Paper 5621-2016

Restricted Cubic Spline Regression: A Brief Introduction

Ruth Croxford, Institute for Clinical Evaluative Sciences

ABSTRACT

Sometimes, the relationship between an outcome (dependent) variable and the explanatory (independent) variable(s) is not linear. Restricted cubic splines are a way of testing the hypothesis that the relationship is not linear or summarizing a relationship that is too non-linear to be usefully summarized by a linear relationship. Restricted cubic splines are just a transformation of an independent variable. Thus, they can be used not only in ordinary least squares regression, but also in logistic regression, survival analysis, and so on. The range of values of the independent variable is split up, with "knots" defining the end of one segment and the start of the next. Separate curves are fit to each segment. Overall, the splines are defined so that the resulting fitted curve is smooth and continuous. This presentation describes when splines might be used, how the splines are defined, choice of knots, and interpretation of the regression results.

INTRODUCTION

We use regression analyses to learn about the relationship between a set of predictors and an outcome. Drawing valid conclusions requires properly adjusting for predictors. Many of the predictors we use are continuous variables, and it can be a challenge to model their relationship to the outcome, balancing the competing goals of model simplicity and goodness of fit. Restricted cubic splines, which are a transformation of a continuous predictor, provide a simple way to create, test, and model non-linear relationships in regression models.

This paper defines restricted cubic splines, and describes how they are used in regression analyses. The paper concludes with a summary of the benefits of this useful method.

OPTIONS FOR MODELING CONTINUOUS VARIABLES IN A REGRESSION MODEL

Including a continuous variable as a predictor in a regression model is not straightforward. Table 1 outlines some of the choices for modeling a continuous predictor. The simplest way to include a continuous variable in a regression model is to assume that it has a linear relationship with the outcome. While this assumption is often a reasonable one, if it is wrong, the model will be misspecified: an influential variable may be overlooked or the assumption of linearity may produce a model that fails in important ways to represent the relationship.

A common approach used to investigate and model non-linearity is to divide the continuous variable into categories (for example, age is often collapsed by 10-year age intervals). While categorization produces models that are attractive from the point of view of decision making (e.g., people 60 years of age or older should get the shingles vaccine), the categories may not do a good job of describing the data. When categories are used, the effect of the predictor is forced to be flat within each category. The use of cut points does not allow a smooth relationship between predictor and outcome – rather, there is a step function at the end of each category. As the number of categories increase, so do the number of degrees of freedom, which is a concern for small datasets. And the use of categories reduces the predictive power of the variable when the model is used for prediction.

Rather than splitting the variable into categories, an alternative is to include the predictor as a continuous variable, finding a transformation that produces a linear relationship. Another alternative is to use a quadratic or cubic polynomial to model the relationship (i.e., adding the square and possibly the cube of the variable to the model). Polynomials have the advantage of producing a smooth fit. However, a drawback is that the curves are not flexible. For example, a quadratic curve must be 'U' (or inverted 'U') in shape – what goes up must come down at the same rate, even if the data don't support the downturn. Polynomials often result in a curve which doesn't fit the data well at the ends.

Lastly, two classes of flexible functions are available: splines and fractional polynomials (which will not be discussed here).

Procedure	Characteristics	Recommendation
Dichotomization	Simple, easy interpretation	Bad idea
More categories	Categories capture prognostic information better, but are not smooth; sensitive to choice of cut-points and hence instable	Primarily for illustration, comparison with published evidence
Linear	Simple	Often reasonable as a start
Transformations	Log, square root, inverse, exponent, etc.	May provide robust summaries of non-linearity
Restricted cubic splines	Flexible functions with robust behaviour at the tails of predictor distributions	Flexible descriptions of non-linearity
Fractional polynomials	Flexible combinations of polynomials; behaviour in the tails may be unstable	Flexible descriptions of non-linearity

Table 1. Options for dealing with continuous predictors in prediction models (adapted from Steyerberg (2009) Clinical Prediction Models)

RESTRICTED CUBIC SPLINES

WHAT IS A SPLINE?

Figure 1 is a photograph of a draftman's spline. Originally used to design boats, the spline consists of a thin flexible strip, held in place with lead weights called ducks (because of their shape, which supposedly resembles a duck). The ducks are used to force the spline to pass through specific points, called knots. Between those points, the flexible material will take the smoothest possible shape. The curves between each set of knots join smoothly at the knot. A delightful photograph of an engineer using a mechanical spline can be seen at http://pages.cs.wisc.edu/~deboor/draftspline.html .

