MODELLING MORTALITY BY THE AGE AND YEAR OF DEATH

by

CHARALAMBOUS MARIA

Master in Actuarial Science

Department of Actuarial Mathematics and Statistics

Heriot – Watt University

EDINBURGH

September 2002

To my family

Contents

Acknowledgements

In writing this project I have been greatly helped by the many constructive comments and criticism I received from my supervisor Iain Currie. I have benefited greatly from the numerous hours of discussion with him over the two-month period, his detailed suggestions for improvements and his active interest and support. I would also like to acknowledge the assistance I received from my colleagues Vassiliadis, Gomez-Hernandez, Galaiya and Kamaruddin.

Abstract

Nonparametric regression, or smoothing consists of using the data to guide the selection of a model from a large class of flexible functions. Eilers and Marx (1996) introduced P-splines as a method of smoothing, in which B-splines and a roughness penalty based on differences are combined. We shall see that this class of smoothers are sufficiently flexible for a wide variety of situations including linear modelling and generalized linear modelling. We extend the method to two dimensions and apply the model to analyse a large set of mortality data indexed by age and year of death.

CHAPTER 1

Introduction

In the recent years, statistical analysis has covered considerable ground in the area of nonparametric smoothing. There has been an upsurge of interest and activity which has led inevitably to a specialization of the subject matter. There are a number of reasons for this growth of popularity: parametric modelling is usually not adequate to fully model large data sets; graphical design of the results has become more desirable and easier to use; exploratory analysis is now a more common approach for data analysis.

Many methods for non-parametric modelling have been proposed and studied. Some of the most popular ones are mainly data analytic and they do not make particular use of statistical models. The roughness penalty method on the other hand, amenable to a very complete and elegant mathematical way, provides the connecting link between classical and non-parametric statistics. Being only one of a number of available curve estimation procedures it has the conceptual advantage to allow explicit specification of the

way in which adherence to the data is to be measured providing a unifying approach to a wide range of smoothing problems. It balances fidelity to the data with smoothness of the fitted curve allowing the latter to be incorporated in a natural way to regression analysis. One of the oldest applications of the roughness penalty approach (Whittaker, 1923) is in the smoothing of life tables for actuarial purposes.

Eilers and Marx (1996) provided a very interesting and flexible addition in nonparametric curve fitting based on the idea of the roughness penalty. They proposed P-splines as an effective tool for flexible smoothing which is basically a combination of B-splines and difference penalties. The proposed method has many attractive and useful properties and a short overview of those will be given later.

Our main theme in this project is the applicability of the roughness penalty approach in modelling human mortality. We use the method to smooth and analyse a large set of mortality data, a key component of actuarial work. We follow Eilers and Marx in using P-splines as our smoothing procedure. Making the rational assumption that mortality

rate varies smoothly with age, we graduate crude rates to obtain estimates of the underlying true mortality.

Although our primary emphasis in this project is on smoothing problems we found it helpful to start in chapter 2, by developing the classical regression models which are closely related to our discussion. Before starting a formal and comprehensive development of the idea of roughness penalty we give a brief description of what is meant by a B-spline since it plays a pivotal role in our smoothing methodology. We then extend the ordinary least squares to incorporate B-splines as a basis for our regression and via some numerical examples we demonstrate the idea of P-spline regression.

In chapter 3 we set out the way in which the roughness penalty method can be applied in the broader context of generalized linear models. We start by explaining some essential background of this class of models and provide an example that arises from the actuarial context. We then consider the theoretical development of P-splines as a method of smoothing and illustrate the idea with the analysis of one-dimensional count data with Poisson errors. The illustrations give some flavour of the analysis of the two-dimensional mortality data covered in chapter 4.

Chapters 2 and 3 cover the theoretical background, simplify and clarify the subsequent discussion of the last chapter, where we consider the extension of P-spline smoothing in two-dimension. We see that there is a natural generalization of the method to two-dimensions and that the attractive features of the one-dimension methodology carry over. We illustrate our remarks with the analysis of a large set of mortality data indexed by age and year of death provided by CMIB.

In theory the estimation procedure might seem quite straightforward but in practise it

involves considerable computation. However recent advances in computer software,
the variety and rapid development of computing facilities and the existence of publicly-available software has made the applications of these statistical techniques feasible. The development of computer programs for numerical optimisation of non-linear functions are now included in statistical packages making the numerical aspect of the problem manageable. For the computation of the regression results we used routines available within the statistical

software R. This is a free downloadable package which provides most of the functionalism of the S-Plus program. The two languages have very similar syntax.

CHAPTER 2

Regression using B-splines

2.1 REGRESSION MODELS

Many statistical investigations are concerned with providing models that are needed to predict one or more variables in terms of others. Although it would be desirable to be able to predict the exact lifetime of a person, the outcome of next week's lottery or the sales of a new product this is not of course possible. However, statistics has managed to provide us with models that in many cases enable us to deal effectively with situations involving uncertainties. The expected lifetime of an insured person for example, would be a satisfactory piece of information for an actuary required to set a life insurance premium.

Regression analysis is an approach that may be used for the study of relations between variables. *Linear regression* is one of the oldest and most widely used statistical techniques. Given data points (*X*,*Y*) the natural way to view linear regression is as a method fitting a model of the form

$$Y = a + bX + e \qquad \textbf{(2.1)}$$

to the observed data, where *a*, *b* represent unknown constant parameters, called "regression coefficients", and *e* is the random "error" of the model which is assumed

to have a normal distribution with mean value of zero and unknown variance σ^2.

The first purpose of regression is to provide a summary of the observed data in order to explore and present the relationship between the design variable *X* and the response variable *Y*. In many problems we are dealing with a data set that gives the indication (possibly through a plot of the observed values) of simple linear regression of the form (2.1). It is obvious and natural then to draw a straight line to emphasize the linear trend. Linear regression automates this procedure and ensures comparability and consistency of

results. The other main purpose of regression is to use the model (2.1) for prediction. Given any point *X* an estimate of the expected value of a new observation *Y* at the point *X* is given by $\hat{a}+\hat{b}X$ where $\hat{a}$ and $\hat{b}$ are estimates of a and b, the parameters of the problem. The fitting of the simple linear relation between *Y*s and *X*s requires us therefore to choose the values of the parameters that give the patterned set $\hat{Y}_i$ closest to the data Y_i. To do so requires some measure of discrepancy to be defined between the observed and fitted values. Classical least squares chooses $\sum_i (y_i - \hat{y}_i)^2$ as the measure of discrepancy. Estimates of the parameters are obtained by minimizing the sum of squares of deviations of the observed *Y*, from the assumed true model. The aim is to minimize the sum of squared differences between observed and fitted values.Let us consider a very simple data set, which will be useful in illustrating the above.

EXAMPLE 2.1

The number of weeks that 10 female patients had been on a diet and the corresponding weights losses are displayed in the following table:

Weeks $\Delta(Weight)$

x	y
1	1.5
2	3.3
2	2
3	3.8
4	7

6	8.5
7	13
9	13.2
12	16
14	20

Plotting these data as in Figure 2.1a, we observe that the pattern followed suggests that the regression is approximately linear and that a straight line would provide a good fit. Using the R-program we find that the values of the regression coefficients are: $\hat{a} = 0.276666$, $\hat{b} = 1.4205556$ **and hence we can calculate the fitted values using the fitted regression line** $\hat{y} = \hat{a} + \hat{b}x$**. The fitted line is shown in Figure 2.1b.**

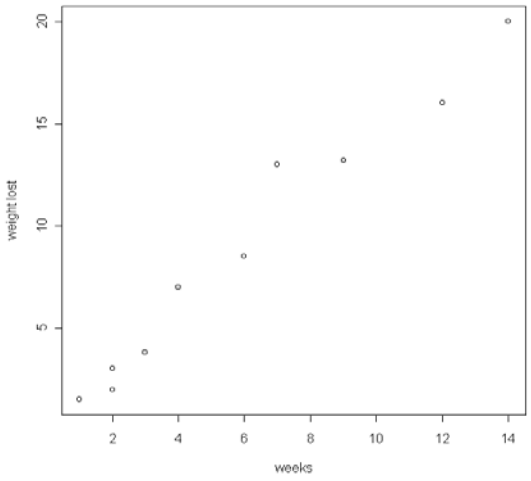

Figure 2.1a: Data on number of weeks being on a diet and observed differences between the weights before and after the diet.

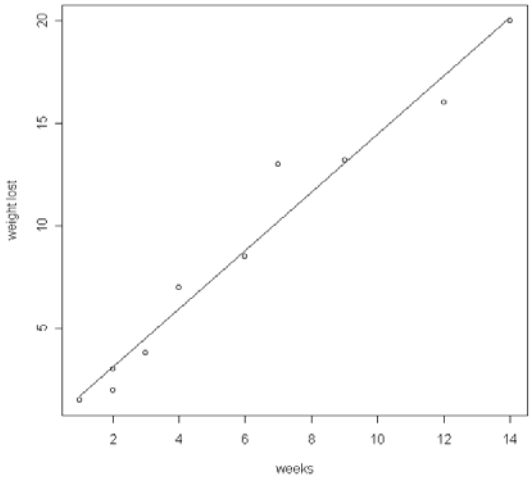

Figure 2.1b: Fitted regression line of example 2.1

Whilst the simple linear model is sufficient to deal with a number of problems of interest, there are many situations in practice where observed responses are influenced simultaneously by several variables. Statistical analysis of the dependence on explanatory variables then usually leads

to the use of *Multiple linear regression*. In general therefore, with a dependent variable *Y* and p independent variables X_1, X_2…,X_P we seek for a relationship of the form:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p + e \qquad \textbf{(2.2)}$$

The same principles of estimation apply with multiple regression as apply with simple linear regression.

It is undesirable to include too many variables in the regression equation since the work of calculation increases rapidly with the number of variables and furthermore a model with many variables it can seldom be easily applied in subsequent prediction.

Another type of regression is the *polynomial regression*. The fitting of polynomial regression equation defined by polynomials of the form

$$Y = b_0 + b_1X^1 + b_2X^2 + ... + b_pX^p + e \qquad \textbf{(2.3)}$$

is in principle no different from the fitting of multiple regression equation as defined previously. The different powers of *X* simply play the role of the different independent variables in the earlier discussion. However, a polynomial is usually fitted in order to smooth out fluctuations in the data caused by random errors and not

because it is thought to represent the actual relationship. It should be pointed out, that the fitting of polynomials of high degree to the observed data is seldom of much value. At first sight it might seem as though a good model is one that fits the data very well. By including enough parameters in our model (high degree of polynomial) we can make the fit as close as we please. In doing so, however we have achieved no reduction in complexity, no simple theoretical pattern of the data. Simplicity is also a desirable feature of a model as it gives better predictions than one that includes unnecessary extra parameters. If a model is made to fit very closely to a particular set of data then it will not be able to encompass the inevitable changes that will be tend to be necessary when another set of data related to the same phenomenon is collected.

