Paper ST09

A Collection of SAS Macros to Calculate Odds Ratios Using Spline Regression

Martin Gregory, Merck Serono, Darmstadt, Germany

1 Abstract

In clinical and epidemiologic research investigating dose-response associations, non-parametric spline regression has long been proposed as a powerful alternative to conventional parametric approaches, since no underlying assumptions of linearity have to be fulfilled. For logistic spline models, however, to date, little standard statistical software is available to estimate any measure of risk, typically of interest when quantifying the effect of a continuous explanatory variable on a binary disease outcome. In the present paper, we describe a set of SAS^{TM} macros which perform non-parametric logistic regression analysis with b-spline expansions of an arbitrary number of continuous exposure variables, estimating adjusted odds ratios with respective confidence intervals for any given value with respect to a supplied reference value. The macros further allow graphical visualization of the shape of the association, retaining the exposure variable under consideration in its initial, continuous form while concurrently adjusting for multiple confounding factors. We discuss the internals of the macros and demonstrate their application macros by investigating the effect of body-mass index on risk of cancer incidence in a large, population-based male cohort.

2 Introduction

Conceivably due to a lack of appropriate tools in standard statistical software packages, the investigation of risk ratios for the effects of explanatory variables on disease outcomes in modern clinical and epidemiologic research is widely based upon splitting the range of a continuous exposure variable into several categories, in order to handle non-linearity in dose-response [1, 2, 3].

Resulting step function analyses have, however, long been associated with several problems, most notably discontinuities in risks at category cut points not being biologically plausible and violating the assumption of actual risks varying smoothly with data [5, 6, 7]. Moreover, it has been shown that categorization may be associated with cut-point bias and considerable misclassification [1, 3, 8, 9]. Step functions are known to provide a rather crude approximation of the true relationship by not using the full range of exposure data available to estimate associations; this, in turn, leads to a reduction of statistical power and less precise regression estimates [2, 5, 6].

Considering the above limitations, more efficient methods of statistical data analysis have been introduced in order to investigate joint covariate effects when linearity in dose-response cannot necessarily be assumed. Particularly, spline regression [10, 11, 12] has been proposed as a cogent alternative to conventional parametric regression approaches, since no underlying assumptions of linearity have to be fulfilled. In contrast to categorical models, spline models use the full range of exposure data to estimate the shape of the association and are known to be especially beneficial when the covariate effect is more pronounced at extremes of the exposure distribution [13, 14].

Despite obvious benefits of spline regression, to date little standard statistical software is available to estimate any measure of risk, typically of interest when quantifying the effects of continuous exposure variables on a disease outcome. In particular, the task of calculating odds ratios for a covariate in a logistic regression when the covariate has been replaced by a spline expansion cannot be done using a single SAS procedure but must employ several procedure and data steps in order to complete the calculations. In the present paper, we describe a set of SAS macros which perform non-parametric logistic regression analyses with b-spline expansions of an arbitrary number of continuous explanatory variables, estimating adjusted odds ratios with their respective confidence intervals for any given value with respect to a supplied reference value.

The collection consists of the following four macros:

%regspline performs the spline expansion, the regression and calculates the odds ratios and associated confidence limits, adding these to the standard output from the regression;

%regspline_plot plots odds ratios and associated confidence limits against the values of the independent variable. Reference lines may also be displayed;

%regspline_or writes the original variable, the odds ratios and associated confidence limits to a file or a SAS dataset. It can also select subsets of the data based on quantiles;

%regspline_subset writes the original variable, the odds ratios and associated confidence limits for specified values of the original variable to a SAS dataset.

The complete set of macros, along with detailed documentation, are available on request from the author.