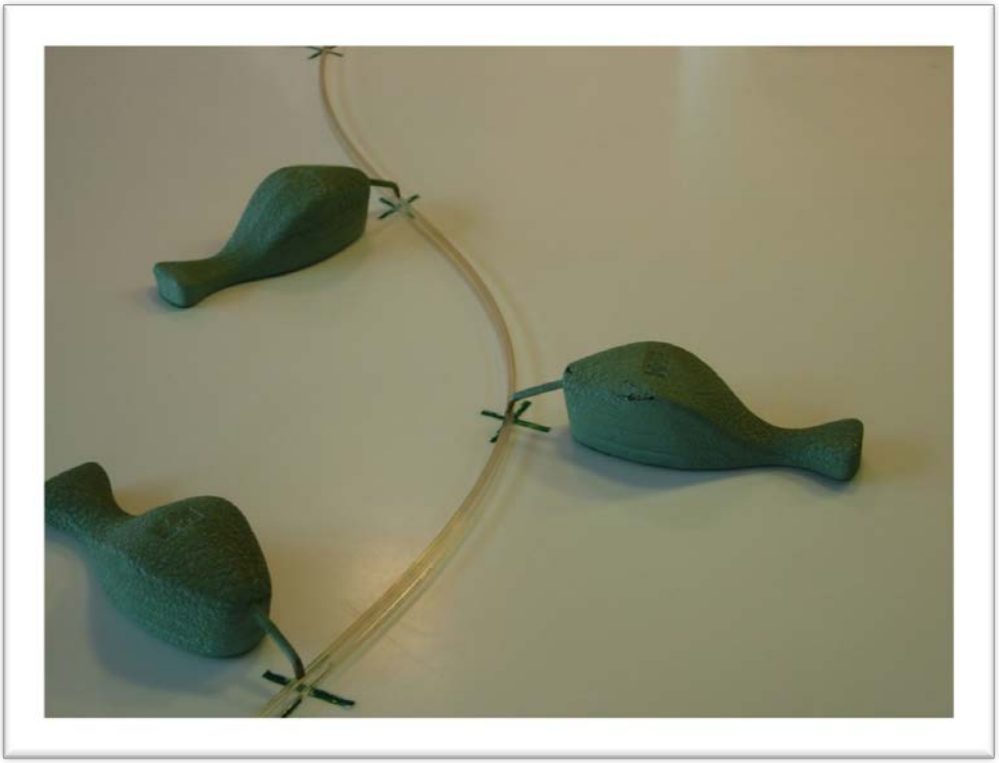

Figure 1. A draftsman's spline, with ducks. Used with permission.

The use of splines for regression is analogous to the spline pictured in Figure 1. The range of values of the predictor is subdivided using a set of knots. Separate regression lines or curves are fit between the knots. Two choices must be made. One is the number and position of the knots. The other is the degree of polynomial to be used between the knots (a straight line is a polynomial of degree 1). Additionally, as suggested by the draftman's spline, it is not enough to fit individual polynomials in between the knots. We require that the individual curves be defined in such a way that they meet at the knots, and in fact that they join "smoothly". A spline function is, therefore, a set of smoothly joined piecewise polynomials. "Smoothly joined" means that for polynomials of degree *n*, both the spline function and its first *n-1* derivatives are continuous at the knots.

If *k* knots are used, fitting a polynomial of degree *n* requires *k + n + 1* regression parameters (including the intercept). In practice, cubic splines are usually used (i.e., a polynomial of degree 3), requiring *k + 3* coefficients in addition to the intercept, compared to only 1 coefficient to fit a linear model. This is the smallest degree of polynomial that allows an inflection, providing sufficient flexibility for fitting data, while not requiring as many degrees of freedom as higher order splines. Cubic splines are smooth in appearance due to the fact that both the first and second derivatives (the slope and the rate of change in the slope) are continuous at the knots.

Cubic splines tend to be poorly behaved at the two tails (before the first knot and after the last knot). To avoid this, restricted cubic splines are used. A restricted cubic spline is a cubic spline in which the splines are constrained to be linear in the two tails. This generally provides a better fit to the data, and also has the effect of reducing the degrees of freedom. Using a restricted cubic spline in a regression analysis will use *k - 1* degrees of freedom (the linear variable *x* and *k – 2* piecewise cubic variables) in addition to the intercept.

SPECIFYING A RESTRICTED CUBIC SPLINE

Specification of a spline uses notation which may be unfamiliar:

Let $u_+ = u$ if $u > 0$

$u_+ = 0$ if $u \leq 0$

If the k knots are placed at $t_1 < t_2 < \ldots < t_k$, then for a continuous variable *x*, a set of (k – 2) new variables are created:

$$x_i = (x - t_i)_+^3 - (x - t_{k-1})_+^3 \frac{t_k - t_i}{t_k - t_{k-1}} + (x - t_k)_+^3 \frac{t_{k-1} - t_i}{t_k - t_{k-1}}, i = 1, .., k-2$$

Thus, the original continuous predictor has been augmented by introducing a set of new variables, each of which is linear in the regression coefficients. The model can therefore be fit using the usual regression procedures, and inferences can be drawn as usual. In particular, we can test for non-linearity by comparing the log-likelihood of the model containing the new variables with the log-likelihood of a model containing *x* as a linear variable.

NUMBER AND LOCATION OF KNOTS

The number of knots is more important than their location. Stone (1986) showed that five knots are enough to provide a good fit to any patterns that are likely to arise in practice. Harrell (2001) states that "for many datasets, *k* = 4 offers an adequate fit of the model and is a good compromise between flexibility and loss of pocesion caused by overfitting a small sample". If the sample size is small, three knots should be used in order to have enough observations in between the knots to be able to fit each polynomial. If the sample size is large and if there is reason to believe that the relationship being studied changes quickly, more than five knots can be used.

