COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 84 (2006) 58–62

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

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

*Jiguo Cao*[a], *Marie-France Valois*[b], *Mark S. Goldberg*[b,*]

[a] *Department of Mathematics and Statistics, McGill University, Montreal, Quebec H3A 1A1, Canada*
[b] *Division of Clinical Epidemiology, Department of Medicine, McGill University Health Center-RVH, 687 Pine Avenue West, R4.29, Montreal, Quebec H3A 1A1, Canada*

ARTICLE INFO

*Article history:*
Received 27 March 2006
Received in revised form
5 August 2006
Accepted 17 August 2006

*Keywords:*
Biometrics
Regression
Natural cubic splines
Generalized linear models
Exposure-response functions

ABSTRACT

We provide for generalized linear regression models that use natural cubic splines to model predictors an S-Plus function to calculate relative risks (RR), log relative risk (log RR), mean percent change (MPC) for continuous covariates modeled using a logarithmic link as well as adjusted means differences (MD) for the identity link. The function makes explicit use of the natural spline basis functions, the estimated coefficients for each natural spline basis function, and the fitted correlation matrix for the estimated coefficients and can thus accommodate any number of degrees of freedom. The main function produces a publication–quality graph of all of these quantities as compared to a user-specified reference value as well as the associated confidence limits. In another function, specific values of these statistics comparing a vector of values of the independent variable to the reference value can be calculated rather than plotted.



## 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 efficiency, 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 fit 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

* *Corresponding author*. Tel.: +1 514 934 1934x36917; fax: +1 514 843 1493.
E-mail address: mark.goldberg@mcgill.ca (M.S. Goldberg).
*URL:* http://www.med.mcgill.ca/epidemiology/goldberg/ (M.-F. Valois).
0169-2607/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2006.08.004

variables, with the result that the estimated association may be biased if its confounding effects are not removed entirely [1,16].

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 a range of 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 converge 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, Saez 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.

## 2. 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 $[x_k, x_{k+1}]$, 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:

$$
\begin{aligned}
&1\quad S(x) = S_k(x) = s_{k0} + s_{k1}(x - x_k) + s_{k2}(x - x_k)^2 + s_{k3}(x - x_k)^3 \\
&\qquad \text{where } x \in [x_{k-1}, x_k], \text{ for } k = 1, \cdots, n \\
&2\quad S_k(x_k) = S_{k+1}(x_k), \text{ for } k = 1, \cdots, n-1 \\
&3\quad S'_k(x_k) = S'_{k+1}(x_k), \text{ for } k = 1, \cdots, n-1 \\
&4\quad S''_k(x_k) = S''_{k-1}(x_k), \text{ for } k = 1, \cdots, n-1 \\
&5\quad S(a) = 0 \\
&6\quad S''(a) = S''(b)
\end{aligned}
\tag{1}
$$

The GLM function using natural splines can be expressed as:

$$E(g(Y)) = \alpha + c_1 S_1(x) + c_2 S_2(x) + \cdots + c_n S_n(x) \tag{2}$$

where Y is the dependent variable (for normally distributed data it will be a continuous variable, for Poisson data it will be either counts or rates, and for a logistic model it will represent the odds of developing the outcome), $g(.)$ 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

$$\widehat{\mathrm{MD}}(x, x_{\mathrm{ref}}) = \sum_{i=1}^{n} \hat{c}_i [S_i(x) - S_i(x_{\mathrm{ref}})] \tag{3}$$

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 $x_{\mathrm{ref}}$ is estimated as

$$\widehat{\mathrm{RR}}(x, x_{\mathrm{ref}}) = \exp\left\{ \sum_{i=1}^{n} \hat{c}_i [S_i(x) - S_i(x_{\mathrm{ref}})] \right\} \tag{4}$$

where $x_{\mathrm{ref}}$ is a reference value of the predictor, and $\hat{c}_i$, $i = 1, 2, \ldots, n$ are 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):

$$\widehat{\mathrm{MPC}}(x, x_{\mathrm{ref}}) = \exp\left\{ \sum_{i=1}^{n} \hat{c}_i [S_i(x) - S_i(x_{\mathrm{ref}})] - 1 \right\} \times 100\%. \tag{5}$$

The asymptotic distribution of $\widehat{\mathrm{MD}}(x, x_{\mathrm{ref}})$ is

$$N(\mathrm{MD}(x, x_{\mathrm{ref}}),\ \sigma^2_{\mathrm{MD}})$$

and for $\log \widehat{\mathrm{RR}}(x, x_{\mathrm{ref}})$, it is

$$N(\log \mathrm{RR}(x, x_{\mathrm{ref}}),\ \sigma^2_{\log \mathrm{RR}})$$

where

$$\sigma^2_{\mathrm{MD}} = \mathrm{Var}(\widehat{\mathrm{MD}}(x, x_{\mathrm{ref}})) = [S(x) - S(x_{\mathrm{ref}})]' \mathrm{Cov}(\hat{C})[S(x) - S(x_{\mathrm{ref}})], \tag{6}$$

$$\sigma^2_{\log \mathrm{RR}} = \mathrm{Var}(\log \widehat{\mathrm{RR}}(x, x_{\mathrm{ref}})) = [S(x) - S(x_{\mathrm{ref}})]' \mathrm{Cov}(\hat{C})[S(x) - S(x_{\mathrm{ref}})], \tag{7}$$