An S-Plus function to calculate relative risks and adjusted means for regression models using natural splines

Jiguo Cao\textsuperscript{a}, Marie-France Valois\textsuperscript{b,\text{*}}, Mark S. Goldberg\textsuperscript{b}

\textsuperscript{a} Department of Mathematics and Statistics, McGill University, Montreal, Quebec H3A 1A1, Canada
\textsuperscript{b} Division of Clinical Epidemiology, McGill University Health Center-RVH, 687 Pine Avenue West, R4.29, Montreal, Quebec H3A 1A1, Canada

\section*{1. Introduction}

Modern statistical analysis of data arising from epidemiologic studies make extensive use of multiple regression techniques to estimate associations between the dependent and explanatory variables [1,2]. Regression models incorporated in the generalized linear models (GLM) [3] provide estimates of the association between various types of dependent and independent variables while accounting for the effects of covariates that may confound the associations under study. For continuous independent variables, researchers have relied generally on parametric representations to model effects. For example, it is often assumed that the relationship between a covariate and the outcome variable is a linear function. If the data are not expected to follow the assumed linear relationship, then some simple parametric transformations of the independent variable are often considered [4]. The limitation of estimable relationships to a few parametric curves may lead to bias, loss of decency, or incorrect conclusions [5–8]. An alternative strategy to avoid the linearity assumption is to simply break the range of the continuous variable into categories and then at the model using the newly created categorical variable [4,2]. This method also entails a loss of information and can introduce considerable misclassification, especially if the selected cutpoints do not follow the empirical response function [9,1,10–13]. In addition, selecting cutpoints to optimize the fit results in a systematic over-estimation of the covariate effect and inflates the type I error of testing the hypothesis of no association [14,15]. Similar strategies are also taken while modeling the effects of continuous confounding variables, with the result that the estimated association may be biased if its confounding effects are not removed entirely [1,16].

\textsuperscript{\text{*} Corresponding author. Tel.: +1 514 934 1934x36917; fax: +1 514 843 1493.
E-mail address: mark.goldberg@mcgill.ca (M.-F. Valois).
0169-2607/$ – see front matter © 2006 Published by Elsevier Ireland Ltd.
doi:10.1016/j.cmpb.2006.08.004

Please cite this article as: Jiguo Cao et al., An S-Plus function to calculate relative risks and adjusted means for regression models using natural splines, Computer Methods and Programs in Biomedicine (2006), doi:10.1016/j.cmpb.2006.08.004
Numerous investigators have recognized these issues and consequently have developed methods that allow the analyst a number of interesting options to both visualize response functions as well as obtain quantitative estimates of association, including parametric natural cubic splines [17], penalized splines [18], and arrange non-parametric smoothers incorporated within the context of the generalized additive models (GAM) framework [6]. The GAM models gained considerable use in the 1990s in the air pollution field, where non-parametric smoothers were used to filter time series of mortality (in order to remove unwanted long-term variations in the dependent series and to remove serial autocorrelation and overdispersion) [19] as well as to characterize non-linear response functions. However, a few years ago, it was discovered that the backfitting algorithm used to maximize these GAMs did not convergence appropriately and did not account adequately for non-linearities between explanatory variables (concurvity), with the result that the estimates of effect were biased and the standard errors were underestimated [20–22]. Although the former problem has been fixed satisfactorily, but not the problem with bias in the parameter estimates, we and our colleagues [23,24] and other investigators have replaced the non-parametric smoothers in the GAMs with natural cubic splines.

In interpreting the findings from analyses that show non-linear response functions, it is useful to present the results in graphical or tabular format so that the estimates of effect across a range of values of the independent variable are compared to one common reference value. For the GAMs, Sæz et al. developed an S-Plus function that can be used for this purpose [25]. Following their lead, we have developed two analogous S-Plus functions when natural cubic splines are used in a GLM that uses a logarithmic link function or the identity link.