For some studies, theory may suggest the location of the knots. More commonly, the location of the knots are prespecified based on the quantiles of the continuous variable. This ensures that there are enough observations in each interval to estimate the cubic polynomial. Table 2 shows suggested locations. (Harrell, 2001).

Number of knots K	Knot locations expressed in quantiles of the *x* variable						
3	0.1	0.5	0.9				
4	0.05	0.35	0.65	0.95			
5	0.05	0.275	0.5	0.725	0.95		
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.5	0.6583	0.8167	0.975

Table 2. Location of knots. From Harrell (2001), Regression Modeling Strategies.

INCLUDING THE SPLINE VARIABLES IN A REGRESSION

The new variables were calculated using only the continuous predictor *x* and the values of the knots. Since they are simply a restatement of the predictor, restricted cubic splines can be used in any type of regression (ordinary least squares, logistic, survival).

EXAMPLES OF THE USE OF RESTRICTED CUBIC SPLINES

Two studies conducted at the Institute for Clinical Evaluative Sciences nicely illustrate two uses of restricted cubic splines. Kendzerka et al. (2014) used restricted cubic spline transformations to test the assumption of a linear relationship between continuous predictors and the risk of hospitalizations related to cardiovascular events in people tested for sleep apnea. When evidence of a non-linear relationship was found, splines were used to model the effects of those predictors.

In another study, the focus was on the shape of the relationship between predictor and outcome, as revealed by the cubic splines. Ravi et al. (2014) used restricted cubic splines to discover the shape of the relationship between surgeon experience (measured using the surgeon's procedure volume in the preceding year) and surgical outcomes. This is a very instructive analysis, because a common approach used to model the effect of surgeon experience on outcomes is to categorize experience by quintiles of volume. As has been pointed out (see, for example, Fang, Austin, & Tu, 2009), categorizing a continuous variable risks missing important relationships).

If you need to include source code, introduce it with a sentence that ends with a colon:

```
proc ds2;
data _null_;
  method init();
    dcl varchar(16) str;
    str = 'Hello World!';
    put str;
  end;
enddata;
run;
quit;
```

CALCULATING RESTRICTED CUBIC SPLINES USING A SAS MACRO

A number of people have written SAS® macros that calculate the variables needed to perform a restricted cubic spline analysis, or that perform the analysis for a specific type of regression. As an example, a number of macros written by Frank Harrell are available on the website of the Department of Biostatistics at Vanderbilt University (http://biostat.mc.vanderbilt.edu/wiki/Main/SasMacros).

CONCLUSION

Restricted cubic splines provide a useful tool for the analysis of the effect of a continuous predictor on an outcome. They allow for great flexibility in the form of the relationship between predictor and outcome.

Their use does not depend on the form of the outcome variable - they can be used in multiple linear regression, logistic regression and survival analysis – and all of the techniques used to draw inferences about parameter estimates can be applied to the cubic spline polynomials.

Inclusion of a restricted cubic spline provides a way to formally test the assumption of a linear relationship between a predictor and the outcome using standard methods. Failure to identify nonlinearity and include it in a model can result in an overestimated or underestimated relationship, or a relationship that is missed altogether. When non-linear relationships exist, splines allow it to be modelled well, reducing model misspecification and providing insight into the relationship between predictor and outcome.

REFERENCES

Fang, J., Austin, P. C., & Tu, J. V. (2009), "Test for linearity between continuous confounder and binary outcome first, run a multivariate regression analysis second", published in the *Proceedings of the SAS® Global Forum 2009 Conference*, Cary, NC: SAS Institute Inc., paper 252-2009. Available at http://support.sas.com/resources/papers/proceedings09/252-2009.pdf

Harrell, F.E. (2010) Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis. Springer-Verlag New York, Inc. New York, USA.

Kendzerska, T., Gershon A. S., Hawker, G., Leung, R. S., & Tomlinson, G. (2014) Obstructive sleep apnea and risk of cardiovascular events and all-cause mortality: A decade-long historical cohort study. *PLOS Medicine, 11(2)*, e1001599. doi:10.1371/journal.pmed.1001599. Available at http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001599.

Ravi, B., Jenkinson, R., Austin, P. C., Croxford, R., Wasserstein, D., Escott, B., Paterson, J. M., Kreder, H., & Hawker, G. A. (2014) Relation between surgeon volume and risk of complications after total hip arthroplasty: propensity score matched cohort study. 2014. *BMJ*, 348:g3284. Available at http://www.bmj.com/content/348/bmj.g3284.

Steyerberg, W.W. Clinical Prediction Models. Springer-Verlag, 2009

Stone, C. J. (1986). [Generalized Additive Models: Comment]. *Statistical Science, 1(3)*, 312-314.

CONTACT INFORMATION <HEADING 1>

Your comments and questions are valued and encouraged. Contact the author at:

Ruth Croxford
Institute for Clinical Evaluative Sciences, Toronto, Ontario
ruth.croxford@ices.on.ca
www.ices.on.ca