3 Theoretical Background

Splines are generally defined to be piecewise of polynomials of degree d whose function values and first $d-1$ derivatives agree at the points they join. Using splines in simple or multiple regression analyses allows investigation of non-linear effects of continuous covariates. Let $(\xi_0 =)a < \xi_1 < \xi_2 < \ldots < \xi_k < b(= \xi_{k+1})$ be a subdivision by K distinct points on the interval $[a, b]$ on which the x variable is valued. These points are called "knots" of the spline function $s(x)$ used to transform the x variable. Knots correspond to break points in linearity which have to be optimized; in medical or epidemiological applications, knot points may represent threshold values of a risk factor for which the probability of a disease suddenly changes. The degree of polynomial pieces and the number as well as location of knots may vary, with knot points giving the curve freedom to bend and more closely follow the data. Since analyses using splines is often cumbersome and interpretation may be complex, it is necessary to compare the trade-off between model complexity (i.e. degrees of freedom of the spline and number of knots) and model fit in order to assess whether a much more complex model provides a significantly better fit. For a fixed sequence of knots $\xi = (\xi_1, \xi_2, \ldots, \xi_k)$, the set of splines is a linear space of functions with $K + d + 1$ free parameters. A useful basis for this linear space is given by Schoenberg's b-splines, or basic-splines, which are numerically well-conditioned and achieve local sensitivity to the data. A more detailed introduction to spline regression can be found elsewhere. An extension of spline methods to logistic regression models was introduced by Devlin and Weeks [15]. The formal calculation of odds ratios from logistic regression models using a b-spline expansion of a continuous, independent variable was described in detail by Cao and colleagues [4] and implemented by them as an S-Plus function. The odds ratio (or) for the predictor x with respect to a reference value x_{ref} is calculated as

$$\widehat{or}(x, x_{ref}) = exp\Big(\sum_{i=1}^{n} \hat{\beta}_i[s_i(x) - s_i(x_{ref})]\Big) \tag{1}$$

where n is the number of degrees of freedom of the spline expansion, $\hat{\beta}_i$ is the coefficient of the i^{th} spline basis function estimated by the logistic regression and $s_i(x)$ is the value of the i^{th} spline basis function at x. The limits of the related 95% confidence interval are calculated as:

$$exp(log(\widehat{or}(x, x_{ref})) \pm 1.96 \times \hat{\sigma}_{log(or)}) \tag{2}$$

where $\hat{\sigma}^2_{log(or)}$ is the variance of $log(\widehat{or}(x, x_{ref}))$ calculated as $s'Cs$ where s is the vector

$$(s_1(x) - s_1(x_{ref}), \ldots, s_n(x) - s_n(x_{ref})) \tag{3}$$

and C is the covariance matrix of the regression coefficients.

4 Application of the Macros

We demonstrate the use of our SAS macros in estimating the effect of body-mass index (BMI) on the risk of cancer incidence in a population-based male cohort. Our data are derived from the Vorarlberg Health Monitoring and Promotion Program (VHM&PP), one of the world's largest ongoing population-based risk factor surveillance programs, linked to disease specific morbidity and mortality outcomes. The VHM&PP is routinely performed by the Agency for Social and Preventive Medicine and addresses all adults of the entire province. It routinely includes the recording of socio-demographic data and a physical examination with a fasting blood sample. A more detailed description of the program methodology and study population has been reported previously [18, 19, 20].

Briefly, the study sample consisted of 73,624 men and 5,305 incident cancers during 19 years of follow-up. Baseline information on age and BMI (both continuous variables), smoking status (current, former, never) and occupational status (blue collar, white collar, self-employed) was collected for each study participant. In our example, we considered b-spline expansions of both BMI and age in order to demonstrate the use of our macros in simultaneously employing spline methods on more than one continuous covariate and to gain sufficient flexibility, also in modelling of the effect of the continuous confounding variable age. All reported estimates additionally were adjusted for smoking status and occupational status (both categorical variables). A BMI of 21.75 kg/m2, as the mid-point of the reference range of normal BMI values, was used for the calculation of odds ratios in all models.

Considering the illustrative character of our analyses and the fact that issues of model selection in general remain a major challenge far beyond the scope of the present paper, we confined our analysis on fitting linear (df=1) and quadratic (df=2) spline models with knots at consecutive quartiles (3 knots) and quintiles (4 knots) of the BMI and age distribution and abstained from fitting more complex spline models to our data. Note however, that it was reported that in practice, few well-located knots generally suffice in most cases, although deciding on their optimal location is difficult and no generally accepted criteria yet exist [21]. The best-fitting combination of spline expansions of BMI and age was defined as the one minimizing the Akaike's Information Criterion (AIC), a penalized likelihood that takes into account the number of parameters estimated in the model, compromising between a good fit and a parsimonious model [22]. It has been shown that AIC may perform poorly for small samples or in presence of too many parameters in relation to the size of the sample; however, none of these conditions affected our investigation.

