International Statistical Review (1987), **55**, 3, pp. 245–259. Printed in Great Britain

Penalized Likelihood for General Semi-parametric Regression Models

Peter J. Green

Department of Mathematical Sciences, University of Durham, Durham DH1 3LE, U.K.

Summary

This paper examines penalized likelihood estimation in the context of general regression problems, characterized as probability models with composite likelihood functions. The emphasis is on the common situation where a parametric model is considered satisfactory but for inhomogeneity with respect to a few extra variables. A finite-dimensional formulation is adopted, using a suitable set of basis functions. Appropriate definitions of deviance, degrees of freedom, and residual are provided, and the method of cross-validation for choice of the tuning constant is discussed. Quadratic approximations are derived for all the required statistics.

Key words: Basis functions; Composite likelihood function; Cross-validation; Decoupled likelihood; Maximum penalized likelihood estimation; Nonlinear regression; Roughness penalty; Smoothing.

1 Introduction

It frequently arises that a statistician has some faith in the validity of a certain parametric statistical model for his data, but for some suspected inhomogeneity with respect to one or more extraneous variables. Typically, such variables might represent space or time, the relationship between them and the response is not of primary interest, and the statistician is inhibited from extending his parametric model to encompass them because of a lack of experience, information or theory about the form of their relationship. A simple example might arise with binomial data from different geographical locations where it might be quite reasonable to model the response probability as a logistic regression on various explanatory factors or covariates, but influenced also by environmental effects, unknown in form but believed to vary smoothly with location. In such situations, procedures derived from penalized likelihoods (Good & Gaskins, 1971; Silverman, 1985a) may well be appropriate. The purpose of the present paper is to examine properties of such methods in the context of the rather general class of regression models used by Green (1984), characterized as probability models expressed as composite likelihood functions. (This is not to claim that such a view of regression is universally appropriate.) The methods discussed combine the ideas of fitting the parametric part of the model by maximizing the likelihood whilst smoothing with respect to the extraneous variables.

Any regression model consists of random and systematic components. We consider models in which the random component is specified by a log likelihood function $L(y; \theta)$ for the responses y in terms of an n-vector of predictors θ: a concrete example would be a particular exponential family

$$L(y; \theta) = \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i) + c(y_i)\}.$$

The systematic component of our models is a specification of the general form of dependence of θ on explanatory factors and covariates. The focus of attention is on this dependence, and we suppose that having recorded the explanatory variables, θ is known but for a p-vector of parameters β and a real-valued function γ lying in some prescribed linear space $\mathcal{G}$. A concrete example would be the simple additive semi-parametric regression function:

$$\theta_i = x_i^{\mathrm{T}}\beta + \gamma(t_i), \tag{1.1}$$

where x_i and t_i (both in general vector-valued) are the observed values of explanatory variables for the ith case.

Thus the complete model we consider is the composite likelihood function

$$L(y; \theta(\beta, \gamma)) = L(\theta(\beta, \gamma)) \tag{1.2}$$

in which the responses y, and the explanatory variables, can be suppressed from the notation. Our principal interest will be in estimating β.

For an explicit example, consider a logistic regression model in which the 'intercept' term varies in time. Then if the ith case has y_i successes out of m_i trials, with covariates x_i recorded at time t_i, we could write

$$L(\theta) = \sum_{i=1}^{n} \{y_i \log \pi_i + (m_i - y_i) \log (1 - \pi_i)\},$$

where $\pi_i = \exp(\theta_i)/\{1 + \exp(\theta_i)\}$ and θ_i is given by (1.1).

This simple problem is typical of many where maximum likelihood leads to over-fitting, in the absence of any restriction on the form of the function γ; in particular, parameters will be unidentifiable. In the context of density estimation, Good & Gaskins (1971) proposed maximizing instead the penalized likelihood

$$L(\theta(\beta, \gamma)) - \tfrac{1}{2}\lambda J(\gamma), \tag{1.3}$$

where J is a roughness penalty, increasing as the function γ becomes less smooth, and λ is a nonnegative tuning constant or hyperparameter which may be adjusted to control the smoothness of the fitted γ. There is of course a Bayesian interpretation; see § 4. For an excellent account of the roughness penalty approach to nonparametric linear regression, the reader is referred to Silverman (1985b). The present work is in the same spirit but aims to extend these ideas by introducing parametric terms into the regression function as well, and by allowing an arbitrary log likelihood L.

If the log likelihood L is that of independent observations y_i, normally distributed with means θ_i given by (1.1), then this penalized likelihood approach is equivalent to a semi-parametric linear regression as proposed by Green, Jennison & Seheult (1983, 1985) in the context of agricultural field experiments, Engle et al. (1986) in an econometric problem, and Wahba (1984), who with co-workers has developed a considerable body of theory for 'partial-spline' methods. Use of penalized likelihood in simple generalized linear models is discussed by O'Sullivan in his thesis (1983), by O'Sullivan, Yandell & Raynor (1986) and by Silverman (1985b). Leonard (1982) considers such methods for a variety of curve estimation problems from a full-blooded empirical Bayesian perspective. In all of these papers the parametric part of the model is not present, but Wahba (1985) remarks that the ideas of partial splines may be combined with penalized likelihood for generalized linear models. In none of these papers is the general regression model (1.2) addressed, and typically estimation of γ is not regarded as of subsidiary importance to that of β.