2.2 B- SPLINE REGRESION

For situations where a simple polynomial or nonlinear regression model is not suitable, the spline provides a flexible smooth function for a set of observations.

Smoothing splines have been extensively used for regression type problems. They are formed mathematically from piecewise polynomial functions satisfying continuity

properties. There is an enormous literature on splines, most of it concerning their numerical-analytic rather than statistical properties. In this project we are dealing with B-splines, a type of spline that are perhaps the most popular in computer graphic applications. They are attractive as base functions for non-parametric modelling and are constructed from polynomial pieces connected at certain values of x, the knots. Very simple examples of B- splines with degree 1 and

2 respectively are shown in figures 2.2a and 2.2b. The first figure illustrates 4 B-splines of degree one, each one based on 3 knots and consists of two linear pieces, whilst the second figure displays 4 B-splines of degree two consisting three quadratic pieces, joined at two knots. The splines here are based on four adjacent knots and at the connecting points the first derivatives of the polynomial pieces coincide. We note that first degree B-splines overlap with two neighbours whereas second degree B-splines overlap with four neighbours (provided that they are not the leftmost or rightmost splines). We could construct as large set of B-splines as we want by simply introducing more knots.

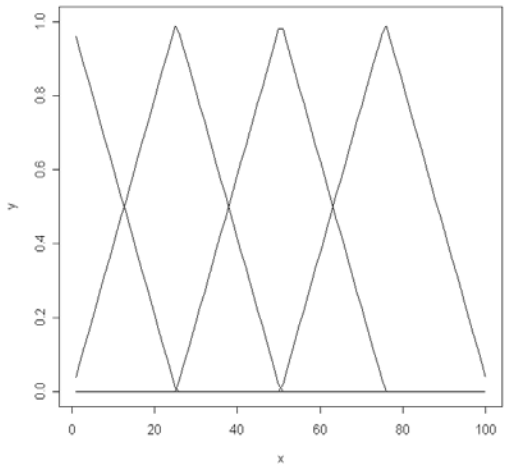

Figure 2.2 a: B-splines of degree 1

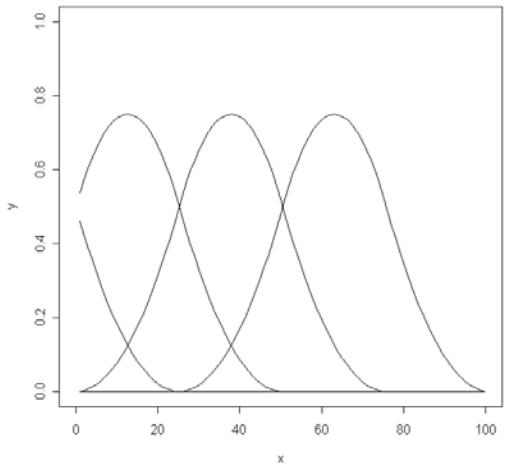

Figure 2.2 b: B-splines of degree 2

To put B-splines in a more formal and general framework we consider the properties of a B-spline of degree q as given by Eilers and Marx (1996):

- it consists of q+1 polynomial pieces, each of degree q
- the polynomial pieces join at q inner knots
- at the joining points, derivatives up to order q-1 are continuous

- the B-spline is positive on a domain spanned by q+2 knots; everywhere else is zero
- except at the boundaries, it overlaps with 2q polynomial pieces of its neighbours
- at a given x, q+1 B-splines are nonzero.

Once we are given the number of knots it is easy to compute B-splines recursively for any desired degree of the polynomial. However, B – splines of third degree are by far the most commonly used in practice. They interpolate the range between the knots by a third order polynomial ensuring that the first and second derivatives are continuous at these points. Cubic B-splines have a fairly pleasing appearance since discontinuities in third and higher order derivatives are not visible.

Many papers and a number of books have appeared discussing how we could cope with the delicate problem of choosing the number and positions of knots. The optimal choice is a complex numerical workload and in this project we adopt what Eilers and Marx proposed (1996) i.e. a relatively large number of knots in an equidistant grid.

2.3 ORDINARY LEAST SQUARES WITH B-SPLINES BASIS

In this section we return to the fundamental problem of computing the minimiser of the residual sum of squares when we are given a data points and we need to condense them by fitting them to a model in the form of a parametric equation.

We consider the case where we are dealing with a simple linear regression of the form (2.1).

The aim is to find the regression coefficients by using the criterion of least squares where we want to minimize the function $S(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ .By differentiating with respect to a and b we get the normal equations of the form:

$$\sum_{i=1}^{n} y_i = n\hat{a} + \hat{b}\sum_{i=1}^{n} x_i \ , \quad \sum_{i=1}^{n} y_i x_i = \hat{a}\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^{\,2}$$

We can write these in matrix form as $\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^{\,2} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$

Let's define the basis matrix X, as $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix}$ and $\Theta = \begin{pmatrix} a \\ b \end{pmatrix}$, $\quad y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix}$

Then $\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^{\,2} \end{pmatrix} = X^T X$, $\begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = X^T y$ and so we get $X^T X\hat{\Theta} = X^T y$.

The estimate of Θ is therefore given by

$$\hat{\boldsymbol{\Theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.4}$$

and the fitted values by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\Theta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{2.5}$$

where H is known as the *hat matrix* since it maps the vector of observed values to their predicted values. It can be proved that the same relationship holds for regression

of the form (2.2) and (2.3) where $X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$ and

$\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_n & x_n^2 \end{pmatrix}$ respectively for p=2.

Note that the trace of the hat matrix gives the number of parameters of the model. For example in the simple linear regression model using the general relationship for matrices, tr(AB)=tr(BA) we get: $tr(H) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_2) = 2$

By simply replacing the basis matrix X with the B-spline matrix it can be shown that the above relationships also apply when we do regression of the form $y = B\Theta + \varepsilon$ using B-spline basis, where Θ is a vector of the regression coefficients. Splines are therefore linear smoothers since the fitted values $\hat{y} = (\hat{y}_1,\hat{y}_2...\hat{y}_n)^T$ can be written as a linear function of the observed values $y=(y_1, y_2, ..., y_n)^T$, that is $\hat{y} = Hy$. The connection with classical regression motivates the calculation of the degrees of freedom that give an indication of the effective number of parameters fitted. Just like the case of ordinary regression the degrees of freedom of the model are given by the trace of the hat matrix and that arises as a natural property of these linear models.

We illustrate the least squares smoothing method in the following example where we choose the basis function to be a set of cubic B-splines on a fixed grid of knots equally spaced to cover the range of the points x.

EXAMPLE 2.2

We generate some data (x,y) as follows: Let x=1,2,…,50, $E(y)=x^3 - 7x + 11$, y=(E(y) – min(E(y))/[max(E.y) – min(E(y)]+rnorm(50,0,0.1). Let ndx denote the number of intervals in the domain of x which when added to the degree of the B-spline basis gives the effective number of parameters fitted (number of B-splines). The B-spline base matrix can be generated with the aid of the R-program using a certain algorithm.

Considering the notation used in the normal equations before and using the B-spline matrix instead of *X,* the estimates of y are given by $\hat{y}=B(B^TB)^{-1}B^Ty$.

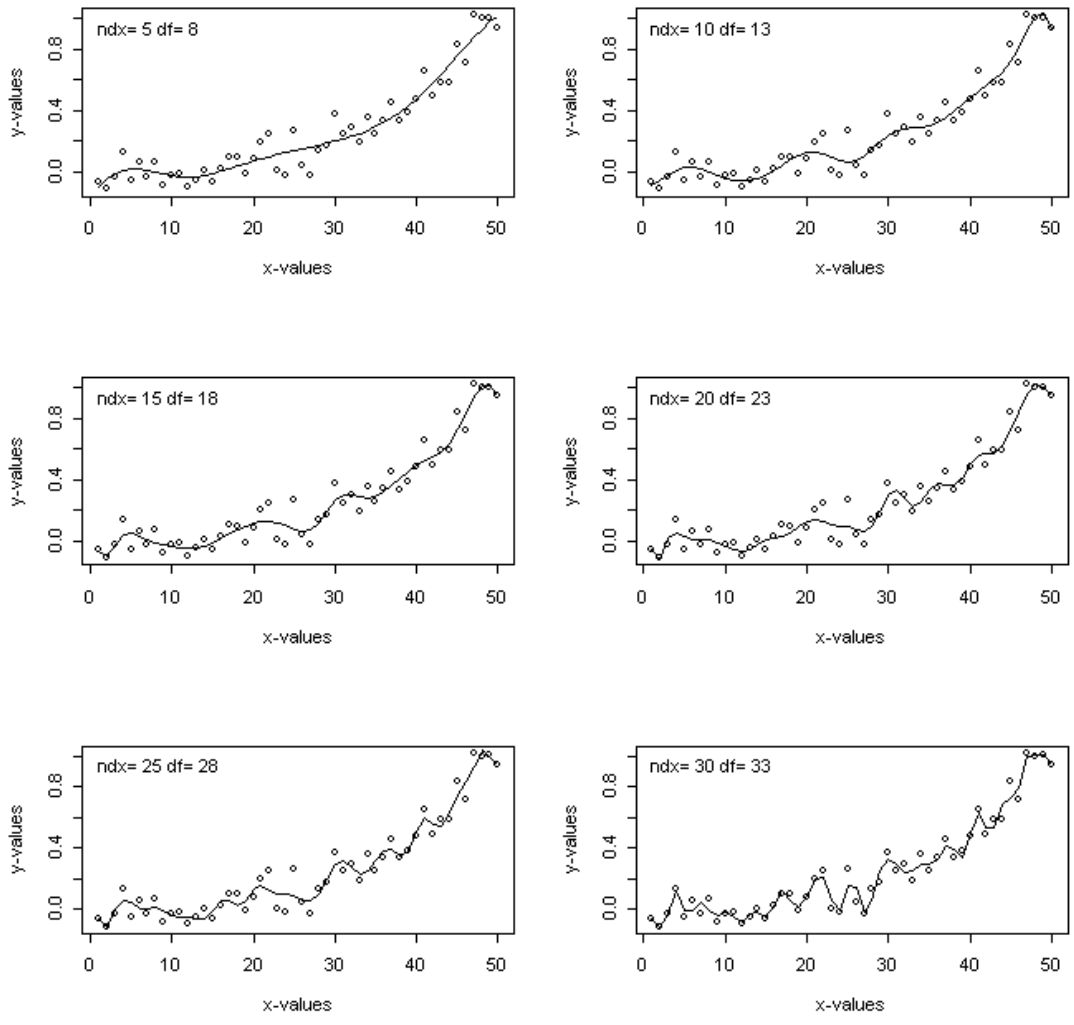

Figure 2.3: **. . .** observed data set (x,y)

___ observed x-values against fitted y-values

We observe that by using large number of intervals in the domain i.e. a large number of B-splines, we obtain a very good fit of the data. However, the model is not that appealing, for such large numbers, if we consider the high complexity of the model associated with that values. The increased number of the degrees of freedom indicates that complexity. A natural need that therefore arises is to cope with the unappealing characteristic of overparameterization which tends to undersmoothing and in practise yield curves that are too "wiggly". We should consider a way that best combines smoothness of the fitted model and closeness to the data so that the choice of ndx would not influence the necessary compromise.