In order to determine the AIC-minimizing, best-fitting combination of spline expansions of BMI and age, we called the %regspline macro for each of the combinations of 3 and 4 knots with degrees 1 and 2. A representative call, for 3 knots and degree 2 of BMI and 4 knots and degree 1 of age, is:

```
%regspline(data=sd.bmi_males,out=sd.k3d2_k4d1,
           depvar=event,splinevar=bmi age,indepvars=smoker emptype,
           nknots=3 4,degree=2 1,fit=fit_k3d2_k4d1,or=or_bmi or_age,
           xref=21.75 39,xref_fuzz=0 0) ;
```

This call reads the input dataset, carries out a logistic regression of the binary dependent variable on a second degree spline expansion with three knots of the BMI variable and a first degree spline expansion with four knots of age, with smoking status (3 categories) and occupational status (3 categories) as covariates. The parameters used in this call are as follows:

data names the input dataset, here sd.bmi_males;

out names the output dataset, here sd.k3d2_k4d1. This dataset contains the original variables along with the b-spline expansion variables, the predicted values, the estimated regression coefficients for each of the independent variables (original variables prefixed with a "cf_"), the estimated adjusted odds ratios compared to the specified BMI reference value of 21.75 (or_bmi) and the associated lower (or_bmi_lcm) and upper (or_bmi_ucm) confidence limits for each given value of BMI.The dataset has one observation for each observation on the input dataset;

depvar names the dependent variable on the input dataset. In order to have the LOGISTIC procedure model the event of cancer occurence (event=1), we had to assign a format with the values 0='No disease' and 1='Cancer' as described in the LOGISTIC documentation [17];

splinevar lists the independent variables which are to be replaced by a spline expansion;

indepvars lists the remaining independent variables to be included in the model;

nknots lists the number of internal knots to be used when generating each spline expansion. The value should satisfy the requirements for the NKNOTS model option of the TRANSREG procedure;

degree lists the degree of the polynomial to use in generating each of the spline expansions;

or names the variables on the output dataset in which to write the estimated odds ratio value for the spline expansion variables. The default is to prefix the spline expansion variable names with *or_*. This parameter also determines the names of the upper and lower confidence limit variables which are named by adding the suffix *_ucm* and *_lcm*, respectively;

fit saves the model fit values to the named dataset;

xref lists the reference values for each spline variable to be used as denominator when calculating the odds ratios;

xref_fuzz Specifying a non-zero value for $xref\_fuzz_i$ causes the macro to search for a value in the interval $[xref_i - xref\_fuzz_i, xref_i + xref\_fuzz_i]$. If any are found, the one closest to $xref_i$ is used. If none are found, the macro stops executing.

Using the second macro, %regspline_plot we can graph the estimated adjusted odds ratios for BMI along with the corresponding confidence limits, for the AIC-minimizing, best-fitting spline model (AIC = 31388.264; BMI: df=1, knots=3; age: df=2, knots=2), using the following call:

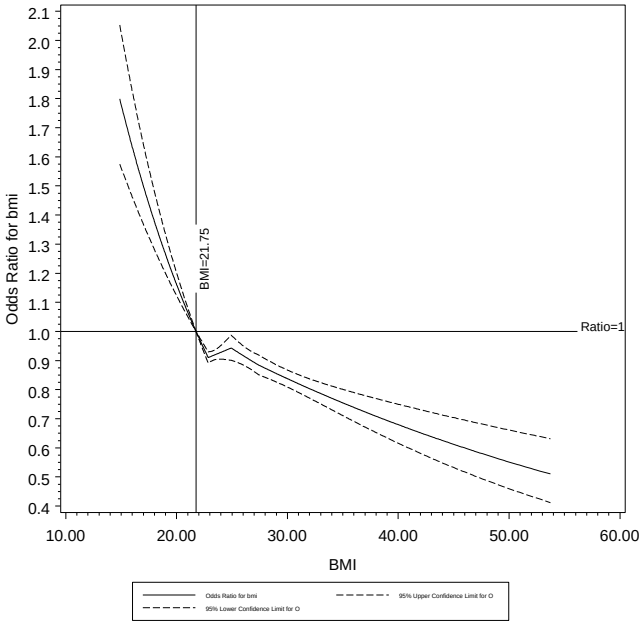

Figure 1: Output of the %regspline_plot macro

```
%regspline_plot(data=sd.k3d1_k2d2,color=n,
                xref=21.75,xref_label=BMI,
                splinevar=bmi,riskvar=or_bmi,
                risklcl=or_bmi_lcm,riskucl=or_bmi_ucm,
                axisheightv=1.25,axisheightl=1.25,legendheight=0.5,
                linewidth=1.75) ;
```

The graph produced by this call is shown in Figure 1. The input dataset was produced by the %regspline macro. The *xref* and *xref_label* parameters specify the position and label for the vertical reference line. The *splinevar*, *riskvar*, *risklcl* and *riskucl* parameters allow specification of the covariate, risk and confidence limit variables, respectively. The *color* parameter controls whether coloured lines are used instead of dashed lines. The *axisheightv*, *axisheightl* and *legendheight* parameters control the text size of the axis values, labels and the legend text, respectively. These values specify a factor to be applied to the default text height. Finally, we have used the *linewidth* parameter to draw lines 75% thicker than the default used by SAS/GRAPH^{TM}.

For application of the macros investigating the association of gamma-glutamyltransferase and risk of cancer incidence, see the papers by Strasak and colleagues [23, 24].

5 Details of the %regspline Macro

This section concentrates on the core macro, %regspline. In brief, this macro uses the TRANSREG procedure [16] to generate the b-spline expansion, the LOGISTIC procedure [17] to carry out the regression and a DATA step in order to calculate the odds ratios for the original covariates. The remaining macros are fairly straight-forward and are provided mainly for convenience. The following sections describe the structure of the %regspline macro in some detail.

5.1 Parameter Processing and Validation

The first section of the macro is concerned with processing and validating the parameters passed. The macro takes a conservative approach, avoiding where possible running the code with invalid parameters. Validation includes:

- checking that required parameters are present;
- checking that one and only one of the parameters *knots* and *nknots* are present;

- verifying that the *knots*/*nknots*, *degree*, *xref* and *xref_fuzz* parameters have the same number of values as the *splinevar* parameter and that each of the values are numeric;
- checking that the names for the output datasets are valid dataset names. This is necessary because the datasets may not exist when the macro is called. The check is done using macro %sasname included with the package. While there is a SAS-provided (SCL) function, sasname(), this has a number of problems. First it treats a two-level dataset name as invalid because it contains a period. Second, sasname() assumes that 32 characters are allowed in names, but both library and file references are still limited to 8 characters.

Parameter processing also includes defining a number of lists of variable names and quoted variable names. The former are using in KEEP and RENAME lists and in arrays in the DATA steps which perform calculations, while the latter are used in a WHERE clause when reading the covariance matrix. Since the macro performs spline expansions of multiple variables, multi-dimensional arrays are used. This lead to an issue where a different number of knots or degree are specified for different variables resulting in a different number of expansion elements. Since a variable-based multi-dimensional array requires the same number of elements for each value of a dimension, it was necessary to work with a possibly repeated dummy variable to pad out the arrays while at the same time storing which variables should actually be used in the calculations. We will elaborate on this issue in the discussion on the calculation of the odds ratios in section 5.4

5.2 Calculating the Spline Expansions

The spline expansions are done using the TRANSREG procedure which is capable of performing regressions with spline expansions. It does not, however, offer logistic regression. Using the DESIGN option on the PROC TRANSREG statement allows us to specify a "model" without a dependent variable and causes the TRANSREG procedure to simply output the spline expansions. For example the call to %regspline:

```
%regspline(data=sd.gbsg2,out=sd.Ak2d1_Sk3d2,
 depvar=cens,splinevar=tsize age,indepvars=pnodes progrec estrec,
 nknots=2 3,degree=1 2,or=rrsize rrage,xref=25 52,xref_fuzz=0 0) ;
```

generates the following code:

```
proc transreg data=sd.gbsg2 design ;
 model bspline(tsize / nknots=2 degree=1) bspline(age / nknots=3 degree=2) ;
 output out=__spline ;
 id cens pnodes progrec estrec ;
run ;
```

The temporary output dataset *__spline* contains all variables listed on the ID statement, the *tsize* and *age* variables and their spline expansions. Since the expansion produces $nknots + degree + 1$ variables, the output dataset will contain variables *tsize_0* to *tsize_3* and *age_0* to *age_5*. The number of observations is identical to the input dataset.