The remainder of the paper is organized as follows. In § 2 we discuss maximum

penalized likelihood estimation for regression models, including a prototype algorithm for obtaining such estimates numerically. Sections 3 and 4 respectively describe some important special cases, and discuss appropriate choice of the roughness penalty J. Sections 5, 6 and 8 provide information on various auxiliary statistics needed in inference about these models, including standard errors for the parameter estimates, goodness-of-fit statistics, and residuals. Section 7 refers to the problems of automatic data-based choice of the tuning constant λ, and the final section discusses complications caused by nuisance parameters.

The peculiar nature of penalized likelihood methods combines their strong plausibility with at best sketchy mathematical justification. As it ranges over all of these aspects of a complete analysis, this paper aims both to make recommendations to the practical statistician on how to use these methods, and to offer to the mathematical statistician suggestions of some profitable lines for further research.

2 The estimation procedure

We begin by apparently compromising the generality of prescription of our problem. As it stands, (1.2) allows the predictor θ to depend on the infinitely many values of the function γ. In practice, therefore, discretization will be necessary at some stage. Following a suggestion of Leonard (1982), in maximizing (1.3), we will restrict γ to lie in a finite-dimensional linear subspace $\mathcal{F}$ of $\mathcal{G}$, namely $\mathcal{F} = \text{span}\{\varphi_j : j = 1, 2, \ldots, q\}$ for a prescribed set of q basis functions. These functions may depend on the explanatory variables but not on the responses. We use the abbreviation

$$\theta(\beta, \xi) = \theta\Big(\beta, \sum_{j=1}^{q} \xi_j \varphi_j\Big), \tag{2.1}$$

and will further restrict attention to roughness penalties of the form

$$J\Big(\sum_{j=1}^{q} \xi_j \varphi_j\Big) = \xi^{\mathrm{T}} K \xi$$

for some fixed $q \times q$ nonnegative-definite matrix K.

It may seem that we are abandoning our intended semiparametric framework, but it should be stressed that q, while it may be somewhat less than n, will still be 'large', and parametric estimation of ξ will not be appropriate. Further, the intention is that $\mathcal{F}$ and $\mathcal{G}$ should in practical terms be indistinguishable. This will entail appropriate choice of $\{\varphi_j\}$ as, for example, a large class of orthogonal polynomials or trigonometric functions in t. This choice will depend on the observed values of t, and will also determine K. The precise quadratic form of the roughness penalty accords with standard practice in several special cases; it is hardly necessary in what follows, but it simplifies the algebra and is likely to make little practical difference. This is discussed further in § 4.

There may in fact be no restriction at all. In nonparametric regression, J is typically a squared semi-norm on a reproducing kernel Hilbert space $\mathcal{G}$; see Aronszajn (1950). Suppose θ depends on γ only through $\{\gamma(t_i), i = 1, 2, \ldots, q\}$ and the $\{t_i\}$ are distinct. Then since

$$\min\{J(\gamma) : \gamma(t_i) = \xi_i, i = 1, 2, \ldots, q\} = \xi^{\mathrm{T}} K \xi$$

for a certain K, and we may choose a basis of spline functions with $\varphi_j(t_i) = \delta_{ij}$, the original and the restricted problem have the same solution so far as values of β and of $\{\gamma(t_i)\}$ are concerned.

We therefore maximize, in place of (1.3), the expression

$$L(\theta(\beta, \xi)) - \tfrac{1}{2}\lambda\xi^{\mathrm{T}}K\xi \tag{2.2}$$

over $\beta \in \mathbb{R}^p$, $\xi \in \mathbb{R}^q$, where θ is a prescribed $\mathbb{R}^n$-valued function. This revised formulation has the further advantage of allowing certain new problems into our framework, that could not otherwise be naturally described with a vector of predictors of finite length.

We will only be concerned with problems where likelihood methods are appropriate: we suppose sufficient regularity that L is approximately quadratic near the 'true values' β_0, ξ_0. A modification to an iteratively reweighted least-squares algorithm derived from the Newton–Raphson method, with Fisher scoring (Green, 1984) should therefore be appropriate. Use of scoring is by no means essential, and observed information could be used in place of expected. Presence of the nonparametric component of the model does not usually complicate the choice of initial estimates.

Write

$$u = \frac{\partial L}{\partial \theta}, \quad A = E\left[-\frac{\partial^2 L}{\partial \theta \partial \theta^{\mathrm{T}}}\right], \quad D = \frac{\partial \theta}{\partial \beta}, \quad E = \frac{\partial \theta}{\partial \xi}.$$

The scores u form an n-vector, and the matrices A, D and E are $n \times n$, $n \times p$ and $n \times q$. All of these quantities in general depend on β and ξ, in the case of u and A only through θ, but these dependencies will be suppressed from the notation. The expectation is taken at the current values of β and ξ. Differentiating (2.2) gives the modified likelihood equations

$$D^{\mathrm{T}}u = 0, \quad E^{\mathrm{T}}u = \lambda K\xi. \tag{2.3}$$

Their solution gives our required maximum penalized likelihood estimates $\hat{\beta}$, $\hat{\xi}$. Typically these equations are nonlinear and require iterative solution. The Newton–Raphson method with expected second derivatives involves successively replacing trial estimates (β, ξ), at which u, A, D and E are evaluated, by (β^*, ξ^*), where

$$\begin{bmatrix} D^{\mathrm{T}}AD & D^{\mathrm{T}}AE \\ E^{\mathrm{T}}AD & E^{\mathrm{T}}AE + \lambda K \end{bmatrix} \begin{bmatrix} \beta^* - \beta \\ \xi^* - \xi \end{bmatrix} = \begin{bmatrix} D^{\mathrm{T}}u \\ E^{\mathrm{T}}u - \lambda K\xi \end{bmatrix}, \tag{2.4}$$

or, equivalently,

$$G\begin{bmatrix} \beta^* \\ \xi^* \end{bmatrix} = (D \quad E)^{\mathrm{T}}AY, \tag{2.5}$$

where

$$H = (D \quad E)^{\mathrm{T}}A(D \quad E), \quad G = H + \begin{bmatrix} 0 & 0 \\ 0 & \lambda K \end{bmatrix}, \quad Y = A^{-1}u + D\beta + E\xi. \tag{2.6}$$

