Using P-splines to extrapolate two-dimensional Poisson data

Iain Currie[1], Maria Durbán[2] and Paul Eilers[3]

[1] Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland
Email: I.D.Currie@ma.hw.ac.uk
[2] Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Madrid, Spain
[3] Department of Medical Statistics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands

Abstract: Eilers & Marx (1996) used P-splines to smooth one-dimensional count data with Poisson errors. In this paper we consider the extrapolation problem and show that P-splines are well suited to extrapolating in both one and two dimensions. The role of the order of the penalty is highlighted. We illustrate our remarks with the analysis of a large set of mortality data indexed by age of death and year of death.

Keywords: Mortality; P-splines; extrapolation; smoothing; two-dimensions.

1 Introduction

The method of P-splines (Eilers and Marx, 1996) is now well established as a method of smoothing in generalized linear models (GLMs). A succinct summary of the method is: (a) use B-splines as the basis for the regression, and (b) modify the log-likelihood by a difference penalty on the regression coefficients. Wand (2003) gives a most useful overview which highlights the wide class of models that can be fitted with the P-spline approach.
Durban, et al. (2002) introduced a two-dimensional P-spline model for Poisson data in which the regression matrix was defined in terms of the Kronecker product of the regression matrices of two one-dimensional P-spline models. The present paper shows that P-splines provide a natural method of extrapolating the fitted mortality rates forward in time. The role of the order of the penalty is shown to be of particular importance. We illustrate our remarks with the analysis of the same set of mortality data as our 2002 paper.

2 Description of the data

The failure to predict accurately the fall in UK mortality rates from the 1970s to date has had far-reaching consequences for the pensions and annuity business of the UK insurance industry. The Continuous Mortality Investigation Bureau (CMIB) has responsibility for monitoring and predicting mortality rates. In this paper we consider one of the CMIB data sets, namely that for male assured lives. For each calendar year (1947 to 1999) and each age (11 to 100) we have the number of years lived (the exposure) and the number of policy claims (deaths). We use a Kronecker product P-spline model (Durban, et al., 2002) and a system of prior weights to predict mortality rates for 1975-1999 using the data from 1947-1974. The comparison between the observed rates for 1975-1999 and our predicted rates provides a good test of our method.

3 Extrapolating mortality tables

Our data consists of two matrices, $\boldsymbol{Y}$ and $\boldsymbol{E}$, whose rows are indexed by age (here 11 to 100) and whose columns are indexed by year (here 1947 to 1999). The matrix $\boldsymbol{Y}$ contains the number of claims (deaths) and the matrix $\boldsymbol{E}$ contains the exposures. Thus $\boldsymbol{R} = \log(\boldsymbol{Y}/\boldsymbol{E})$ is the matrix of raw log hazards. Durban, et al. (2002) showed how to smooth $\boldsymbol{R}$ by using a 2-dimensional extension of the P-spline model of Eilers and Marx (1996). The smoothing is achieved by using a penalized generalized linear model (PGLM) for $\boldsymbol{Y}$ with Poisson errors and appropriately defined regression and penalty matrices.

We define the regression matrix in terms of the Kronecker product of two 1-dimensional regression matrices. Let $\boldsymbol{B}_a = \boldsymbol{B}(\boldsymbol{x}_a)$, $n_a \times c_a$, be a regression matrix of B-splines based on the explanatory variable $\boldsymbol{x}_a$; in our example, $\boldsymbol{x}_a' = (11, \ldots, 100)$ so $n_a = 90$ and c_a is typically about 20. Similarly, let $\boldsymbol{B}_y = \boldsymbol{B}(\boldsymbol{x}_y)$, $n_y \times c_y$, be a regression matrix of B-splines based on the explanatory variable $\boldsymbol{x}_y$; in our example, $\boldsymbol{x}_y' = (1947, \ldots, 1999)$ so $n_y = 53$ and c_y is typically about 10. The regression matrix for our 2-dimensional model is the Kronecker product

$$\boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a. \tag{1}$$

This formulation assumes that the vector of observed claim numbers $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{Y})$, (this corresponds to how Splus stores a matrix). Note that $\boldsymbol{B}$ has $n_a n_y$ rows and $c_a c_y$ columns, so is typically 4770 by 200. The model is, at present, a standard GLM: $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ where $\log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{B}\boldsymbol{a}$ and $\log \boldsymbol{e}$, $\boldsymbol{e} = \mathrm{vec}(\boldsymbol{E})$, is the usual offset in a log linear model for mortality data.