In order to calculate the odds ratios, we also need to determine the spline expansions for the reference values. This is done in a data step which searches the output of the TRANSREG procedure for the reference values. If an exact matching value is not found, the macro stops execution. Using the $xref\_fuzz$ parameter, it is possible to specify an approximate match. Specifying a non-zero value for $xref\_fuzz_i$ causes the macro to search for a value in the interval $[xref_i - xref\_fuzz_i, xref_i + xref\_fuzz_i]$. If any are found, the one closest to $xref_i$ is used. If none are found, the macro stops executing.

5.3 Performing the Regression

A LOGISTIC step is used to perform the regression. The MODEL statement is automatically generated using the value of the *depvar* parameter as dependent variable with the spline expansions of the variables named in the *splinevar* parameter and the non-expanded covariates named in the *indepvars* parameter as independent variables. Since the calculation of the confidence intervals for the odds ratio estimates requires the covariance matrix, the COVOUT option is used on the PROC LOGISTIC statement.

Note that the dependent variable must be coded appropriately in order for the LOGISTIC procedure to choose the correct value as the event. See the documentation of the MODEL statement [17] for details. Otherwise the macro provides the flexibility of allowing the caller to specify any desired model options using the *model_opts* parameter.

Additionally, if the fit parameter is used an ODS OUTPUT statement is generated to save the fit statistics for the model. This can be useful in comparing different models.

5.4 Calculating the Odds Ratios and Confidence Intervals

This section of the macro merges the estimates output by the LOGISTIC step with the covariance matrix and the dataset containing the spline expansions for the reference values. This is done in a DATA step using arrays and nested loops to implement the matrix operations implicit in the formulas (1) to (3). While use of SAS/IML™ would have made it possible to write more elegant and easy to read code, the implementation in a DATA step means that macro is not dependent on the caller having access to SAS/IML.

The calculation of the odds ratio, $\_or\{k\}$, for the k^{th} variable is done using the following statements to implement formula (1):

```
__md=0 ;
do i=1 to se_n1{k} ;
   __md=sum(__md,cf_ls{k,i}*(se_ls{k,i}-xr_ls{k,i})) ;
   end ;
_or{k}=exp(__md) ;
```

where

- $se\_n1\{k\}$ is the number of variables in the spline expansion for the k^{th} variable;
- $cf\_ls\{k,i\}$ is the i^{th} regression coefficient of the k^{th} variable;
- $se\_ls\{k,i\}$ is the i^{th} spline expansion variable of the k^{th} variable; and
- $xr\_ls\{k,i\}$ is the i^{th} spline expansion variable of reference value for the k^{th} variable;
- $\_or\{k\}$ is the name, specified in the parameter *or*, for the variable in which to store the odds ratio for the k^{th} spline expansion variable.

The array *se_ls* is defined using the statement:

```
array se_ls {&sv_n,&se_max} &se_ls ;
```

with one row for each variable for which a spline expansion is to be performed. The list of expansion variables in the macro variable *se_ls* may include dummy variables if the dimension of the spline bases differ. For example, the parameter values

```
splinevar=ht wt,deg=2 2,nknots=2 3
```

would result in the following array definitions for *se_ls* and *se_n1*:

```
array se_ls {2,6} ht_0 ht_1 ht_2 ht_3 ht_4 __rs_dummy
                  wt_0 wt_1 wt_2 wt_3 wt_4 ht_5 ;
array se_n1 {2} _temporary_ (5 6) ;
```

The confidence intervals for the k^{th} spline expansion variable are calculated according to formula 2 as follows:

```
__var_md=0 ;
do i=1 to se_n1{k} ;
   do j=1 to se_n1{k} ;
      __var_md=sum(__var_md,
         (se_ls{k,i}-xr_ls{k,i})*cov{k,i,j}*(se_ls{k,j}-xr_ls{k,j})) ;
      end ;
   end ;
__se_md=sqrt(__var_md/se_n1{k}) ;
_orl{k}=exp(__md - 1.96*__se_md) ;
_oru{k}=exp(__md + 1.96*__se_md) ;
```

where $cov\{k,i,j\}$ is the covariance matrix for the k^{th} spline expansion variable and $\_orl\{k\}$ and $\_oru\{k\}$ are the names on the output dataset for the lower and upper confidence interval variables, respectively.