These equations have the form of a combination of weighted normal equations, for β^*, and generalized ridge regression equations, for ξ^*. The two ingredients are given separately by Green (1984) and O'Sullivan et al. (1986). See also Silverman (1985b, § 8.1).

We can now state conditions on the model (1.2), (2.1) for these equations to be soluble. First represent K as $L^{\mathrm{T}}L$, where L is $r \times q$ of full rank r, which is usually less than q. If so, then K has a nontrivial null space: let T be $q \times (q - r)$ such that $LT = 0$ and $[L^{\mathrm{T}} : T]$ is nonsingular. Our conditions are that for all β, ξ, the matrix A is nonsingular, and D, E and $[D : ET]$ have full rank p, q and $p + q - r$ respectively. We may then proceed with any positive finite λ: the matrix G is readily seen to be nonsingular. Convergence of the iteration (2.4) is not guaranteed, but in practice will usually occur rather rapidly for sensible initial values. The algorithm has at least a fixed-point justification: if (2.4) gives $\beta^* = \beta$ and $\xi^* = \xi$ then (2.3) is satisfied.

Jointly with Dr. Brian Yandell, the author is developing various implementations of the basic algorithm (2.4). Details appear elsewhere (Green & Yandell, 1985; Yandell & Green, 1986).

3 Special cases

The general model (1.2) makes no assumptions about the independence of random terms, or additivity and linearity among systematic components. Of course such simplifications are sometimes available. If the log likelihood L is that of n independent observations $\{y_i\}$ each indexed by the corresponding θ_i, then A is diagonal. If θ is linear in β or ξ then D or E will be constant. Such properties may be exploited in algorithms, but do not affect a general treatment.

(a) *The linear independent normal case.* If the observations y are independently normally distributed, $y \sim N(\theta, \sigma^2 I)$ with a linear parameterization $\theta = D\beta + E\xi$, then D and E are constant and $A = \sigma^{-2} I$. The scale factor σ^2 factorizes from both sides of (2.4), so may be ignored on redefining λ: this is an example of a more general phenomenon: see § 9. The artificial response Y in (2.6) is identically y, and no iteration is necessary. If E, ξ, λ and K are omitted, we have the ordinary linear model. If D and β are omitted instead then we have a model including spline smoothing (as described in §2) and ridge regression: when $K = I$ we obtain $\hat{\xi} = (E^{\mathrm{T}}E + \lambda I)^{-1} E^{\mathrm{T}} y$. With both D and $E = I$ present, this covers the least-squares smoothing approach to the analysis of agricultural field trials due to Green et al. (1983, 1985). They used a roughness penalty based on differencing from neighbouring plots, for example

$$\xi^{\mathrm{T}} K \xi = \sum_i (\xi_i - 2\xi_{i+1} + \xi_{i+2})^2. \tag{3.1}$$

In this application, D represents the design matrix in a designed experiment, and the resulting methodology may be related to other more classical approaches (Green, 1985).

In all these special cases, it may be more natural to focus on least-squares rather than normal theory/maximum likelihood as the basic principle.

(b) *Logistic regression.* To continue the example from § 1, we now have $\theta_i = x_i^{\mathrm{T}}\beta + \xi_i$, and find that $u_i = y_i - m_i\pi_i$, A is diagonal with $A_{ii} = m_i\pi_i(1 - \pi_i)$, E is the identity and D has ith row equal to x_i^{T}. The equations (2.4) are no longer fixed and iteration is necessary. An appropriate form for K will depend on the temporal or spatial configuration of the $\{t_i\}$: see § 4.

(c) *A grouped continuous model.* For nonparametric regression of ordered categorical data on a single explanatory variable, the following model may be appropriate. For $r = 1, 2, \ldots, R$ we have an S-vector multinomial response $\{y_{r,s}, s = 1, 2, \ldots, S\}$ with associated probabilities $p_{r,s}$ assumed to satisfy

$$\sum_{i=1}^{s} p_{r,i} = \Psi(\beta_s - \xi_r)$$

for some prescribed distribution function Ψ, where $\xi_r = \gamma(t_r)$ and t_r is the value of the explanatory variable for this response. This grouped continuous model is equivalent to the assumption of a latent continuous variable with distribution function $\Psi(. - \xi_r)$ which is categorized into S classes at the unknown cutpoints $\{\beta_1, \beta_2, \ldots, \beta_{S-1}\}$ to yield the observed frequencies $\{y_{r,s}\}$. See McCullagh (1980) for a complete discussion. This falls into our present framework if we take θ as $\{\theta_{r,s} = \beta_s - \xi_r\}$, so that $p_{r,s} = \Psi(\theta_{r,s}) - \Psi(\theta_{r,s-1})$.