2.4 PENALIZED LEAST SQUARES REGRESSION

The underlying principle in penalized least squares estimation is to estimate the unknown smooth (regression) function by explicitly trading off fidelity to the data with smoothness of the estimate. We obtain this by adding a roughness penalty term to the residual sum of squares. The addition of the roughness penalty ensures that the penalized least squares estimator is determined not only by its goodness-of-fit to the data as quantified by the residual sum of squares but also by its smoothness enforced by the penalty.

Eilers and Marx (1996) defined the penalty function to be based on finite differences of the regression coefficients of adjacent B-splines. The penalized spline estimator is then called a P-spline. This approach reduces the problem of choosing the smoothing or interpolating curve from being infinite-dimensional to finite-dimensional and encourages smoothness by forcing the coefficients to be close.

The difference penalty is easily introduced into the regression equations. The parameters are estimated by minimizing the function:

$S(\Theta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} b_{ij}\Theta_j)^2 + \lambda[(\Theta_1 - \Theta_2)^2 + (\Theta_2 - \Theta_3)^2 + ...(\Theta_{k-1} - \Theta_k)^2]$ which can also be written as

$$S(\Theta) = (y - B\Theta)^T (y - B\Theta) + \lambda \Theta^T D^T D\Theta \qquad (2.6)$$

where λ is the smoothing parameter, its coefficient the roughness penalty and D a matrix of differences with k-1 rows and k columns of the form

$$D = \begin{pmatrix} 1 & -1 & 0 & . & 0 & 0 \\ 0 & 1 & -1 & . & 0 & 0 \\ . & 0 & 1 & . & . & . \\ . & . & . & . & 0 & 0 \\ 0 & 0 & 0 & . & 1 & -1 \end{pmatrix}$$ (k is the number of columns in B depending on the number of knots and the degree of the B-spline).

We shall not for the moment discuss the choice of the smoothing parameter. This will be considered in detail later on. Note that we could also have used quadratic penalty of the form $(\theta_1-2\theta_2-\theta_3)^2+...+(\theta_{\kappa-2}-2\theta_{\kappa-1}+\theta_\kappa)^2$ and with proper modification of the D matrix equation (2.6) would also hold.

Using standard calculus as in 2.3, it can be shown that the penalized estimator is given by

$$\hat{\Theta}(\lambda) = (B^T B + \lambda D^T D)^{-1} B^T y \qquad (2.7)$$

The effective dimension of the P-spline fit is approximately the trace of the hat matrix as proposed by Hastie and Tibshirani (1990).

We applied penalized least squares estimation on the data set of example 2.3. We used cubic B-splines as base functions and a second order penalty on the regression coefficients. We considered several values of the smoothing parameter and kept the number of intervals fixed at 35. The graphical presentation is shown in figure 2.4.

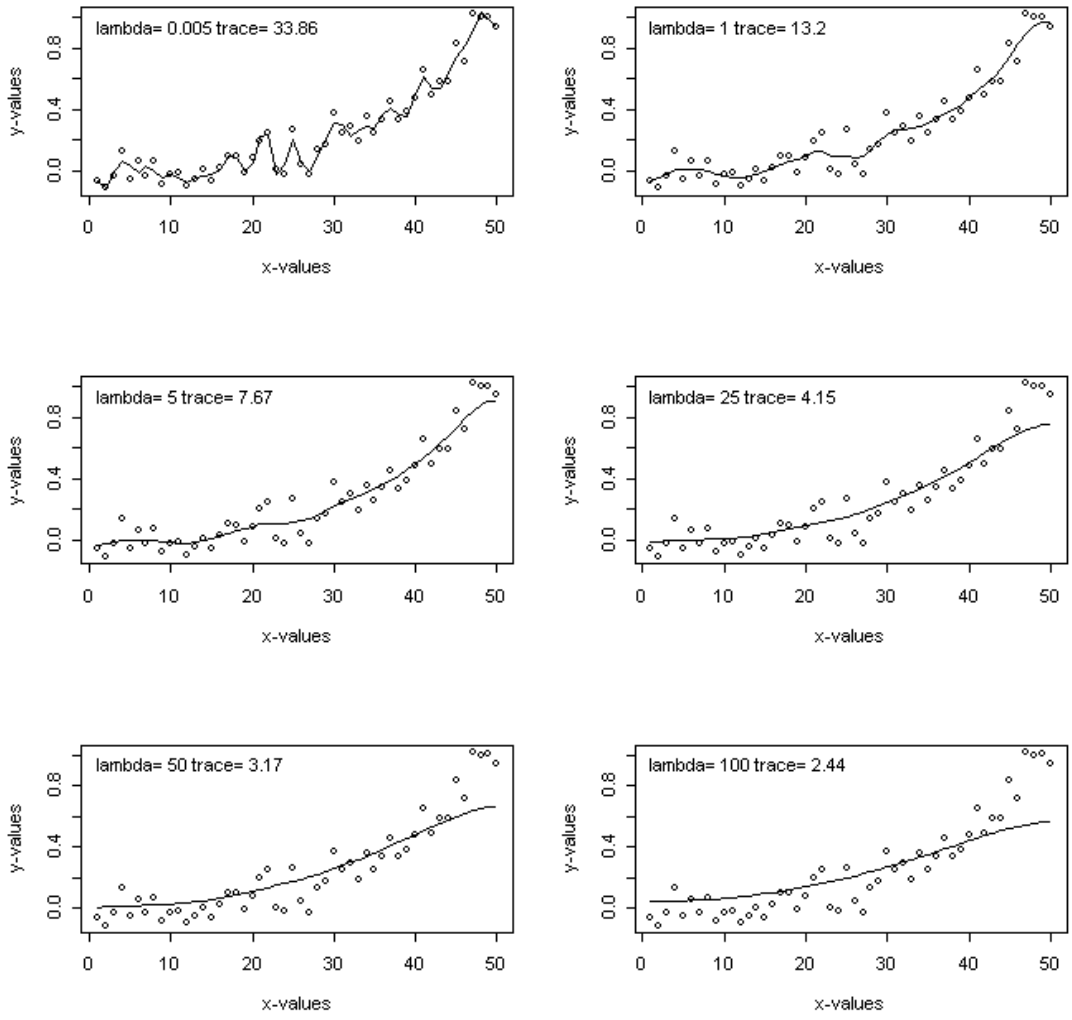

Figure 2.4 : Penalized least squares estimation with cubic B-splines base function and second order roughness penalty.

We note that for small values of λ, the estimated curve tracks the data closely. In the limiting case where $\lambda \to 0$ the fitted curve will approach that shown in figure 2.3 (for ndx=30), yielding a wiggly undersmooth curve. For large values of the smoothing

parameter the estimated curve displays little curvature but fairly poor fit. In the limiting case as $\lambda \to \infty$ the curve will approach the linear least squares regression line. Note also that as lambda increases the trace of the hat matrix i.e the effective dimension of the fitted curve decreases.

The above remarks lead us to the conclusion that the smoothing parameter determines

the strength of the penalty and balances the two criteria of smoothness of the fitted model and closeness of the fit of the model to the data. We need therefore some way to determine the optimal value of it and subsequently the "adequate" level of smoothing.

2.5 OPTIMAL SMOOTHING IN PENALIZED LEAST SQUARES REGRESSION

The problem of choosing the smoothing parameter can be approached by an automatic method whereby its value is chosen by the data. In order to determine an "optimal" level of smoothing one should consider measures for the complexity and the quality of fit of a model. We already mentioned that the trace of the hat matrix could be interpreted as the effective dimension of the fitted curve. For the quality of fit we could consider the residual sum of squares:

$SSE=(y-\hat{y})^T(y-\hat{y})$ $\quad \hat{y}=Hy \Rightarrow SSE= [(I-H)y]^T[(I-H)y] = y^T(I-H)^T(I-H)y = y^T(I-H)^2y$
and dividing by n (the number of observations) we get $SSE/n=\sigma^2$.

Several model selection criteria were evaluated to select the smoothing parameter in penalized regression splines using basically the above arguments and giving quite similar results. One of the most acceptable and successful is the Akaike information criterion (AIC) which requires the value of λ that minimizes $\mathbf{\log \sigma^2 + 2\, tr(H)/n}$.

Spline smoothing estimation with various smoothing parameter selectors is available in various statistical packages including S-PLUS, JMP, SAS/INSIGHT and XploRe.

With the aid of the S-PLUS program we applied the AIC for the data set of example 2.2. Using cubic B-splines as our basis for the regression and a second order penalty on the regression coefficients we obtained the values of lambda that minimizes AIC for different values of ndx. The results are displayed in the following table and the plots of the fitted curves are shown in figure 3.5

NDX	**Lambda**	**Trace**	**AIC**
5	0.59	3.98	-4.545460
10	0.63	4.1	-4.534544
15	22.46	4.14	-4.532176
20	54.52	4.16	-4.531066
25	100	4.23	-4.530431
30	187.05	4.28	-4.530257

Note how the optimal value of lambda increases as ndx increases. Moreover the dimensionality of the problem is remarkably reduced. Note the corresponding values of the trace of the hat matrix when we used ordinary least squares for our estimates for example 2.2. The effective number of parameters fitted would increase as the number of intervals in the domain increases reaching the value of 33 for ndx=30. With the introduction of penalties we manage to keep that value close to 4. Furthermore all ndx give pretty much the same fitted values and that solves the problem of choosing the optimal value of B-splines for our regression.

