Spline Regression in Clinical Research

ZD Mulla

ABSTRACT

Objective: *The objective of this article is to illustrate the statistical technique of spline regression, an under-utilized tool in clinical research. Spline regression was used to assess the dose-response association between serum albumin and hospital mortality.*

Methods: *Data from a previous study of patients hospitalized throughout Florida, United States of America (USA), for invasive group A streptococcal disease were accessed. For the current analysis, serum albumin (SA) at the time of admission was the risk factor of interest. The outcome was unadjusted hospital mortality among 117 patients. First, a traditional, suboptimal approach was employed by breaking SA into three groups and calculating the crude hospital mortality rate in each SA category. The second approach involved the creation of a curve using a quadratic spline model.*

Results: *The traditional approach yielded only three points of information: the hospital mortality rate for the three SA groups. Among patients whose SA upon admission was < 2.5 g/100 mL, 2.5 to 3.4, and 3.5 or greater, the hospital mortality rate was 40.7%, 14.8%, and 8.3%, respectively. The spline model, however, resulted in a smooth curve which was more clinically plausible.*

Conclusion: *The goal of this paper is to expose clinicians to splines. Spline regression, unlike categorical analysis, does not impose the unrealistic assumption of a homogenous risk within categories. Another disadvantage of categorical analyses is that they allow large changes in risk between categories.*

La Regresión por Spline en la Investigación Clínica

ZD Mulla

RESUMEN

Objetivo. *El objetivo de este artículo es ilustrar la técnica estadística de regresión por spline, una herramienta subutilizada en la investigación clínica. La regresión por spline fue usada para evaluar la asociación dosis-respuesta entre la albúmina de suero y la mortalidad hospitalaria.*

Métodos. *Se accedió a datos de un estudio anterior de pacientes hospitalizados a lo largo de la Florida, Estados Unidos de América, en relación con la enfermedad estreptocócica invasiva por estreptococos del grupo A. Para el presente análisis, la albúmina de suero (AS) a la hora del ingreso, fue el factor de riesgo de interés. El resultado fue una mortalidad hospitalaria no ajustada entre 117 pacientes. Primeramente, se empleó un enfoque sub-óptimo, tradicional, dividiendo el AS en tres grupos y calculando la tasa cruda de mortalidad hospitalaria en cada categoría de SA. El segundo enfoque implicó la creación de una curva, usando un modelo de spline cuadrático.*

Resultados. *El enfoque tradicional arrojó sólo tres puntos de información: la tasa de mortalidad hospitalaria para los tres grupos de AS. Entre pacientes cuya AS al momento del ingreso fue < 2.5 g/100 mL, 2.5 á 3.4, y 3.5 o mayor, la tasa de mortalidad hospitalaria fue 40.7%, 14.8%, y 8.3%, respectivamente. Sin embargo, el modelo spline trajo por resultado un curva llana, clínicamente mís plausible.*

From: Department of Obstetrics and Gynecology, Texas Tech University Health Sciences Center, School of Medicine at El Paso, El Paso, Texas, USA.

Correspondence: Dr ZD Mulla, Department of Obstetrics and Gynecology, Texas Tech University Health Sciences Center, School of Medicine at El Paso, 4800 Alberta Ave, El Paso, Texas 79905, USA. e-mail: zuber.mulla@ttuhsc.edu.

Conclusión. *El fin de este trabajo es exponer a los clínicos al uso de splines. La regresión por spline, a diferencia del análisis categórico, no impone el postulado poco realista de un riesgo homogéneo dentro de las categorías. Otra desventaja de los análisis categóricos es que permiten grandes cambios de riesgo entre categorías.*

West Indian Med J 2007; 56 (1): 78

INTRODUCTION

Dichotomous (binary) outcomes are common in clinical research. Clinical investigators, for example, may be interested in determining if there is a dose-response association between a continuous risk factor, such as body mass index (BMI) and pre-eclampsia, a dichotomous outcome (1). The patient either developed pre-eclampsia or she did not.

A traditional approach to dose-response analysis has been to convert the continuous independent variable (BMI in the example above) into a categorical variable and then examine the risk of the outcome by category. To clarify, one would break BMI into quartiles or quintiles *etc*, or preferably, one would create categories that are clinically meaningful (2) and then calculate the proportion who developed the outcome of interest in each group and plot the results. Furthermore, a multivariate analysis could be performed in this situation. If BMI was broken down into four categories, then three indicator (dummy) variables could be entered into a regression model along with confounders.