This regression model will usually be over-parameterized ($\mathrm{len}\,(\boldsymbol{a}) \approx 200$) so we introduce a penalty on $\boldsymbol{a}$. (Durban, et al., 2002) show that an appropriate penalty matrix is

$$\boldsymbol{P} = \lambda_a \boldsymbol{I}_{c_y} \otimes \boldsymbol{D}_a' \boldsymbol{D}_a + \lambda_y \boldsymbol{D}_y' \boldsymbol{D}_y \otimes \boldsymbol{I}_{c_a} \tag{2}$$

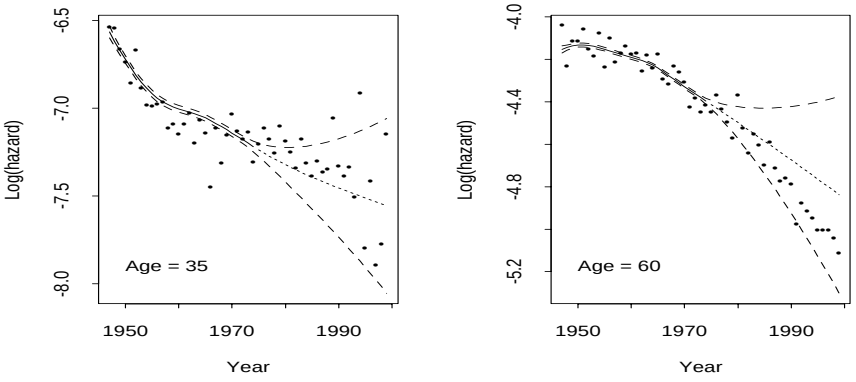

FIGURE 1. Observed, fitted and extrapolated log(hazard) with 95% confidence intervals for $p_a = p_y = 2$. Left panel: age 35, right panel: age 60.

where $\boldsymbol{I}_{c_a}$ is an identity matrix of size c_a and $\boldsymbol{D}_a$ is a difference matrix with dimension $(c_a - p_a) \times c_a$ where p_a is the order of the penalty on age; similar definitions apply for $\boldsymbol{I}_{c_y}$ and $\boldsymbol{D}_y$. For given values of the smoothing parameters λ_a and λ_y the model is fitted by penalized likelihood and the penalized version of the scoring algorithm

$$(\boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})\hat{\boldsymbol{a}} = \boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B}\tilde{\boldsymbol{a}} + \boldsymbol{B}'(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}). \tag{3}$$

Here, $\tilde{\boldsymbol{a}}$, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{W}} = \text{diag}(\tilde{\boldsymbol{\mu}})$, the diagonal matrix of weights, denote current estimates, and $\hat{\boldsymbol{a}}$ denotes the updated estimate of $\boldsymbol{a}$; additionally, $\log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{B}\boldsymbol{a}$, the canonical link. Finally, the smoothing parameters can be selected by optimising with respect to the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), for example. We perform extrapolation with the following simple device: we define a weight matrix $\boldsymbol{V} = \text{blockdiag}\{\boldsymbol{I}, \boldsymbol{0}\}$ where $\boldsymbol{I}$ is an identity matrix of size $n_a n_{y_1}$ and $\boldsymbol{0}$ is a square matrix of 0's of size $n_a(n_y - n_{y_1})$. We have in mind using n_{y_1} years of data as a training set and extrapolating the remaining $n_y - n_{y_1}$ years. Alternatively, we can take $\boldsymbol{I}$ to have size $n_a n_y$ and extrapolate into the future. To accommodate the weight matrix $\boldsymbol{V}$ we modify the scoring algorithm (3) as follows:

$$(\boldsymbol{B}'\boldsymbol{V}\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P})\hat{\boldsymbol{a}} = \boldsymbol{B}'\boldsymbol{V}\tilde{\boldsymbol{W}}\boldsymbol{B}\tilde{\boldsymbol{a}} + \boldsymbol{B}'\boldsymbol{V}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) \tag{4}$$

where any unknown values in $\boldsymbol{y}$ and $\boldsymbol{e}$ can be given arbitrary values.