5.5 Error Handling

Since the %regspline macro is likely to be used in non-interactive processing, for example, in a loop over different numbers of knots and degrees to determine the best fitting models, it is important that the calling program can detect errors which occur during execution and take appropriate action in order to avoid wasting time. As an illustration, comparison of six different combinations of number of knots and degrees of freedom for spline expansion of a single variable on a dataset with 75,000 observations, running on a 2.4GHz AMD Athlon^{TM} 3800+ processor took 3.5 minutes of real time. While the SAS macro language does not directly support the provision of return codes, it is possible to achieve satisfactory results using global macro variables. In order to avoid possible naming collisions, %regspline allows the caller to name the global macro variable to use. During initialisation, this variable is set to a value of 9999 so that unexpected errors causing termination of the macro will not provide a false positive result. The following code fragment illustrates the use of this technique:

```
%regspline(data= ..., rc=check) ;
%if &check ^= 0 %then %do ;
```

In addition to providing return codes related to invalid or missing parameter specification, the value of the automatic macro variable SYSERR is checked after each step is executed. If SYSERR has a value greater than or equal to 4, an appropriate message is printed in the log, the return code is set to a non-zero value and %regspline terminates.

Unless %regspline terminates unexpectedly, all intermediate datasets are deleted. It is possible to prevent this deletion by setting the parameter _DEBUG to 1. This has, however, the side effect of switching on the MPRINT, MLOGIC and SYMBOLGEN options and of printing various intermediate values used in the calculations. On termination the values of these options are reset to those in effect before %regspline was called.

Finally, all macro variables defined in %regspline are declared local to avoid possible collisions with macro variables defined in the calling environment.

6 Conclusions

Using the features provided by the macro, it is possible to easily perform non-parametric logistic regression analysis with b-spline expansions of an arbitrary number of continuous exposure variables, to estimate adjusted odds ratios for any given value with respect to a supplied reference value for each exposure variable and to calculate the corresponding confidence intervals. Calls to the macro can be included in an enclosing macro which loops over different degrees and numbers of knots for the spline expansion, permitting the user to automate selection of the best model.

The macro does have a number of limitations. The reference values for calculating the odds ratios must exist on the input dataset in order that the spline expansion for the reference value can be calculated. While the $xref\_fuzz$ parameter allows the selection of a value close to the desired one, this is not an entirely satisfactory solution. A further limitation is that it is necessary to code the response variable appropriately, or apply a format, in order that the LOGISTIC procedure models the desired event. Finally, it is not possible to specify a confidence interval other than 95%. The latter two limitations could be relatively easily overcome by providing additional parameters, the first, however, may require substantial effort to fix.

References

[1] D.G. Altman, Categorizing continuous variables, *Br. J. Cancer* 64 (1991) 975.

[2] K.M. Boucher, M.L. Slattery, T.D. Berry, C. Quesenberry, K. Anderson, Statistical methods in epidemiology: a comparison of statistical methods to analyze dose-response and trend analysis in epidemiologic studies, *J. Clin. Epidemiol.* 51 (1998) 1223-1233.

[3] S. Greenland, Dose-response and trend analysis in epidemiology: alternatives to categorical analysis, *Epidemiology* 6 (1995) 356-365.

[4] Cao J, Valois MF, Goldberg MS. An S-Plus function to calculate relative risks and adjusted means for regression models using natural splines. *Comput Methods Programs Biomed* 2006;84:58-62.

[5] C. Weinberg, How bad is categorization?, *Epidemiology* 6 (1995) 345-347.

[6] S. Greenland, Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis, *Epidemiology* 6 (1995) 450-454.

[7] L.P. Zhao, L.N. Kolonel, Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies, *Am. J. Epidemiol.* 136 (1992) 464-474.