Disadvantages of the traditional categorical approach include the assumption of a homogenous risk within the categories and the fact that this technique allows for large changes in risk between categories (2). These points will be illustrated using an example by Greenland (3). Assume the risk factor of interest is daily intake of ascorbic acid and the outcome is mortality risk. The researcher chooses the following boundaries for the categories: 20, 50 and 100 milligrams of ascorbic acid intake per day. The categorical dose-response model then dictates that there is no difference in the risk of mortality between 0 and 20 mg per day but then permits an arbitrarily large jump in the risk between 20 and 21 mg per day. According to Greenland [3], "This is biologically absurd, given that 0 mg per day represents a relatively rapidly fatal deficiency state, 20 mg per day does not, and the difference between 20 and 21 mg per day is biologically trivial." Dividing the risk factor into categories is acceptable if the sample size is large enough to permit the creation of very narrow groups (3).

In this article, I provide an introduction to spline regression and give an example using data from a study of patients who were hospitalized in Florida, USA, for invasive group A streptococcal disease (IGASD). The dose-response association between serum albumin and the risk of hospital mortality in this case series is evaluated using both the traditional categorical approach and the more desirable spline model.

METHODS

Data from a previous study of the clinical epidemiology of IGASD were utilized to illustrate spline regression (4). Briefly, the original study was an analysis of patients hospitalized throughout the state of Florida, USA, for IGASD between August of 1996 and August of 2000 and reported to the Florida Department of Health. Secondary analysis of this database was approved by the Institutional Review Board of the Texas Tech University Health Sciences Center, School of Medicine at El Paso.

For the current study, the risk factor of interest was serum albumin at the time of admission or shortly thereafter. The dichotomous outcome was the crude (unadjusted) hospital mortality. A total of 117 patient records were available for the dose-response evaluation. Serum albumin in this case series ranged from 1.1 to 5.1 g/100 mL.

The first step is to categorize the serum albumin variable, as one would do in a categorical analysis (5). The boundaries between the categories are called knots or join points (5). Knots at 2.5 and 3.5 g/100 mL were chosen because of their clinical significance. A serum albumin value of 2.5 g/100 mL or less is associated with oedema while 3.5 g/100 mL is the lower limit of normal.

There are various types of splines including linear, quadratic and cubic. Quadratic splines are smoother than linear splines and are more easily interpreted than cubic splines and therefore are more popular in epidemiology (3). This paper will focus on the quadratic spline. Quadratic splines can suffer from odd behaviour in the tails (5). This instability may be attenuated by restricting either the lower tail, upper tail or both tails to a line segment rather than a curve (2, 3, 5). The spline model that was developed for the current analysis was a quadratic spline with the upper tail restricted to a line segment.

The next several steps encompass the creation of several variables and the execution of a regression model, usually a logistic regression model. The reader is referred to Rothman (5) for these simple formulae. Any software package that can perform logistic regression may be used. The investigator then chooses a range of values that will be displayed in the graph. For this analysis, values of serum albumin ranging from 1.0 to 5.0 g/100 mL in increments of 0.1 g/100 mL were used. The SAS System for Windows 9.1.3 (SAS Institute, Inc, Cary, North Carolina, USA) was used to perform logistic regression. The coefficients from the logistic regression model were used to calculate predicted proba-

bilities of hospital mortality. Finally the predicted probabilities were plotted against the desired range of serum albumin using Microsoft Office Excel. Confounders can be accommodated and the reader is directed to reference 5 for a discussion on this topic. The confounders should be centred before they are included in the analysis. This will allow the investigator to use the plotting method described here without any modification. For example, maternal age at first birth may be a potential confounder in a particular analysis. This confounder cannot have a value of zero and therefore centering is indicated. The technique of centering would require the investigator to simply subtract a frequently observed value of the variable, such as 25 years of age, from the actual value for each subject. The new variable (age at first birth minus 25 years) rather than the original confounder would be entered into the regression model.