Example: We illustrate our methodology by using the 1947-1974 data to predict the 1975-1999 rates. Figure 1 shows the fitted and extrapolated

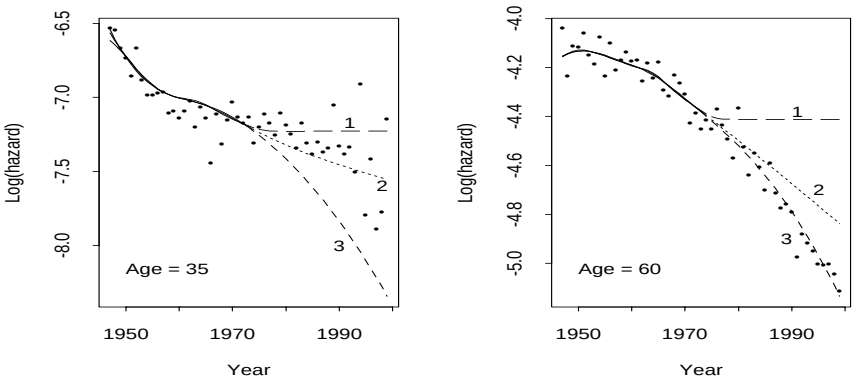

FIGURE 2. Observed, fitted and extrapolated log(hazard) for $p_a = p_y = 1$, 2 and 3 in turn. Left panel: age 35, right panel: age 60.

log(mortality) values for ages 35 and 60. The fit used cubic B-splines and second order difference penalties; the smoothing parameters were chosen using BIC. Confidence intervals are also included and we note that the observed rates for 1975-1999 for both ages are comfortably within their respective 95% confidence funnels.

4 The role of the order of the penalty

In the previous section we used a quadratic penalty, $p_a = p_y = 2$. In this section we examine the conventional wisdom that the order of the penalty has only a small effect on any smoothed values. Figure 2 shows the results of fitting and extrapolating using first order ($p_a = p_y = 1$), second order ($p_a = p_y = 2$) and third order penalties ($p_a = p_y = 3$). We make two comments: first, the order of the penalty has no discernible effect on the smooth of the training data; second, the order of the penalty has a dramatic effect on the extrapolated values. In this paper we have concentrated on the 2-dimensional problem but it is clear from (4) that the method can be applied in 1-dimension. In this case it can be shown that the extrapolation works by extrapolating the regression coefficients and these extrapolations are constant, linear or quadratic depending on the order of the penalty. This result is approximately true in 2-dimensions, as is evident from Figure 2. We make some further comments on this property in our concluding remarks.

5 Conclusions

The failure to predict accurately the fall in mortality rates has had far-reaching consequences for the UK pensions and annuity business. What comfort can be drawn from the results presented in this paper? We compare the predicted mortality rates from 1975-1999 with the observed rates over the same period and draw two main conclusions.

First, the predicted rates are higher than the observed rates for nearly all ages. Visual inspection of the observed rates suggests that it is unlikely that the sharp fall in mortality that occurred from the 1970's to the present could have been predicted back in the 70's.

Second, from 1975 to date, the observed rates lie at about one standard error below the predicted rates and are comfortably within the confidence funnel of the predicted rates. In view of the variation in the mortality rates observed before 1975 this suggests that a prudent course is to allow for this variation by discounting the predicted rates by a certain amount. If this discount had been set at one standard error then the resulting 'prudent' predictions would have been very close to what actually happened. Our view is that some such discounting procedure is the only reasonable way of allowing for the uncertainty in these, or indeed any, predictions.

We also make two general remarks on our method. First, we emphasise the critical role of the order of the penalty, *pord*. The choice of the order of the penalty corresponds to a view of the future pattern of mortality: $pord = 1$, 2 or 3 corresponds respectively to future mortality continuing at a constant level, improving at a constant rate or improving at an accelerating (quadratic) rate. We not only used BIC to choose the values of the smoothing parameters for given value of *pord* we also used BIC to choose the value of *pord*; the preferred value of *pord* was 2 and this was used to produce Figure 1.

Second, in this paper we have been concerned with extrapolation forward in time. However, the method is quite general. In one dimension we can extrapolate both forward and backward while in two dimensions we can extrapolate a rectangular data set in any direction. All that is required are the regression and penalty matrices, and the appropriate weight matrix. The extrapolation is then effected by (4).

References

Durban, M., Currie, I. & Eilers, P. (2002). Using P-splines to smooth two-dimensional Poisson data. *Proc. 17th IWSM.* Chania, Crete. 207-214.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11**, 89-121.

Wand, M. P. (2003). Smoothing and mixed models. *Comput. Stat.*, to appear.