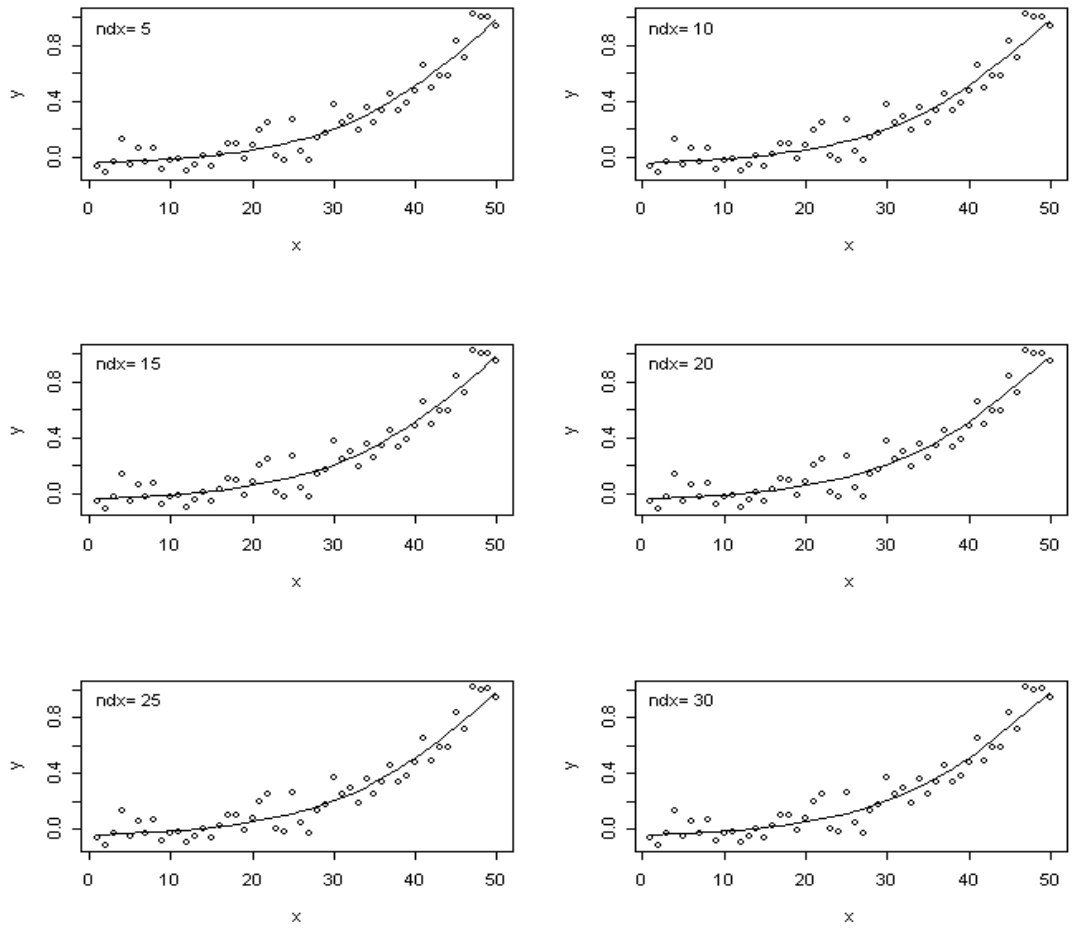

Figure 2.5: Penalized regression using AIC criterion for several values of ndx .

CHAPTER

3

Generalized linear models

3.1 AN OUTLINE OF GLM

3.2 AN ACTUARIAL APPLICATION OF GLM

3.3 ESTIMATING THE HAZARD RATE

3.4 SMOOTHING MORTALITY DATA WITH PENALTIES

3.5 OPTIMAL SMOOTHING

3.1 AN OUTLINE OF GLM

In chapter 2 we considered standard linear regression models where we assumed that the response variables are independent, normally distributed, with constant variance. However, that is not always the case. A wide variety of models with a categorical response is a typical example where the assumption of continuity and normality cannot be accepted as reasonable. Consider for example a researcher interested in predicting the number of children in a family as a function of income, education and several other socio-economic variables. The dependent variable in this case (number of children) is discrete and thus the linear multiple regression model would be inadequate to use for interpreting the data.

Although several non-linear or non-normal regression models have been studied individually for years, it was only in 1972 that Nelder and Wedderburn demonstrated the idea of Generalized Linear Models (GLM) by providing a unified framework for a class of such models. Since then, the class of GLM is one of the most frequently used statistical tools of the applied statistician, exerting an enormous influence in this area of mathematics. GLM are essentially an extension of classical linear models, flexible enough to be used for regression modelling for non-normal data but also allow for most of the familiar ideas of normal linear regression to carry over. The important assumption of independent observations made in linear models of classical regression analysis is also a characteristic of GLM.

Let's introduce the exponential family of distributions and then define generalized linear models.

If a random variable Y, has probability function (if discrete) or probability density function (if continuous) that can be written in the form

$$\mathbf{f_Y(y,\theta;\varphi) = \exp\{[y\theta - b(\theta)]/\alpha(\varphi) + c(y,\varphi)\}} \qquad (3.1)$$

for some functions a( ), b( ) and c( ) and some parameters θ, φ then Y is said to have a distribution that belongs to the exponential family. The parameter θ is called the natural parameter of the exponential family, specific to Y, which will carry information from the explanatory variables and φ is the scale parameter. The specific form of the distribution is determined by the functions a,b and c.

The exponential family just defined includes as special cases the normal, binomial, Poisson, exponential and gamma distributions. GLM are based on the exponential family.

A generalized linear model has the following components:

(i) response variables $Y_1,\ldots,Y_n$ which are assumed to be independent and having the same distribution coming from the exponential family

(ii) a set of parameters $\beta = (\beta_1 \quad . \quad . \quad \beta_p)^T$ whose values are unknown and have to be estimated from the data.

(iii) a set of explanatory variables

$$X = \begin{pmatrix} x_{11} & x_{12} & . & . & x_{1p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n1} & x_{n2} & . & . & x_{np} \end{pmatrix}$$

(iv) a link function g, monotonic and differentiable such that

$g(\mu_i) = \eta_i = \sum_{j=0}^{p} x_{ij}\beta_j$, $i= 1,2,...,n$ where $\mu_i = E(Y_i)$ and η_i is the *linear predictor.*

The general form of a GLM is $Y_i = \mu_i + \varepsilon_i$, $i = 1,...,n$.

While in theory the link function may be any monotonic and differentiable function, in practice it is chosen depending on the assumed distribution of the response variables. So for the Normal distribution we use the identity link g(z)=z, for the Binomial the logit link g(z)=log(z/1-z), for the Poisson the log link g(z)=log(z) etc. Other functions of course may be used. The particular choice of link function for each particular exponential family density (3.1) has particular significance mathematically and to a lesser extent statistically. If *g* is the inverse of the function b' then θ coincides with the linear predictor and we call this particular *g* the *canonical* link function for the model. This choice slightly simplifies the algebra and the algorithms.

The assumptions made under a GLM are loose enough to accommodate a wide class of models useful in statistics but

tight enough to allow the development of a unified methodology of estimation.

To summarize, a generalized linear model differs from the general linear model in two important aspects. Firstly, the response distribution can be non-normal and furthermore discrete and secondly the response variable values are estimated from a linear combination of predictor variables, which are related to the dependent variable via a link function. The general linear model can be considered a special case of the GLM where the dependent variable values follow the normal distribution and the link function is a simple identity function.

Parameter estimates are obtained using the principle of maximum likelihood which requires iterative re-weighted least square procedure. There are many iterative methods for Maximum Likelihood (ML) estimation in the GLM of which the Newton-Raphson and Fisher Scoring method are among the most efficient and widely used. GLM can be fitted and evaluated using SPLUS, SAS, GLIM,R and a number of other statistical packages.

3.2 AN ACTUARIAL APPLICATION OF GLM

A fundamental concept in survival models is the force of mortality μ_x, also known as the hazard rate. Intuitively for small dx, $\mu_x dx$ is approximately the probability that a life who survived to age x dies in the small interval of age x to x+dx. Most attempts to model human mortality are based upon the observation that for reasonably long periods the probability of death increases as age increases. Furthermore, our experience suggests the force of mortality (at least at older ages) increases exponentially. The Gompertz Model (1825), well known to actuaries, describes human mortality over the whole range of ages based on these observations. The formula for this model is given by $\mu_x = Bc^x$.

The Gompertz Model of mortality is a nice actuarial example of GLM. Consider a group of lives and let E_x be the number of lives age x, and y_x the realization of the random variable Y_x corresponding to the number of deaths. The Poisson model of mortality suggests that $Y_x \sim P(E_x\lambda_x)$ **with mean** $\mu_x = E_x\lambda_x$ **where** λ_x **stands for the force of mortality. For the expected number of deaths we get**

$\log(\mu_x) = \log(E_x\lambda_x) = \log E_x + \log\lambda_x = \log E_x + a + bx$ **which is the loglink**

function incorporating the linear relationship between log(λ_x) and age(Gomperz Model).

3.3 ESTIMATING THE HAZARD RATE

Consider again the Gompertz model for mortality and the Poisson model for deaths introduced in 3.2. We aim to obtain estimates of the parameters a and b using the

principle of maximum likelihood. The likelihood function is given by $L(a,b;d) \propto \prod_{i=1}^{n} e^{-\mu_i} \mu_i^{y_i}$ and the log-likelihood

$l(a,b) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \log \mu_i = -\sum_{i=1}^{n} E_i e^{a+bx_i} + \sum_{i=1}^{n} y_i(\log E_i + a + bx_i)$ for $\mu_i = \mathrm{E}_i e^{a+bx_i}$

We could then write this as:

$$\boldsymbol{l(a,b) = -\sum_{i=1}^{n} E_i e^{a+bx_i} + \sum_{i=1}^{n} y_i(a + bx_i) + const.} \qquad (3.2)$$

The maximisation of the log-likelihood leads to the following equations:

$-\sum_{i=1}^{n} \hat{\mu}_i + \sum_{i=1}^{n} y_i = 0$, $-\sum_{i=1}^{n} x_i \hat{\mu}_i + \sum_{i=1}^{n} x_i y_i = 0$ or in matrix form $\begin{pmatrix} 1^T \\ x^T \end{pmatrix}(y - \hat{\mu}) = 0$ for

$\mu^{\mathrm{T}} = (\mu_1 \ . \ . \ \mu_n)$, $y^{\mathrm{T}} = (y_1 \ . \ . \ y_n)$, $1^{\mathrm{T}} = (1 \ . \ . \ 1)$, $x^{\mathrm{T}} = (x_1 \ . \ . \ x_n)$ and $0^{\mathrm{T}} = (0,0)$. For *X=(1:x),* we can also write these as

$$\boldsymbol{X^T(y - \hat{\mu}) = 0} \qquad (3.3)$$

The above equation holds for multiple regression, polynomial regression or B-spline regression with appropriate modification of the *X* matrix.

The resulting ML equation is not linear but can be solved using the iterative Newton-Raphson scheme or the Fisher scoring method giving

$$\hat{\Theta} = \tilde{\Theta} + (X^T \tilde{W} X)^{-1} X^T (y - \tilde{\mu}) \tag{3.4}$$

where $\Theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, $\tilde{\Theta}, \tilde{\mu}$ are current approximations to the problem and $W = diag(\mu)$ the diagonal matrix of weights. The estimates can be obtained by using the fitting function *glm* in S-Plus or an iterative scheme.

In this project we apply the Gompertz assumption for mortality for Poisson regression. We attempt to analyse a set of mortality data provided by the Continuous Mortality Investigation Bureau (CMIB) which gathers information on claims from UK life insurance companies. The data set we consider is for 20 to 90 years-old male policyholders for each of the calendar year 1958-1968. We have the number of years lived which are our exposure data and the number of policy claims, the death data. Our aim is to obtain estimates of the hazard function at this range of ages. The observed mortalities are called *crude mortality rates* in actuarial language and we expect random fluctuations of those rates, since any estimate is only a sample from the sampling distribution of the estimator, as well as irregularity since they will not progress smoothly as ages rise due to the random sampling involved. From our experience however, we expect, for large enough samples, the true underlying mortality rates to proceed smoothly as we move to higher years of age. Mortality rates moving in discrete steps would not be a reasonable pattern as human mortality is merely affected by the gradual aging progress. Particularly in the actuarial business there is an extra reason we prefer to use smooth set of mortality rates and that is the smoothly progressing rates of the premiums implied.