The matrix A is no longer diagonal, but D and E have a very simple form. For identifiability, one component of ξ must be held fixed and omitted from equations (2.4).

4 Choice of roughness penalty

Various authors have remarked in different contexts that choice of the amount of smoothing, λ in equation (2.2), is more important than the form of the smoothing kernel K itself. This might be amended by adding that the null space of the roughness, the space $\{\xi : \xi^T K \xi = 0\}$ may also be important; for vectors in this null space, since they are not penalized in (2.2), are implicitly also fitted as covariates.

Advocates of spline smoothing would choose a penalty of the form

$$J(\gamma) = \int_a^b (\gamma^{(m)}(t))^2 \, dt \tag{4.1}$$

for a curve on a single-dimensional variable. As mentioned in § 2 this is an example of our approach; the kernel K is given by

$$K_{ij} = \int_a^b \varphi_i^{(m)}(t) \varphi_j^{(m)}(t) \, dt \quad (i, j = 1, 2, \ldots, q).$$

The rank of K will be $q - m$ for any spline basis, the null space of K consisting of those ξ for which $\sum \xi_j \varphi_j$, where the sum is over $j = 1, \ldots, q$, is a polynomial of degree $(m - 1)$ or less.

Wahba (1978) derives spline smoothing from a Bayesian model in which an appropriate prior, which is partially improper, is constructed on a space of smooth functions; see also Silverman (1985b). In the notation above, the prior is a multivariate normal distribution for ξ with mean 0 and inverse variance matrix λK. Impropriety of the prior is equivalent to rank deficiency in K. In our present partially parametric context, with an uninformative prior for β as an additional ingredient, the maximum penalized likelihood estimate for (β, ξ) is the mode of the corresponding posterior distribution. We can make the prior more explicit, whilst avoiding impropriety, as follows. Let L and T be as constructed in § 2; then we can generate the prior for ξ as

$$\xi = T\delta + L^T (LL^T)^{-1} \varepsilon, \tag{4.2}$$

where δ is a fixed $(q - r)$-vector and ε an r-vector of zero mean, uncorrelated normally distributed random variables with variance λ^{-1}. We can see that the penalty term $\lambda \xi^T K \xi$ is indeed twice the negative of the appropriate log likelihood term.

Leonard (1982) provides a more completely Bayesian approach for the nonparametric case, again using a Gaussian process as a prior for γ: specifically he recommends an Ornstein–Uhlenbeck process, with two hyperparameters in place of λ, for the difference between the derivative of γ and a prescribed or estimated base curve. The full empirical Bayesian approach allows estimation of the hyperparameters.

If the observations are located at equally-spaced $\{t_i\}$ on a line, the squared mth derivative penalty (4.1) will in practice be indistinguishable from that involving mth differences, for example (3.1), with a basis such that $\varphi_j(t_i) = \delta_{ij}$. Use of other roughness penalties was also considered by Green et al. (1985).

When, and only when, the log likelihood L is that of a normal distribution with expectation θ linear in ξ, addition of the roughness penalty corresponds to use of a 'random effects' model for ξ, or equivalently, as far as β is concerned, to modification of the assumed variance structure for y (Green, 1985).

Whatever form of K is chosen, the tuning constant λ controls the relative impact of roughness, as judged by K, and error, determined by the likelihood. The extreme cases $\lambda \to \infty$ and 0, between which we wish to compromise, can be interpreted as follows. As $\lambda \to \infty$, the likelihood is maximized subject to a roughness penalty of 0; that is we restrict to the purely parametric model $L(\theta(\beta, T\delta))$. As $\lambda \to 0$, the problem degenerates to the minimization of $\xi^T K \xi$ subject to the 'interpolation' condition that $\theta(\beta, \xi)$ maximizes $L(\theta(\beta, \xi))$; whether this is a constraint on ξ depends on the form of the regression function θ.

5 Asymptotic theory

A rigorous asymptotic theory for the general class of regression models considered here is not yet available. From the home ground of the normal linear model we have simultaneously relaxed five assumptions: normality, the exponential family, independence, the linear parameterization and the purely parametric nature of the model. As we have seen, these five relaxations do not disturb the conceptual and pedagogical unity of these models and adds little complication to computation, but they are a considerable barrier to mathematical tractability. The practical statistician must therefore tread warily in applying the heuristics and 'rules of thumb' suggested here, and the mathematical statistician may find some fruitful problems to be tackled. Much of this section is therefore speculative.

For the purely parametric regression model, the general results of McCullagh (1983) show that the other relaxations mentioned above do not prohibit a familiar-looking asymptotic theory. Working in the framework of quasi-likelihood, we obtain asymptotic normality for parameter estimates, and the associated results for score statistics and likelihood ratios, under the main condition that $D^T A D$ has determinant diverging to ∞ but a nonsingular limit when properly normalized. Thus as expected, for example, the same asymptotic theory applies in logistic regression, example (b) in § 3 without the nonparametric component, whether n or the binomial denominators $\{m_i\}$ tend to infinity. This distinction is explored further by Jørgensen (1986) with his 'small-dispersion asymptotics'. It might be expected that this type of result could be extended to semi-parametric models in a framework where the space of basis functions remains stable, and in particular q is fixed.

However such a framework will not usually be adequate. For example, in discussing nonparametric smoothing the usual situation is that the observed $\{t_i\}$ become increasingly dense in a finite interval as $q = n \to \infty$ (Craven & Wahba, 1979; Cox, 1984). Results have been extended to nonnormal models (O'Sullivan, 1983; Cox & O'Sullivan, 1983).