[8] F.E. Harrell, K.L. Lee, B.G. Pollock, Regression models in clinical studies: determining relationships between predictors and response, *J. Natl. Cancer Inst.* 80 (1988) 1198-1202.

[9] M. Maclure, S. Greenland, Tests for trend and dose response: misinterpretations and alternatives. *Am. J. Epidemiol.* 135 (1992) 96-104.

[10] E.J. Wegman, I.W. Whright, Splines in statistics, *J. Am. Stat. Assoc.* 78 (1983) 351-365.

[11] R.L. Eubank, *Spline Smoothing and Nonparametric Regression* (Marcel Dekker, Berlin, 1988).

[12] K.J. Rothman, S. Greenland, *Modern epidemiology, 2nd ed.* (Lippincott-Raven Publishers, Philadelphia, 1998).

[13] P.S. Rosenberg, H. Katki, C.A. Swanson, L.M. Brown, S. Wacholder, R.N. Hoover, Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge, *Stat. Med.* 22 (2003) 3369-3381.

[14] J. Polesel, L. Dal Maso, V. Bagnardi, A. Zucchetto, A. Zambon, F. Levi, C. La Vecchia, S. Franceschi, Estimating dose-response relationship between ethanol and risk of cancer using regression spline models, *Int. J. Cancer* 114 (2005) 836-841.

[15] T.F. Devlin, B.J. Weeks, Spline Functions for Logistic Regression Modeling, in *Proceedings of the 11th Annual SAS Users Group International Conference*, pp. 646-649 (SAS Institute Inc., Cary, NC, 1986).

[16] SAS Institute Inc., The TRANSREG Procedure, *SAS Online Documentation, 9.1.3* (2003), http://support.sas.com/onlinedoc/913/getDoc/en/statug.hlp/transreg index.htm.

[17] SAS Institute Inc., The LOGISTIC Procedure, *SAS Online Documentation, 9.1.3* (2003), http://support.sas.com/onlinedoc/913/getDoc/en/statug.hlp/logistic index.htm.

[18] H. Ulmer, C. Kelleher, G. Diem, H. Concin, Long-term tracking of cardiovascular risk factors among men and women in a large population-based health system: the Vorarlberg Health

[19] A. Strasak, K. Rapp, W. Hilbe, W. Oberaigner, E. Ruttmann, H. Concin, G. Diem, K.P. Pfeiffer, H. Ulmer & the VHM&PP Study Group, Serum uric acid and risk of cancer mortality in a large prospective male cohort, *Cancer Causes Control* 18 (2007) 1021-1029.

[20] A. Strasak, E. Ruttmann, L. Brant, C. Kelleher, J. Klenk, H. Concin, G. Diem, K.P. Pfeiffer, H. Ulmer & the VHM&PP Study Group, Serum uric acid and risk of cardiovascular mortality: A prospective long-term study in 83683 Austrian men, *Clin. Chem.* (2007, Nov 26, epub ahead of print).

[21] N. Molinari, J.P. Daurs, J.F. Durand, Regression Splines for threshold selection in survival data analysis, *Stat. Med.* 20 (2001) 237-247.

[22] H. Akaike, Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, pp. 267-281 (Budapest, Second International Sypmposium of Information Theory, 1973).

[23] AM Strasak, K Rapp, LJ Brant, W Hilbe, M Gregory, W Oberaigner, E Ruttmann, H Concin, G Diem, KP Pfeiffer, H Ulmer and the VHM&PP Study Group. Association of gamma-glutamyltransferase and risk of cancer incidence in men: a prospective study. *Cancer Res.* 2008 May 15;68(10):3970-7.

[24] AM Strasak, RM Pfeiffer, J Klenk, W Hilbe, W Oberaigner, M Gregory, H Concin, G Diem, KP Pfeiffer, E Ruttmann, H Ulmer; Vorarlberg Health Monitoring and Promotion Program Study Group, Prospective study of the association of gamma-glutamyltransferase with cancer incidence in women. *Int J Cancer* 2008 Oct 15;123(8):1902-6.

Contact Information

Martin Gregory
Head of Statistical Programming Germany
Merck KGaA
Frankfurterstr 250
64293 Darmstadt
Germany
Martin.Gregory@merck.de