In Figure 3.1 the dots represent the crude rates and the lines the graduated rates which are our estimates of the actual rates of mortality. The displayed observed and fitted mortalities, both presented in the log-scale, are for policyholders aged 20-90 for the calendar years 1958,1960,1962,1964,1966,1968. We aimed to provide a smooth mortality curve and to reduce the random sampling error (by basically using estimates

at ages close to age x to improve our initial estimate). It can be seen that the estimated log-mortality is a linear function of age. The data suggest that there is a little improvement in mortality over the decade, particularly at higher ages. We also note that mortality rates tend to be underestimated at young ages between 20-30. That would not of course be desirable for any life assurance company as it could lead to wrong calculations of financial functions of interest (e.g. A_x) and therefore the impacts on the assurance company could be disastrous. The above remarks lead us to the conclusion that we need to improve the fit at young ages as the Gompertz model fails to describe efficiently their mortalities.

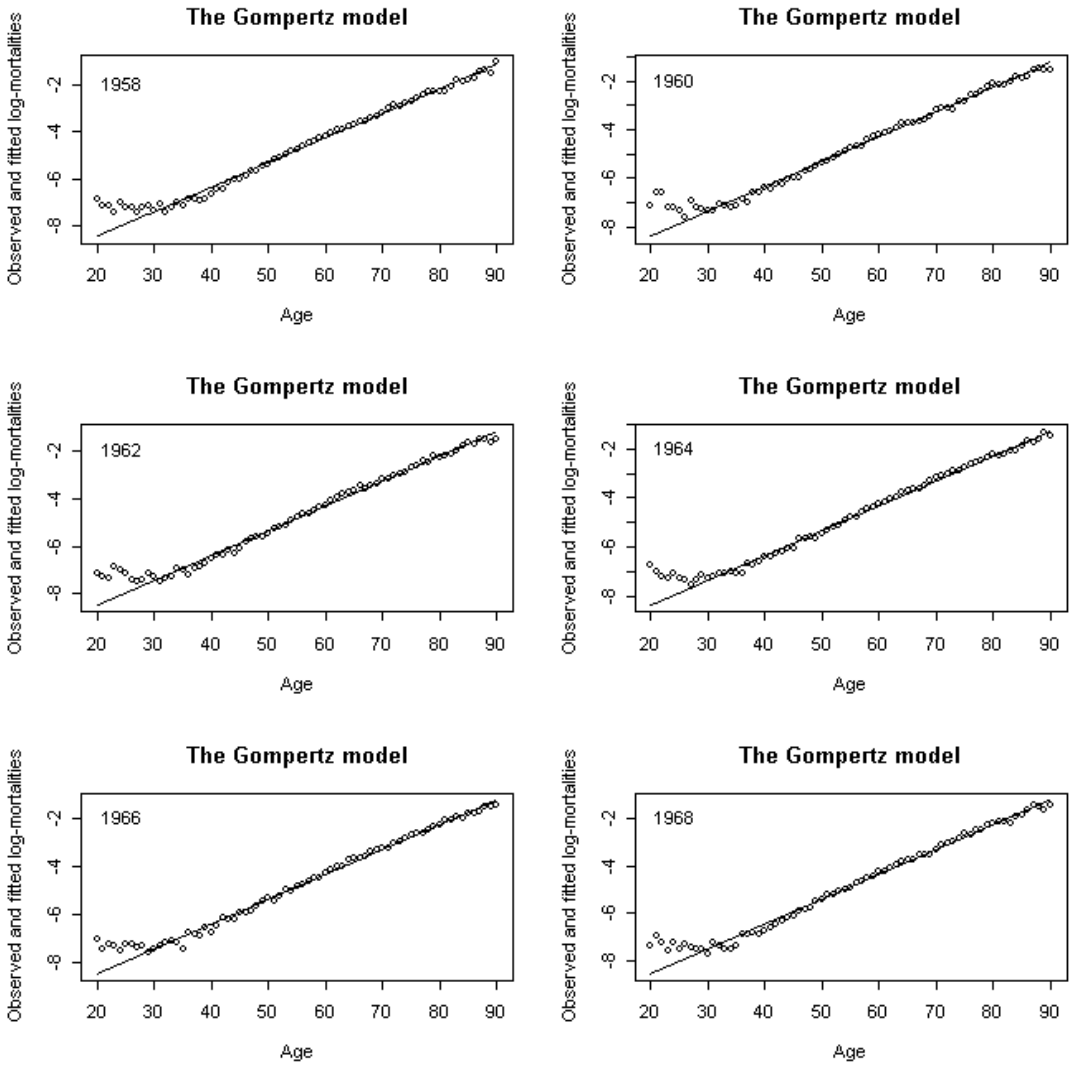

Figure 3.1: Log-mortalities against age.

In Figure 3.2 we interpolated the data set for the calendar year 1968 by using cubic(bdeg=3) B-spline functions as the basis for our regression with number of internal knots 5, 15, 25, 35. The crude estimates have been graduated using the maximum likelihood concept .We note that at ages between 20-25, the crude death rates are higher than those at ages between 30-35. This could be possibly explained by the increased motor accidents and terminal illnesses (e.g. AIDS) observed at these ages as people at younger ages tend to be less sensitive and mature at these issues. Another point we should make is that as we increase the number of intervals in the domain i.e. the number of B-splines the fitted curve tracks the data more closely. But the shape of the resulted curve is not appealing at all. In fact it presents a rather poor graduation. That underlines the point that a good fit to the data is not the one and only aim in curve fitting. Another, often conflicting aim is to obtain a smooth estimated curve. We should therefore find a way so that the number of ndx used would not influence much the shape of the fitted curve and a way that best compromises goodness of fit and smoothness.

It should be stressed that there are situations where the wiggly curves observed for large values of ndx would not be unsatisfactory as explanations of the given data. It may well be that the phenomenon under study is known to vary rapidly and that the given observations are known to be extremely accurate. However, even in this case it is of interest to regard the very local variation in the curve as random “noise” in order to study the more slowly varying trend in the data.

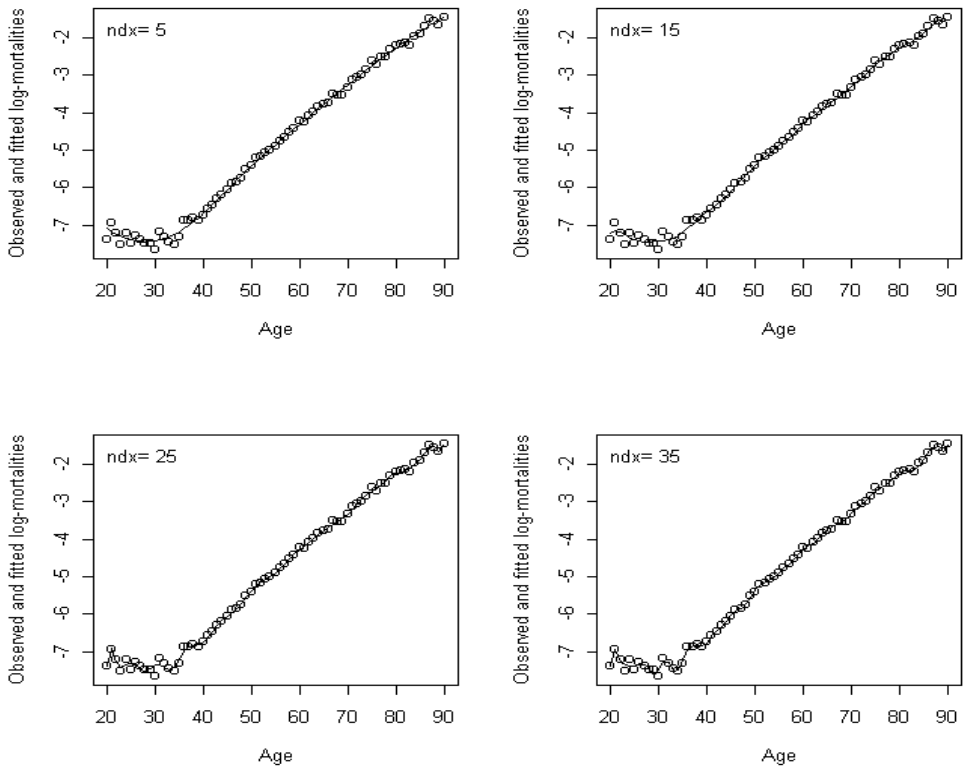

Figure 3.2: Likelihood estimate(-) of the death rate and the crude death rates(o) for the year 1968, using B-splines, both presented on the log-scale.

3.4 SMOOTHING MORTALITY DATA WITH PENALTIES

When we deal with a regression problem, a good fit to the data is not the only aim of curve fitting. Another aim, usually conflicting to some extent, is to obtain a curve estimate that does not involve too much rapid fluctuation. We therefore target to maximize goodness of fit and minimize roughness.

The unconstrained maximization of the log-likelihood will not provide a sensible estimate of the parameters. The likelihood will be maximized by any smooth function that interpolates the data, giving a fairly useless result as it leads

to overfitting and subsequently an implausibly rough fitted curve.

The idea set out in section 2.4 was that the curve estimation requires balance between goodness of fit and roughness. Penalizing the residual sum of squares by adding a roughness penalty term is an obvious way of obtaining this. The approach can be seen as a particular case of the more general concept of penalized likelihood. Good and Gaskins, who first introduced that concept (1971), suggested subtracting from the log-likelihood a roughness penalty. Instead of maximizing the pure log-likelihood itself, we choose to maximize a modified form of it, the *penalized log-likelihood*. Therefore in the regression context the penalized likelihood is equal to

$$\boldsymbol{l(\Theta) - 1/2\lambda\Theta^T D^T D\Theta} \tag{3.5}$$

where D a difference matrix.

If the smoothing parameter λ is large, then the main component of the modified likelihood will be the roughness penalty term and therefore the fitted curve will display very little curvature. In the limiting case where $\lambda \to \infty$, given that we use a second order penalty, the estimated curve will approach the linear regression fit. For relatively small values of the smoothing parameter the main contribution to the penalized likelihood will be the probability function. In this case the fitted values will

track the data very closely even if this is at the expense of giving a rather variable, wiggly curve. For illustration, see the graphs in Figure 3.3. In the left hand panel where we have used a small value of lambda, the graduated rates adhere too closely to the crude rates and approach the interpolating curve shown in figure 3.2. In the right hand panel we have used a relatively large value of lambda and the graduation seems satisfactory. In both cases we used cubic B-splines for our regression with a second order penalty and value of ndx=35. The data set shown is from the calendar year 1968.