For a more specific discussion, let us first consider the more straightforward linear normal case, and then indicate the way in which moving to the general case will complicate matters.

(a) *The linear normal case.* Suppose that the matrices A, D, E and K do not depend on β or ξ. Then the log likelihood is quadratic, and we are dealing with the linear normal model:

$$y \sim N(D\beta + E\xi, A^{-1}).$$

The working response Y is identically y so the solution to the equations (2.4) does not depend on the initial estimates, and the maximum penalized likelihood estimates $\hat{\beta}$ and $\hat{\xi}$ are obtained without iteration. Concentrating on the parametric component, standard manipulations of partitioned matrices then reveal that

$$\hat{\beta} = (D^T A(I - S)D)^{-1} D^T A(I - S)Y, \tag{5.1}$$

where

$$S = E(E^{\mathrm{T}}AE + \lambda K)^{-1}E^{\mathrm{T}}A, \tag{5.2}$$

and so $\hat{\beta}$ is exactly normally distributed with

$$E(\hat{\beta}) = \beta + (D^{\mathrm{T}}A(I-S)D)^{-1}D^{\mathrm{T}}A(I-S)E\xi, \tag{5.3}$$

$$\mathrm{var}\,(\hat{\beta}) = (D^{\mathrm{T}}A(I-S)D)^{-1}D^{\mathrm{T}}A(I-S)^2D(D^{\mathrm{T}}A(I-S)D)^{-1}. \tag{5.4}$$

Note that in general $\hat{\beta}$ is biased.

Informally, the aim from this point will be to choose λ so that, for suitably 'smooth' ξ, the bias is negligible compared with the standard errors implied by (5.4): this must hold in a suitable asymptotic framework in which the form of A, D, E and K and usually even the dimension q will depend on n.

However, to make progress rather more specialization is needed. As an example, consider one-dimensional spline smoothing. The relevant model is

$$y \sim N(D\beta + \gamma(t),\ \sigma^2 I),$$

where $\gamma(t)$ denotes the vector with ith component $\gamma(t_i)$, and the roughness penalty assumed will be

$$J(\gamma) = \int (\gamma^{(m)}(u))^2\, du.$$

An appropriate basis for theoretical study is that due to Demmler & Reinsch (1975) who demonstrated the existence of a basis $\{\varphi_k,\ k = 1, 2, \ldots, q\}$, where $q = n$, such that $E = \sqrt{n}$ times an orthogonal matrix U, and K is diagonal with $K_{ii} = \nu_i$, say, depending also on n, and ranked so that $0 = \nu_1 = \ldots = \nu_{m-2} \leqslant \nu_{m-1} \ldots \leqslant \nu_n$. Then

$$S = E(E^{\mathrm{T}}AE + \lambda K)^{-1}E^{\mathrm{T}}A = U\,\mathrm{diag}\left[\frac{n\sigma^{-2}}{n\sigma^{-2} + \lambda\nu_i}\right]U^{\mathrm{T}}.$$

This explicit expansion, together with estimates of the eigenvalues $\{\nu_i\}$ due to Speckman (1982), enables a study of the bias and variance in (5.3) and (5.4). This is the approach used by Rice (1986) in a study of a simple example where D has $p = 1$ column that is constructed to have a nontrivial regression on t. With a degree of smoothing asymptotically equivalent to that chosen by typical automatic methods (see § 7), he demonstrates that, while the standard errors of $\hat{\beta}$ are of order $n^{-\frac{1}{2}}$, indeed $\mathrm{var}\,(\hat{\beta})$ is asymptotically $(D^{\mathrm{T}}AD)^{-1}$ as in the parametric case, the bias term (5.3) will in general be of larger order. In fact this is intuitively unsurprising. We assume that $(D\ ET)$ is of full rank, not that this is true of $(D\ E)$: indeed with $q = n$ basis functions the latter could not be true. Thus the decomposition $D\beta + E\xi$ is ambiguous. Given the true model $D\beta_0 + E\xi_0$, dissection of Rice's proof reveals conventional $\sqrt{n}$-consistency for some β such that $D\beta + E\xi = D\beta_0 + E\xi_0$: but in general there is no reason for this to be the *correct* β, as the two parts of the model are confounded. Undersmoothing would be necessary for this problem of bias to disappear.

In a similar framework, but with general p and the rows of D chosen independently and identically distributed so that they are not correlated with t, Heckman (1986) indeed obtains the familiar result.

(b) *The general case.* The special case just considered is not as narrow as it may seem, since we know that in many standard statistical models, the matrices A, D and E actually vary rather slowly with β and ξ: that is why the asymptotic theory is true so generally (McCullagh, 1983) and the method of scoring so widely applicable (Green, 1984).

However, technical difficulties are certainly introduced, and the author is not aware of published asymptotic theory for any nonlinear semi-parametric models.

Because of the curvature in such models, even consistency is not certain, and indeed establishing consistency will be the crucial step. Thereafter careful approximations should allow the transfer of results for the linear normal case to give the same asymptotic distribution $\hat{\beta} \sim N(\tilde{\beta}_0, (D^{\mathrm{T}}AD)^{-1})$ for some $\tilde{\beta}_0$ related to β_0 but giving sufficiently small 'asymptotic confounding' between the parametric and nonparametric components of the model. The normality will be obtained using a central-limit type argument under a condition on $(D^{\mathrm{T}}AD)$ such as that quoted earlier.