The calculation of relative risks from models incorporating natural cubic splines

A piecewise cubic spline $S(x)$ on the interval $[a, b]$ is a continuous piecewise curve with cubic functions $S_k(x)$ on each interval $[x_k, x_{k+1}]$ with its first and second derivatives are all continuous on $[a, b]$, where $a = x_0 < x_1 < \ldots < x_n = b$, and $x_0, x_1, \ldots, x_n$ are called knots or break points. A natural spline is a piecewise cubic spline when its second derivative is 0 at the boundary points $[a, b]$. Generally, if we have $n + 1$ knots (including boundary points), we can construct $n$ natural spline basis functions which are orthogonal with each other, and any spline function defined by these knots can be expressed as a linear combination of these basis functions. Here $n$ is called the number of degrees of freedom (df) of the natural spline.

**Definition 2.1.** Suppose we have knots $x_0, x_1, \ldots, x_n$, where $a = x_0 < x_1 < \ldots < x_n = b$, then the function $S(x)$ on $[a, b]$ is called a natural spline defined by these knots, if there exist $n$ cubic polynomials $S_k(x)$ with coefficients $s_{k0}, s_{k1}, s_{k2},$ and $s_{k3}$ that satisfy the following six conditions:

1. $S(x) = S_k(x) = s_{k0} + s_{k1}(x - x_k) + s_{k2}(x - x_k)^2 + s_{k3}(x - x_k)^3$ for $x \in [x_{k-1}, x_k]$, for $k = 1, \ldots, n$
2. $S_k(x_k) = S_{k+1}(x_k)$, for $k = 1, \ldots, n - 1$
3. $S_k'(x_k) = S_{k+1}'(x_k)$, for $k = 1, \ldots, n - 1$
4. $S_k''(x_k) = S_{k+1}''(x_k)$, for $k = 1, \ldots, n - 1$
5. $S(a) = 0$
6. $S''(a) = S''(b)$

The GLM function using natural splines can be expressed as:

$$E(g(Y)) = \alpha + \sum_{i=1}^{n} \hat{c}_i S_i(x)$$

where $Y$ is the dependent variable (for normally distributed data it will be a continuous variable, for Poisson data it will be equal to counts or rates, and for a logistic model it will represent the odds of developing the outcome), $g(\cdot)$ the link function (identity or the natural logarithm), $\alpha$ the intercept, $S_i(x)$, $i = 1, 2, \ldots, n$ natural splines, $c_i, i = 1, 2, \ldots, n$ the associated regression coefficients, and $n$ is the df of natural splines. We assume that the intercept of the natural spline on the left most knot is equal to zero (condition 5 in Eq. (1)).

For the identity link, the adjusted mean difference is

$$\tilde{MD}(x, x_{ref}) = \sum_{i=1}^{n} \hat{c}_i[S_i(x) - S_i(x_{ref})]$$

and for the log link, the ratio of the outcome variable (often referred to as the relative risk, RR) for a predictor $x$ with respect to $x_{ref}$ is estimated as

$$\tilde{RR}(x, x_{ref}) = \exp \left\{ \sum_{i=1}^{n} \hat{c}_i[S_i(x) - S_i(x_{ref})] \right\}$$

where $x_{ref}$ is a reference value of the predictor, and $\hat{c}_i, i = 1, 2, \ldots, n$ is the estimated regression coefficients from the GLM. Another formulation used, when the RR is small, is to express it as a mean percent change (MPC):

$$\tilde{MPC}(x, x_{ref}) = \exp \left\{ \sum_{i=1}^{n} \hat{c}_i[S_i(x) - S_i(x_{ref})] - 1 \right\} \times 100\%.$$
3. **Use of the S-Plus functions**

The main function `plot.ns.ci.prg` produces a publication–quality graph of MD, RR, log RR, or MPC and the associated confidence limits for an independent variable with respect to a specified reference value `x_ref`. This function has 17 arguments, as follows:

- `data`: Original S-Plus data set.
- `variable`: Name of the predictor variable.
- `fit`: The fitted object from the generalized linear model.
- `pos`: The position where the first coefficient of the predictor variable begins in GLM summary.
- `df`: Number of degrees of freedom (df) for the predictor variable. If no specified knots are given, the program will select the df + 1 knots (including boundary knots) at suitably chosen quantiles of the predictor variable.
- `knots`: Knots that define the natural spline, including boundary points. When knots and df are both given, the number of knots is equal to df + 1.
- `xref`: The reference predictor value.
- `typeout`: You can choose “MD” (identity link), “LOGRR”, “RR” or “MPC” (log link) to output log Relative Risk, the Relative Risk, or the Mean Percent Change, respectively. The default is “RR” (log link).
- `step`: The step length to generate a vector of predictor values to obtain their MD, RR, or log RR, or MPC. The smaller the step, the more continuous the graph looks, but the longer it will take to run the program. The default is 0.1.
- `xtitle`: The title of the x-axis. The default is the predictor variable (in uppercase).
- `ytitle`: The title of the y-axis. The default is the argument `typeout` (in uppercase).
- `level`: The level for the confidence interval. The default is 95%.
- `na.action`: It can be T or F. If `na.action = T`, the original regression model will exclude observations that are flagged as missing by `na.exclude`. If `na.action = F`, an error will be caused if missing values are present in data. The default is T.
- `col`: The color of lines.
- `lty`: The types of lines.
- `lwd`: The width of lines.
- `linkf`: The link function used in the model. It can be “log” or “identity”.

The auxiliary function `ns.ci.prg` computes RR, logRR, MPC, or MD and the confidence limit for a specified vector x of the independent variable as compared to a specified reference value `x_ref`. This function has 12 arguments, in which data, variable, fit, pos, df, knots, xref, typeout, level, na.action, and linkf are identical to those in the main function `plot.ns.ci.prg`, and x is a vector of values to be compared to `x_ref`.

4. **Example**

Our example illustrates the use of the log link function, and the data is derived from a time series analysis of daily non-accidental mortality in Montreal, 1984–2001, where we regressed daily counts of death using an overdispersed quasi-likelihood Poisson model on a natural spline functions of maximum daily temperature (maxtemp, df = 4), adjusting for seasonal and subseasonal trends in the mortality time series (day, df = 250), mean daily ozone (mo3, df = 2) and mean daily nitrogen dioxide (mno2, df = 4), day of the week (DOW), and a factor for calendar year (YearFact).

We conducted the analysis in S-Plus in three steps:

- **Step 1**: Generate the fitted object using `GLM`.
- **Step 2**: Produce a graph of the expected logarithm of the number of daily deaths in the fitted model with respect to maximum daily temperature (Fig. 1).
- **Step 3**: Produce a graph of RR in the number of daily deaths as compared to the reference value of 20°C (Fig. 2).

**Fig. 1** – The expected logarithm of the number of daily deaths according to mean daily temperature, with the respective pointwise 95% confidence limits shown as the dashed lines.
To produce the RR for a specified vector of maximum daily temperature (MDT) (e.g., 15, 30 °C) as compared to the reference value of 20 °C, the call to the function was: ns.ci.prg(data, "maxtemp", fit = pos = 2, df = 4, xref = 20, x = c(15, 30), typeout = "RR", level = 0.95, na.action = na.exclude, link = "log")

With the results that the RR = 0.99 and the respective 95% CI of 0.98–1.01 when MDT = 15 °C, and the RR = 1.11 and the respective 95% CI of 1.09–1.14 when MDT = 30 °C.

As one can see from Fig. 2, the value of the maximum likelihood estimate and associated confidence interval at the reference value of 20 °C are equal to unity. In this figure, the daily counts of deaths at maximum temperatures above 30 °C are significantly higher than those at 20 °C.

5. Mode of availability

The functions can be downloaded from http://www.med.mcgill.ca/epidemiology/goldberg/.

Acknowledgements

The work described in this paper was funded through a contract with Health Canada, number 4500083542. Dr. Goldberg, gratefully acknowledges receipt of an Investigator Award, from the Canadian Institutes for Health Research.

REFERENCES