These remarks give rise to the very reasonable question of how to choose the value of λ that would be the most appropriate to a given data set. We attempt to address this problem in the following section.

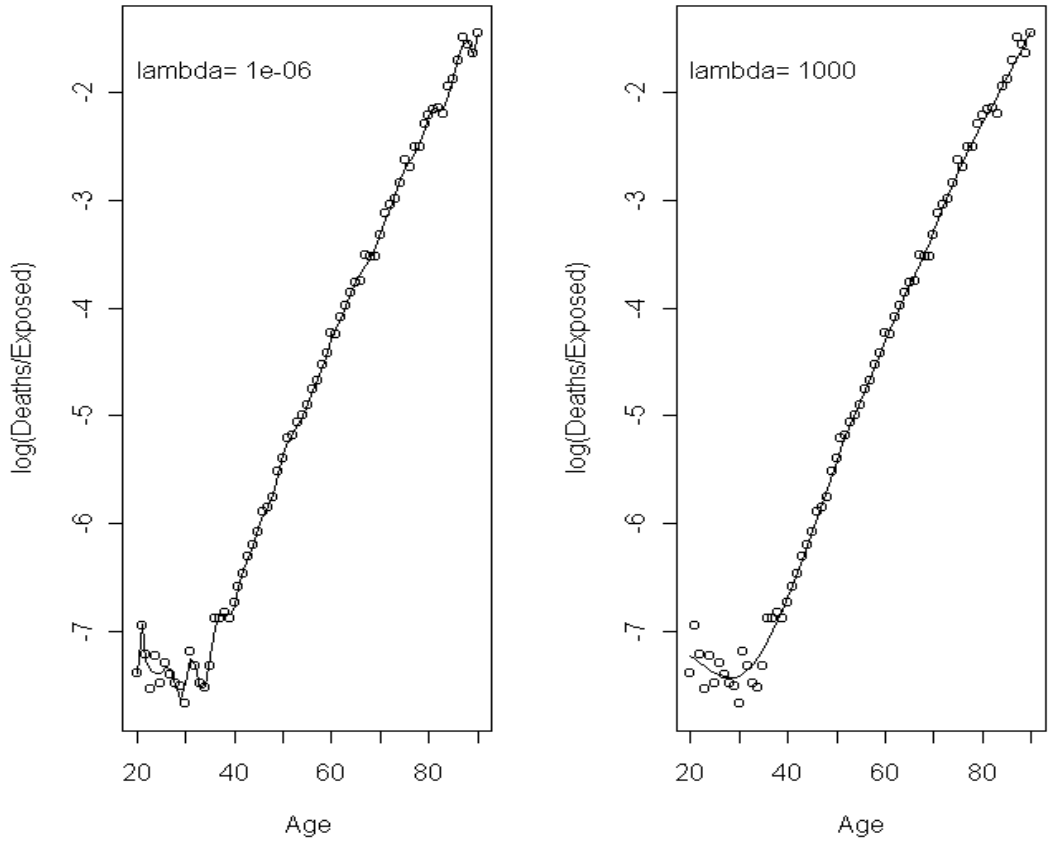

Figure 3.3: A graduation of the log-hazard using the idea of penalized likelihood for a small and large value of lambda

Placing restriction on the likelihood function ensures balance between over-adherence to the data (high values of the log-likelihood) with smoothness of the estimated curve (low values of the roughness penalty). The optimisation of the penalized log-likelihood gives the following system of equations:

$$B^T(y-\mu) = \lambda D^T D\Theta \qquad (3.6)$$

for $B = (B_1(x),...,B_k(x))$ **an** $m \times k$ **matrix of B-splines where k depends on the number of knots and the degree of the B-spline. These equations are solved using iterative weighted linear regressions giving the system**

$$(B^T\tilde{W}B + \lambda D^T D)\hat{\Theta} = B^T(y-\tilde{\mu}) + B^T\tilde{W}B\tilde{\Theta} \qquad (3.7)$$

where $\tilde{\Theta}$ **and** $\tilde{\mu}$ **are current approximations to the problem** $W = diag(\mu)$ **the diagonal matrix of weights.**

The main conceptual advantage of the roughness penalty method applied to the mortality data, is that it considers automatically the fact that the variability of the crude death rates differs greatly as we move in different sections of the

age range and that is a consequence of both the variation in the cohort size and in the underlying mortality rate.

3.5 OPTIMAL SMOOTHING

In the previous section we showed how we can control the smoothness of the fitting curve with λ. A way of getting an optimal value of the smoothing parameter is the application of the Akaike information criterion, which its general definition is given by

$$AIC(\lambda) = Deviance + 2tr(H) \tag{3.8}$$

where

$$H = B(B^T WB + \lambda D^T D)^{-1} B^T W \tag{3.9}$$

The best value of λ is the one that gives AIC the minimum value. The deviance is a measure of the closeness of the fit of the model to the data and it can be interpreted as the residual sum of squares in the linear model case. The trace of the hat matrix determines the effective dimension of the B-spline smoother.

We observe that AIC is easier to compute for a Poisson distribution rather than the normal distribution as the relationship between mean and variance of the former is known. To make the computational part easier is suggested to use $tr(H) = tr((B^T WB + \lambda D^T D)^{-1} B^T WB)$ as $tr(AB) = tr(BA)$ for conformable matrices.

AIC basically corrects the log-likelihood of a fitted model for the effective number of parameters. We have chosen to use this criterion for optimal smoothing as it is computationally easy and fast but it is important to remark that it is likely to result in undersmoothing for data with much variation as the assumed variance of the data may be too low.

In our data set for the year 1968, we search for an optimal value of λ on the geometric grid $10^{-6},10^{-5},\ldots,10^{6}$. For our estimate we used penalized spline fit with a quadratic penalty on the knot coefficients. The value of lambda that minimizes AIC for each of the number of ndx used and the corresponding minimum value of AIC is displayed in table 3.1.

Table 3.1

ndx	lambda	AIC
5	1	103.9
15	100	98.3
25	100	99.7
35	1000	98.3

Figure 3.4 shows the log-hazard against age for the data set of the calendar year 1968. The estimated curves look identical to the naked eye even if we have used different values of ndx for each graduation. That is evidence of the efficiency of the AIC criterion for optimal smoothing. For a right combination of ndx and lambda, so that AIC is minimized we managed to balance goodness of fit and roughness.

Actuaries would not of course work directly from a graph like those in Figure 3.4 but would use the numerical values of the estimated death rates for further calculations.

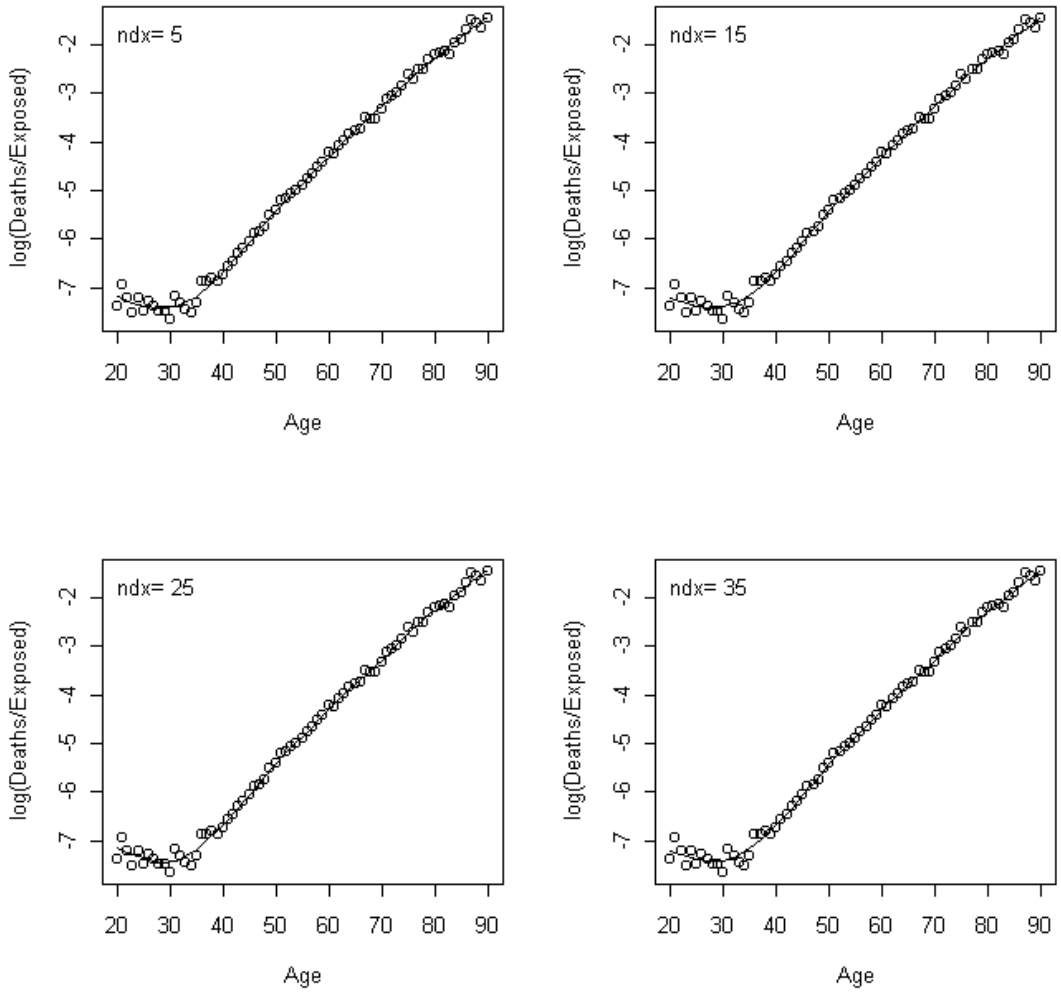

Figure 3.4: Estimated Log-hazard against age for the data set of year 1968.

The idea set out from the outset of this project was the direct concept of balancing the

two aims in curve estimation: goodness of fit and smoothness. Penalizing the log-likelihood by subtracting a roughness penalty term is an obvious way of making the necessary compromise explicit. AIC provides a good method

for “choosing” the smoothing parameter and therefore the “rate of exchange” between deviance and local variation.

CHAPTER 4

Smoothing two-dimensional Poisson data

4.1 INTRODUCTION

In the previous chapters we have considered the applications of roughness penalties in linear modelling and generalized linear modelling. In particular, in chapter 2 we used P-splines as a method of smoothing in linear models. The method was based on using B-splines as the basis for the regression and calculate the regression coefficients by minimizing the penalized least squares which is obtained by adding to the classical least squares a penalty term to control the smoothness of the model. Similarly in chapter 3 we used B-splines to smooth one-dimensional count data with Poisson errors. The method was based on the idea of modifying the log-likelihood by a difference penalty on the regression coefficients. We illustrated the method by modelling the mortality against age for several years between 1958 and 1968.