Apart from theoretical study, there is scope for empirical work in assessing the value of these asymptotics and their utility with finite samples, although it would be difficult to design such studies to generate very widely applicable conclusions. In practice in the meantime, we propose to use (5.3) and (5.4), even with data-dependent choice of smoothing parameter, but to be wary of any formal inference based on these. It may be useful to assess the risk from the bias problem using diagnostics that examine the association between $\gamma(t)$ and columns of D, evaluated at the maximum penalized likelihood estimates.

6 Deviance and degrees of freedom

In generalized linear models, the deviance with its associated degrees of freedom provides a goodness-of-fit statistic, to be referred to the appropriate χ^2 distribution; differences in deviances and degrees of freedom give tests of model adequacy (Nelder & Wedderburn, 1972). Such tests are exact only in the normal linear model with known variances. The practice of performing such tests will not be rehearsed here, but we shall attempt to derive appropriate definitions of deviance and degrees of freedom for the general semi-parametric regression model.

We need the notion of 'saturated model' to be well defined. Suppose that, as a function of θ freed from its functional dependence on β and ξ, L is uniquely maximized at θ'. Then the deviance associated with a particular regression function $\theta(\beta, \xi)$ is

$$\Delta = 2\{L(\theta') - L(\theta(\hat{\beta}, \hat{\xi}))\}. \tag{6.1}$$

The maximum penalized likelihood estimates $\hat{\beta}$ and $\hat{\xi}$ minimize not (6.1) but the penalized deviance $\Delta + \lambda\hat{\xi}^{\mathrm{T}}K\hat{\xi}$. How many degrees of freedom should we associate with Δ? In parametric models, of course, the answer is the null asymptotic expectation.

As in § 5, the first-order approximation is that given by the linear normal case, where A, D, E and K are constant. Then

$$\Delta = \Delta_1 \equiv \hat{u}^{\mathrm{T}}A^{-1}\hat{u}, \tag{6.2}$$

where $\hat{u} = u(\theta(\hat{\beta}, \hat{\xi}))$; here $\hat{u} = A(\theta' - \theta(\hat{\beta}, \hat{\xi})) = A(Y - D\hat{\beta} - E\hat{\xi}))$. In general, (6.2) is only an approximation: we call Δ_1 the linearized deviance. In generalized linear models, it is essentially Pearson's χ^2 statistic; for a discussion of its asymptotic distribution in this case, see McCullagh (1985). Let B be a square root of $A = BB^{\mathrm{T}}$, and define the $n \times n$ matrix

$$M = \left\{I - B^{\mathrm{T}}[D\ E]G^{-1}\begin{bmatrix} D^{\mathrm{T}} \\ E^{\mathrm{T}} \end{bmatrix}B\right\}. \tag{6.3}$$

Simple manipulations using (2.5) establish that

$$\Delta_1 = Y^{\mathrm{T}}BM^2B^{\mathrm{T}}Y, \quad \lambda\hat{\xi}^{\mathrm{T}}K\hat{\xi} = Y^{\mathrm{T}}B(M - M^2)B^{\mathrm{T}}Y.$$

In the present case, $Y \sim N(D\beta_0 + E\xi_0, A^{-1})$ so we can calculate the expectations of these

quadratic forms. Since

$$\begin{bmatrix} D^{\mathrm{T}} \\ E^{\mathrm{T}} \end{bmatrix} BM^{r}B^{\mathrm{T}}[D\ E] = (G-J)(G^{-1}J)^{r}$$

for $r = 0, 1, 2, \ldots$, where

$$J = \begin{bmatrix} 0 & 0 \\ 0 & \lambda K \end{bmatrix},$$

we find that

$$E(\Delta_1) = \mathrm{tr}\,(M^2) + Q_1 - Q_2, \quad E(\Delta_1 + \lambda\hat{\xi}^{\mathrm{T}}K\hat{\xi}) = \mathrm{tr}\,(M) + Q_0 - Q_1,$$

where

$$Q_r = \xi_0^{\mathrm{T}}\lambda K\{[E^{\mathrm{T}}AE + \lambda K - E^{\mathrm{T}}AD(D^{\mathrm{T}}AD)^{-1}D^{\mathrm{T}}AE]^{-1}\lambda K\}^{r}\xi_0.$$

The value of these expressions is limited by the presence of the correction terms, quadratic forms in the unknown true ξ_0. Note that $0 \leqslant Q_{r+1} \leqslant Q_r$, with mutual equality if and only if $K\xi_0 = 0$. Such quadratic forms are inherently difficult to estimate so we are left with two alternatives:

(a) neglecting these terms, leading to an over-estimate of degrees-of-freedom and thence conservative goodness-of-fit tests, or
(b) replacing them by their expectations under the prior distribution (4.2).

