, 1999.