The approach has several merits clearly deserving a respectable place in smoothing methodology. In order to increase the credibility of Eiler's and Marx's claim that P-splines come close to the "ideal smoothers" we should address the many important advantages that this method enjoys. Among others we only mention the simplicity of the idea (roughness penalty based on B-splines coefficients to prevent overfitting) which is certainly appealing, the reduction of the dimensionality of the problem and the computational burden (which can easily be incorporated in standard software), the connection to smoothing spline and polynomial regression and its flexibility to be applied in different modelling situations.

From all the above mentioned attractive properties of P-spline smoothing we stand at the last one, that it is flexible enough to be applied in different modelling situations. It would be interesting to see an application of this methodology in multivariate problems. In this chapter we explore the extension of the method to the more challenging two-dimensional Poisson data, which was our main goal from the beginning of this project. We will see that the ideas of P-spline regression described in the previous chapters apply equally well in the context of the two-dimensional regression model of this chapter. We illustrate that by developing a model and analysing a set of mortality data provided by CMIB considering the mortality of male policyholders aged 20-90 for the years 1958-1968. The method models the age and year factor simultaneously i.e. describes mortality as a surface above the age and year plane. We pay particular attention to mortality trends over the decade and the extent to which these trends are influenced by the age factor. The application shows the very wide potential applicability of P-splines smoothing.

4.2 A MODEL FOR TWO-DIMENSIONAL POISSON DATA

We'll start our discussion for the two-dimensional Poisson data by putting them in the context of multiple regression and its simplest form the Bivariate regression. We have seen in section 2.1 that there are many problems in which predictions of one variable could be improved if we consider additional relevant information. For instance, we should be able to make better predictions for a person's future lifetime if we consider not only the age factor but also the sex, the smoking status, the socioeconomic level etc.

For describing such relationships, we usually use the linear equation of the form

$$\boldsymbol{Y = b_0 + b_1 X_1 + b_2 X_2 \ + ... + b_p X_p + e}$$

where $\boldsymbol{Y}$ is the random variable whose values we want to predict in terms of given values $X_1, X_2 \ ... X_p$ and $b_0, b_1, ... b_p$ the multiple regression coefficients. For our data set

we consider the dependent variable to be the observed mortality at each age and year and the explanatory variables to be the age and year of death. The derivation of the regression coefficients corresponds to the work of section 2.2. In this chapter we depart from the classical procedure and we arrange our two-dimensional data in a way that can be presented in a mathematical sense as one-dimensional data and furthermore can be handled by computers relatively easy.

We set up our notation similar to that used in the one-dimensional case. Let $Y=(y_{ij})$ be the matrix of the observed deaths at age i, i=1,2,…,m and year j, j=1,2,…,n. We arrange this matrix in vector form by considering the column order from left to right. Let the resulting vector of length $N = m \times n$ be denoted by y. We work similarly for the exposure matrix $E=(E_{ij})$ and the mortality matrix $\Theta=(\theta_{ij})$ and let the corresponding vectors be denoted by e and θ respectively. We no longer require the one-dimensional splines obtained by holding either the age variable or the year variable fixed to be simple splines. We therefore consider the B_a matrix of dimension $m \times n_a$ to be the set of B-splines basis for smoothing in the one-dimensional case by age considering a single year (the number of columns of B_a depends on the number of knots and the degree of the B-spline). Similarly we consider B_y be the $n \times n_y$ one-dimensional B-spline basis for smoothing by year for a single age. We then consider the Kronecker/ tensor product of the two B-splines basis, which is a systematic method of using families of smooth functions on one dimension to generate smooth surfaces in higher dimensional spaces. We define our regression B-spline matrix of dimension $mn \times n_a n_y$ to be the tensor product

$$\boldsymbol{B} = \boldsymbol{B}_y \otimes \boldsymbol{B}_a \tag{4.1}$$

Assuming that the number of deaths Y_x has a Poisson distribution with mean $\mu_x = E_x\theta_x$ we get $log\theta = Ba$, where a is the vector of length $n_a n_y$ (number of columns in B) corresponding to the regression coefficients. We write a in matrix form as

$$\boldsymbol{A} = (\boldsymbol{a}_1,...,\boldsymbol{a}_{ny}), \quad \boldsymbol{A}' = (\boldsymbol{a}_1{}^r,...,\boldsymbol{a}_{na}{}^r) \tag{4.2}$$

Similar to the one-dimensional case it is not appropriate to use the log-likelihood directly for calculation of the regression coefficients. Motivation for the penalized likelihood criterion is by now familiar. In the absence of the penalty term the optimization of the likelihood would yield over-fitting. Instead we choose to smooth the entries of A by imposing penalties on each row and column of the matrix. We now find A by maximizing the penalized log-likelihood

$$l(A;Y) - \frac{1}{2}\lambda_a \sum_{j=1}^{n_y} a'_j D'_a D_a a_j - \frac{1}{2}\lambda_y \sum_{i=1}^{n_a} a_i{}^{r'} D'_y D_y a_i{}^{r} \tag{4.3}$$

where D_a and D_y are difference matrices on columns (age) and rows (years) respectively. For $a = (a'_1,...,a'_{n_y})'$ and the penalty matrices for age and year $P_a = I_{n_y} \otimes D'_a D_a$, $P_y = D'_y D_y \otimes I_{n_a}$ expression (4.3) can be written in column form as

$$l(a;y) - \frac{1}{2}\lambda_a a'(I_{n_y} \otimes D'_a D_a)a - \frac{1}{2}\lambda_y a'(D'_y D_y \otimes I_{n_a})a \tag{4.4}$$

$$= l(a;y) - \frac{1}{2}a'(\lambda_a P_a + \lambda_y P_y)a \tag{4.5}$$

We have now obtained a rather elegant model that fits the age and year effect simultaneously. As in section 3.4 when we dealt with the one-dimensional case (equation 3.7) we calculate the regression coefficients by solving the iterative scheme

$$(B^T \tilde{W} B + P)\hat{a} = B^T (y - \tilde{\mu}) + B^T \tilde{W} B \tilde{a} \tag{4.6}$$

where $\tilde{a}$ and $\tilde{\mu}$ are current approximations to the problem, $W = diag(\mu)$, the diagonal matrix of weights, and $P = \lambda_a P_a + \lambda_y P_y$.

Just like in the one-dimensional case we consider the classical selector AIC for choosing the value of the

smoothing parameters, which can be easily extended to deal with the two-dimensional case. However as it has already been pointed out in

section 3.5 AIC is likely to result in undersmoothing for over dispersed data giving a non-satisfactory result. We therefore consider also the Bayesian information criterion (BIC) as a method for smoothing parameter selection, which introduces much heavier penalties and therefore produces smoother curves. Both selectors have the form

$$dev(y; a, \lambda_a, \lambda_y) + \delta tr(H)$$

(4.7)

where *tr(H)* is the trace of the hat matrix and $\delta=2$ for AIC and log(N) for BIC. With our data set N=781 giving log(N) the approximate value of 6.66 and thus the strength of the penalty in BIC is far greater than that used by AIC. The minimization of the function can be approached by a simple grid search.

Conceptually the maximization of the penalized log-likelihood can be considered in two steps. First consider either of the two methods for choosing the smoothing

parameters and then iterating the process to convergence to get the estimates.

Many applications of regression involve large number of fitted parameters. The numerical complications encountered when we are dealing with a large data set may be then enormous. The preceding discussion has introduced a way of overcoming the computational difficulties by mainly take advantage of the important property of P-splines to reduce dimensionality. The methodology described above was originally developed by Durban, Currie and Eilers (2002). Further details can be found in the *Proceedings of the 17th International Workshop on Statistical Modelling, Crete, 207-214* which includes a discussion of other models for two dimensional Poisson data, computational problems that arise, economies of calculations and illustration with the analysis of a large set of mortality data.

4.3 ANALYZING THE CMIB MORTALITY DATA

In this section we demonstrate the methodology introduced above by applying the suggested model to the CMIB mortality data described in 3.3. For fitting the log-hazard against ages and years we calculated both methods for choosing the smoothing parameters but we mostly displayed the ones obtained by BIC, since AIC selects smaller smoothing parameters and produced very rough estimated curves.

For the smoothing operation we chose as basis function a set of cubic B-splines (bdeg=3) and equally spaced internal knots which were taken to cover the range of age and year. Optimization of the penalized log-likelihood was obtained using a second order penalty (pord=2). Table 4.1 summarizes the results giving the minimum values of BIC and AIC (bold) for the different values of ndx_a and ndx_y, the total number of fitted parameters (npar) and the effective dimension of the fitted model (tr). Note the very large difference in the effective dimensions of the fitted model selected by BIC and AIC.

Table 4.1

ndx_a	ndx_y	npar	λ_a	λ_y	tr	BIC	AIC
20	**3**	**138**	**190**	**130**	**28**	1612.46	**1482.5**
			10	**0.1**	**66**	**1720.35**	1410.6
20	**4**	**161**	**180**	**420**	**28**	1612.66	**1483**
			10	**0.1**	**77**	**1765.38**	1405.65
10	**3**	**78**	**14**	**80**	**26**	1589.	**1469.**

						01	**62**
			3	**0.1 1**	**44**	**1626. 76**	1419. 6
10	**4**	**91**	**12. 5**	**31 0**	**25**	1586. 51	**1468. 39**
			1.3	**0.1**	**55**	**1677. 25**	1418. 98

The cross section of the fitted surface corresponding to age 65 is displayed in Figure 4.1. In the left hand panel the number of internal knots for age was 10 (ndx_a=10) and for years 3 (ndx_y=3), while in the right hand panel ndx_a=10 and ndx_y=4. The fit of the model selected by AIC is omitted from this plot since it leads to under smoothing revealing the evidence that this criterion is not suitable for over dispersed data. Inspection of the two plots reveals that the number of knots does not play a crucial

role in curve estimation as they give more or less the same fitted values. It is worth remarking how the mortality falls off quite rapidly after 1962.