Using this latter approach, further matrix manipulations give $E(Q_j) = t_j$, say, where $t_0 = r$, and $t_j = \mathrm{tr}\,(M^j) + p + q - n$ for $j = 1, 2, 3, \ldots$. Denoting $\mathrm{tr}\,(M)$ by ν, we finally have the estimates

$$E(\Delta_1) = \nu, \quad E(\Delta_1 + \lambda\hat{\xi}^{T}K\hat{\xi}) = n - (p+q) + r.$$

It may be shown that for any $\lambda > 0$, ν satisfies the inequality

$$n - \mathrm{rank}\,[D\ E] \leqslant \nu \leqslant n - (p+q) + r.$$

Its exact form as a function of λ is given by Green (1985). In the notation of that paper, R is ET and V is

$$I + \lambda^{-1}E(E^{\mathrm{T}}E)^{-1}L^{\mathrm{T}}(L(E^{\mathrm{T}}E)^{-1}L^{\mathrm{T}})^{-2}L(E^{\mathrm{T}}E)^{-1}E^{\mathrm{T}}.$$

Combining this with the information above about expectations supports the use of ν as a surrogate for degrees of freedom for Δ. Parameters corresponding to the columns of $[D\ ET]$, that is β and δ, are always fitted, requiring $(p+q-r)$ degrees of freedom, and the remaining $n - (p+q) + r - \nu$ are associated with the nonparametric component of ξ permitted by $\lambda < \infty$. Even in this linear normal case, there is no tractable distribution theory for Δ: for discussion of χ^2 approximations to the distribution of error sums-of-squares in nonparametric regression, see Buckley & Eagleson (1986). Further we have neglected the effects of a data-dependent choice of λ.

When we turn to the general nonlinear or nonnormal case, the additional difficulties are similar to those of § 5.

(i) The deviance is not identical to its linearized approximation.
(ii) The dependence of the estimated values of A, D, E and possibly K on the data has been neglected in the expectation calculations.

Notwithstanding these difficulties our tentative recommendation is to use the deviance Δ (without the addition of the penalty term) with degrees of freedom ν in the analysis of

deviance just as for generalized linear models (Nelder & Wedderburn, 1972). Calculation of ν can proceed directly by finding the trace of (6.3), or more economically using the easily-proved identity

$$\nu = n - \mathrm{tr}\,(S) - \mathrm{tr}\,\{(D^{\mathrm{T}}A(I-S)D)^{-1}D^{\mathrm{T}}A(I-S)^2D\},$$

where S is given by (5.2).

The remarks at the end of § 5 advocating further theoretical work are equally applicable here.

7 Cross-validation

In practice, some automatic data-dependent choice of the tuning constant λ will often be required. Model selection by means of cross-validation was discussed in a systematic way by Stone (1974). Its use in determining an appropriate degree of smoothing in nonparametric regression problems has been enthusiastically espoused by Wahba and co-workers in the past ten years and, in a refined form, known as generalized cross-validation (Wahba, 1977), seems to have become the de facto standard approach. In the linear problem, generalized cross-validation has additional invariance over the ordinary version, it has now become computationally practicable, and it is known to provide an asymptotically optimal degree of smoothing in a predictive mean-square sense. O'Sullivan (1983) generalizes the application of generalized cross-validation to generalized linear models by transcribing a formula from the linear case, without deriving the criterion afresh from its plausible first principles. This we attempt to do here, for our more general class of regression problems.

The basic idea in cross-validation is to delete one observation at a time from the data set, and endeavour to predict it from the model as fitted to the remaining observations. The smoothing parameter is chosen to optimize the overall quality of prediction. The appropriate generalization of this 'delete-one' operation in our model $L(\theta(\beta, \xi))$ consists of *decoupling* each component of θ in turn from its dependence on β and ξ. The predictive discrepancy will be measured in likelihood or deviance terms.

The decoupling is achieved by the introduction of dummy covariates. For some generality, let F be an arbitrary $n \times f$ matrix ($f \geqslant 1$), and let $\tilde{\beta}$, $\tilde{\xi}$ and $\tilde{\tau}$ maximize the penalized decoupled likelihood $L(\theta(\tilde{\beta}, \tilde{\xi}) + F\tilde{\tau}) - \frac{1}{2}\lambda\tilde{\xi}^{\mathrm{T}}K\tilde{\xi}$. We define the predictive discrepancy in the column space of F as the nonnegative quantity

$$\Delta^{\dagger}(F) = 2\{L(\theta(\tilde{\beta}, \tilde{\xi}) + F\tilde{\tau}) - L(\theta(\tilde{\beta}, \tilde{\xi}))\}; \tag{7.1}$$

if this is zero then the decoupled estimates $(\tilde{\beta}, \tilde{\xi})$ coincide with $(\hat{\beta}, \hat{\xi})$. We will average $\Delta^{\dagger}(F)$ over an appropriate set of directions to give an overall predictive discrepancy, but first we obtain a quadratic approximation for $\Delta^{\dagger}(F)$.

At $(\tilde{\beta}, \tilde{\xi}, \tilde{\tau})$ we have $D^{\mathrm{T}}\tilde{u} = 0$, $E^{\mathrm{T}}\tilde{u} = \lambda K\tilde{\xi}$ and $F^{\mathrm{T}}\tilde{u} = 0$, where

$$\tilde{u} = u(\theta(\tilde{\beta}, \tilde{\xi}) + F\tilde{\tau}) \doteq u(\theta(\hat{\beta}, \hat{\xi})) - A(D(\tilde{\beta} - \hat{\beta}) + E(\tilde{\xi} - \hat{\xi}) + F\tilde{\tau}).$$