RESULTS

Figure 1 shows the suboptimal approach to modelling the association between serum albumin and the unadjusted risk

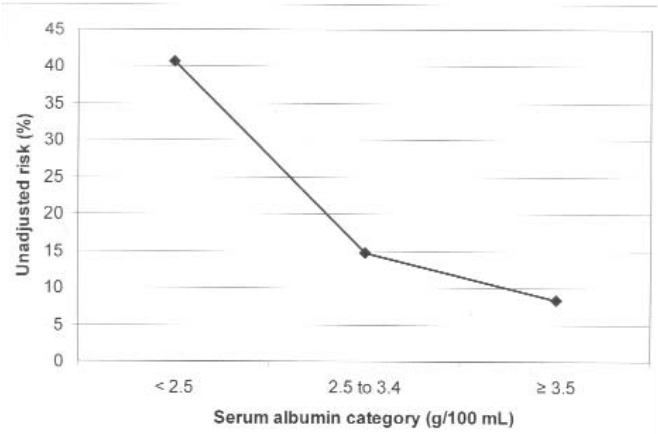

Fig. 1: Association between serum albumin and risk of hospital mortality (n = 117).

of hospital mortality. The hospital mortality in patients whose serum albumin was < 2.5 g/100 mL was 40.7%. This risk decreased to 14.8% in the next category (the subnormal group) and was 8.3% in patients with a serum albumin of $ 3.5 g/100 mL.

Figure 2 displays the spline function summarizing the association between serum albumin and the crude risk of hospital mortality. This curve has a smoother appearance and one that is more biologically plausible compared to Figure 1.

DISCUSSION

The aim of this paper is to provide a simple introduction to spline regression so that clinicians may become savvy consumers of the medical literature. Becoming aware of this modelling technique will facilitate discussion between clinician investigators and members of their research team, such as an

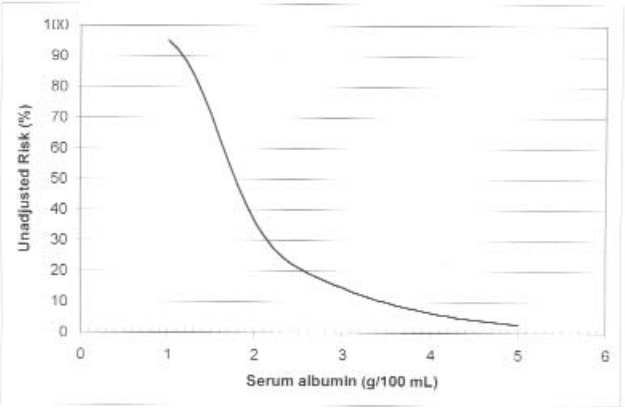

Fig. 2: Quadratic spline model for the association between serum albumin and risk of hospital mortality (n = 117).

epidemiologist or biostatistician, who can assist in the implementation of this under-utilized method.

Three final methodological points need to be addressed. The first issue concerns sample size. The technique described here is advisable when the variable of interest is continuous and the sample size does not permit the use of narrow categories (3); however, logistic regression is a large-sample method. Therefore Greenland recommends that for each category of the exposure variable (serum albumin in the example presented here) one should have at least five subjects with the outcome of interest and five subjects who did not have the outcome of interest (3). This rule of thumb only applies if the investigator's model does contain any product terms. Second, as with all regression methods, one should implement regression diagnostics (model checking) such as tests of fit before using a particular model (3). Third, confidence bands to accompany the spline curve may be generated (3).

In conclusion, spline models represent an improvement in the assessment of dose-response over most categorical analyses since they incorporate information on the variation of the risk of the outcome within the categories of interest (2). Finally, splines can yield a good approximation to more complicated methods such as nonparametric regression and penalized splines (2).

REFERENCES

1. Bodnar LM, Ness RB, Markovic N, Roberts JM. The risk of preeclampsia rises with increasing prepregnancy body mass index. Ann Epidemiol 2005; **15:** 475–82.
2. Witte JS, Greenland S. A nested approach to evaluating dose-response and trend. Ann Epidemiol 1997; **7:** 188–93.
3. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 1995; **6:** 356–65.
4. Mulla ZD, Leaverton PE, Wiersma ST. Invasive group A streptococcal infections in Florida. South Med J 2003; **96:** 968–73.
5. Rothman KJ, Greenland S. Regression splines. In: Modern Epidemiology, 2nd edition. Philadelphia: Lippincott Williams & Wilkins; 1998: 392–4.