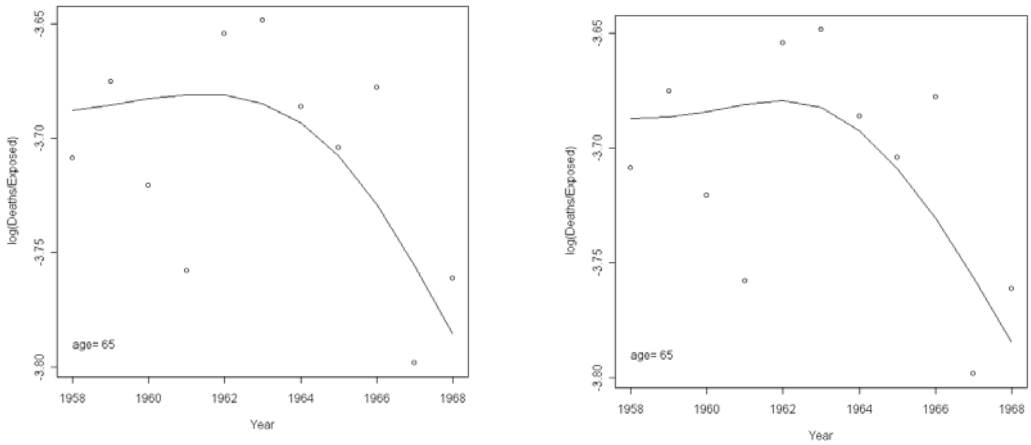

Figure 4.1: **Log hazard against year for age 65.**

Considering the fitted surface constructed for ndx_a=20, ndx_y=4 and BIC as the smoothing parameter selector we also extracted the cross sections for ages 60, 80, 37 and 46. The plots for ages 60 and 80 are shown in Figure 4.2 while for ages 37 and 46 in Figure 4.3. Mortalities in high ages seem to have improved with that of age 60 more rapidly. The same pattern applies to younger ages but the improvement is not so marked.

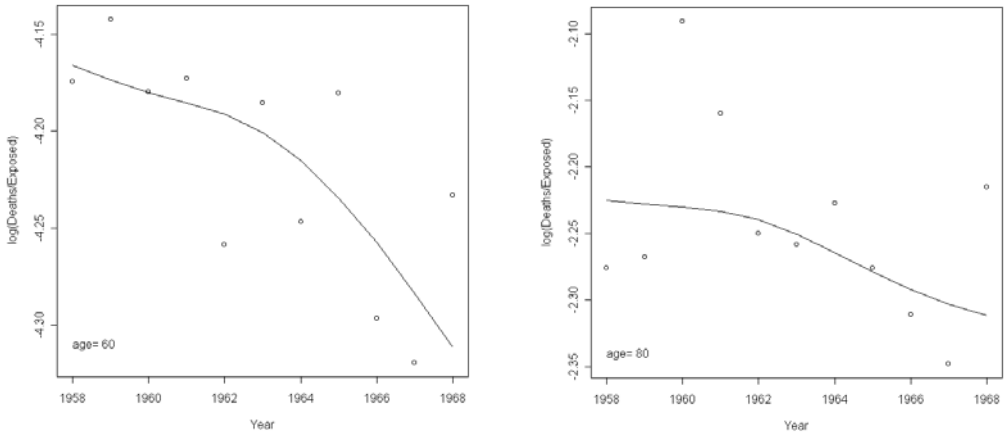

Figure 4.2: **Log hazard against year for ages 60 and 80.**

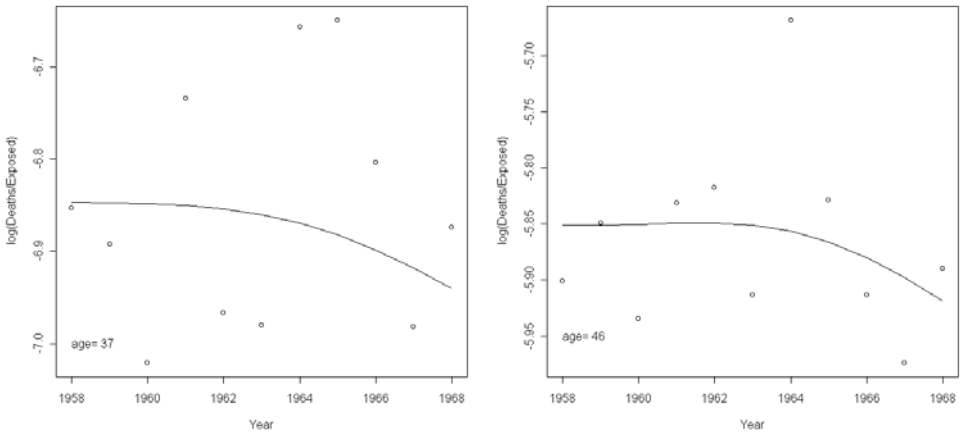

Figure 4.3: **Log hazard against year for ages 37 and 46.**

Figure 4.4 is a plot of the log-mortality against age for years 1958, and 1968. The solid line shows the fitted curve when we chose the degree of smoothing using BIC whilst the dotted line shows the fitted curve when we used AIC as a smoothing parameter selection. For our B-spline basis ndx_a was 20 while ndx_y was 4. There is little to choose between the two fits but we would prefer the slightly stiffer fit suggested by BIC.

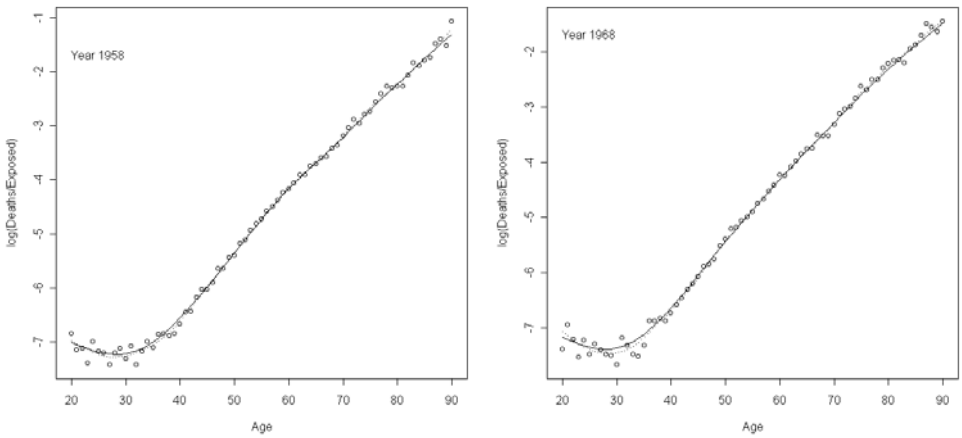

Figure 4.4: Log hazard against age for the year 1958, 1968. Smoothing parameter selection: BIC, solid line; AIC, dotted line.

In Figure 4.5 we produced the fitted mortality surface over the age and year plane. The ability of the model to produce a smooth surface over the age and year plane is clear.

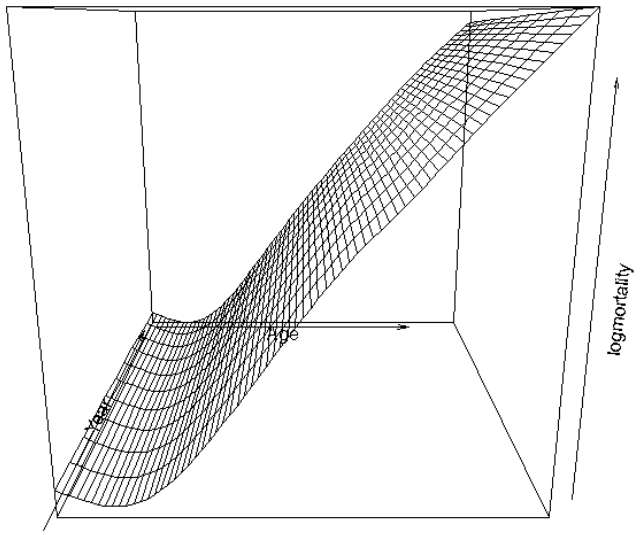

Figure 4.5: Fitted mortality surface

One of the attractions of the model as illustrated above is obviously the happy combination of circumstances both that the estimate is the solution of a neatly expressed and intuitively attractive optimization and that it can be calculated *keeping the computational side of the fitting process under control.* Even though the two-dimensional case poses considerable computational burden it seems that the special structure of the penalized log-likelihood and the

cubic B-splines is sufficient to cope with it producing a satisfactory description of the data.

Overall this analysis, in common with that of other data sets studied, confirms a pattern of improved mortality over the years. Especially in higher ages (after 60) this improvement occurs more rapidly. It is desirable for an actuary to observe these trends carefully since human mortality plays a crucial role in the calculation of financial factors e.g. A_x , a_x and acknowledge the financial implications that this improvement has for the insurance business.

CHAPTER 5

Concluding remarks

One of the beauties of the P-spline approach in curve estimation is its conceptual versatility. We demonstrated that by considering its applications in situations arising from linear modelling and generalized linear modelling and by exploring the extension of the method to consider the more challenging two-dimensional case. We have shown how P-spline smoothing provides a flexible data-fitting methodology constituting a respectable addition to smoothing techniques that are still a driving force in the development of nonparametric regression.

In addition to explanation of the formal statistical background and the theory we also embarked on discussing the method in the light of its applicability for the analysis of real data. Since human mortality is prominent in the field of actuarial science and a subject of interest to an actuary required to carry out the tasks of pricing and valuation we considered a set of mortality data provided by CMIB and we analysed the trends over the years. The fitted model has proven to perform well in describing the underlying mortality pattern and we demonstrated that by presenting graphs which are of great importance for any exploratory analysis. Although the estimation of the fitted curves was computationally demanding the special structure of P – splines simplified and accelerated significantly the computational process. Even though we discussed the model in the context of smoothing of 2-dimensional Poisson data it is clear that the work presented applies equally well in a broader glm setting which includes other important models like the Binomial.

The crucial fact arising from the analysis of the CMIB data confirms that mortality has improved rapidly over the last years. For comparison we plot several cross- sections of the fitted mortality surface showing either estimated mortalities

against age for several years or estimated mortalities against year for certain ages. As

expected, we observed the increase in mortality rates as we moved in higher ages with some pattern of high mortality at young ages possibly due to accidental causes which are much more significant at those ages. What all of these plots had in common though, was the suggestion of a significant fall in mortality rates over time and that was especially visible at older ages (over 60). This feature of course implies important financial consequences for the insurance industry and it should be reflected in the mortality assumptions used by life offices when calculating financial functions of interest.

References

Paul H.C.Eilers and Brian D.Marx (1996) *Flexible smoothing with B-SPLINES and penalties*, Statistical Science, 89-101, US

Maria Durban, Iain Currie and Paul Eilers (2001) *Using P-splines to smooth two-dimensional Poisson data*, Proceedings of the 17th International Workshop on Statistical Modelling, 207-214, Crete

W.N.Venables and B.D.Ripley (1999) *Modern Applied Statistics with S-PLUS* Springer

Annette J.Dobson (1983) *An introduction to statistical modeling,* Chapman and Hall, London

J.H.Ahlberg, E.N.Nilson, J.L.Walsh (1967) *The Theory of Splines and Their Applications*, Academic Press, New York

Richard H.Bartels , John C.Beatty and Brian A.Barsky (1987) *An Introduction to Splines for use in Computer Graphics and Geometric Modeling,* Morgan Kaufmann Puplishers, California

Annette J Dobson (1990) *An introduction to generalized linear models,* Chapman and Hall, London

P.J.Green and B.W. Silverman (1995) *Nonparametric Regression and Generalized Linear Models, a roughness penalty approach*, Chapman and Hall, London

McCullagh, P. and Nelder, J.A.(1983) *Generalized Linear Models,* Chapman and Hall, London