But we know that $D^{\mathrm{T}}\hat{u} = 0$ and $E^{\mathrm{T}}\hat{u} = \lambda K\hat{\xi}$, so by subtraction, treating A, D and E as fixed, evaluated at $(\hat{\beta}, \hat{\xi})$, say, we have

$$\begin{bmatrix} D^{\mathrm{T}}AD & D^{\mathrm{T}}AE & D^{\mathrm{T}}AF \\ E^{\mathrm{T}}AD & E^{\mathrm{T}}AE + \lambda K & E^{\mathrm{T}}AF \\ F^{\mathrm{T}}AD & F^{\mathrm{T}}AE & F^{\mathrm{T}}AF \end{bmatrix} \begin{bmatrix} \tilde{\beta} - \hat{\beta} \\ \tilde{\xi} - \hat{\xi} \\ \tilde{\tau} \end{bmatrix} \doteq \begin{bmatrix} 0 \\ 0 \\ F^{\mathrm{T}}\hat{u} \end{bmatrix},$$

whence $\tilde{\beta}$ and $\tilde{\xi}$ may be eliminated to give

$$\tilde{\tau} \doteq \left(F^{\mathrm{T}}AF - F^{\mathrm{T}}A(D\,E)G^{-1}\begin{bmatrix} D^{\mathrm{T}} \\ E^{\mathrm{T}} \end{bmatrix} AF\right)^{-1} F^{\mathrm{T}}\hat{u} = (F^{\mathrm{T}}BMB^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\hat{u}.$$

So, by linearizing the expression (7.1) and noting that $(\tilde{\beta}, \tilde{\xi}, \tilde{\tau})$ maximizes the first term, penalized, we have

$$\Delta^{\dagger}(F) \doteq (F\tilde{\tau})^{\mathrm{T}}A(F\tilde{\tau}) = \hat{u}^{\mathrm{T}}F(F^{\mathrm{T}}BMB^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}AF(F^{\mathrm{T}}BMB^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}\hat{u}. \tag{7.2}$$

This expression measures the result of decoupling any number f of components of θ; for an analogue of delete-one cross-validation we set $f = 1$. For example if $F = e^{(i)}$, the unit vector in the ith coordinate direction,

$$\Delta^{\dagger}(e^{(i)}) \doteq \frac{A_{ii}\hat{u}_i^2}{\{(BMB^{\mathrm{T}})_{ii}\}^2}.$$

If A is not diagonal, we may prefer to orthogonalize the predictor space and use

$$\Delta^{\dagger}((B^{-1})^{\mathrm{T}}e^{(i)}) \doteq (B^{-1}\hat{u})_i^2/M_{ii}^2.$$

In generalized cross-validation the individual predictive discrepancies are combined over different directions by a weighted sum enjoying certain invariance properties. Let $w_i = M_{ii}/\mathrm{tr}\,(M)$, so that $\sum w_i = 1$, and define the generalized cross-validation criterion

$$V(\lambda) = \sum_{i=1}^{n} w_i^2\Delta^{\dagger}((B^{-1})^{\mathrm{T}}e^{(i)}) \doteq (u^{\mathrm{T}}A^{-1}u)/\mathrm{tr}\,(M)^2 \doteq \Delta/\nu^2.$$

We can choose λ to minimize this quantity, which has the same form as that used by O'Sullivan (1983).

That this is the correct weighting of the individual predictive discrepancies can be seen by examining the invariance properties of $V(\lambda)$. Re-parameterization by invertible appropriately differentiable transformations of θ, β and ξ does not change the model; it alters u, A, D and E, but $\hat{u}^{\mathrm{T}}A^{-1}\hat{u}$, Δ, M and ν remain invariant, and so therefore does $V(\lambda)$.

One justification for use of the generalized cross-validation criterion in the linear (spline) case is provided by the result of Craven & Wahba (1979) stating that such a criterion is asymptotically optimal in the sense of minimizing the mean squared error

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}(t_i) - \gamma_0(t_i))^2.$$

Of course, this property may be shared by many other criteria for choosing λ. The only natural expression of $R(\lambda)$ in likelihood terms seems to be via the divergence defined by Kullback & Leibler (1951). We define $R(\lambda)$ so that

$$nR(\lambda) = 2E_{\theta_0}(L(\theta_0) - L(\hat{\theta})) \doteq (\hat{u} - u_0)^{\mathrm{T}}A^{-1}(\hat{u} - u_0)$$

to give an appropriate weighted mean squared error for $(\hat{\beta}, \hat{\xi})$ which makes connections with the linearized deviance apparent. Cross-validation and Kullback–Leibler distance are also discussed by Bowman, Hall & Titterington (1984). O'Sullivan (1983) sketches an argument suggesting that the Craven & Wahba result extends to the generalized linear model case, with a definition of $R(\lambda)$ equivalent to the above; it therefore seems likely that if his proof could be rigorized, it might apply to the present more general set-up as well.

Note that by arguments similar to those of § 6, we have

$$E(R(\lambda)) \doteq Q_1 - Q_2 + \operatorname{tr}((I-M)^2)$$

whose expectation under the prior for ξ_0 is just $\operatorname{tr}(I-M)$.

8 Residuals

How best to define residuals depends very much on the purpose to which they are to be put. The multitude of definitions available even in simple linear regression models (Cook & Weisberg, 1982) strongly suggests that even more alternatives will be available in our present general context. Here we attempt only a limited discussion. We seek residuals primarily for diagnostic purposes, and, in view of our reliance on the likelihood function $L(\theta(\beta, \xi))$ prefer these to be likelihood based and associated with the predictors θ rather than directly with the observations. Use of such residuals for diagnosis of data inadequacy will require inspection of the likelihood function to determine the data points instrumental in giving a particular component of θ a large residual. Detection of model inadequacy can proceed more directly, and note that we do not desire invariance of residuals to transformation of θ itself.

The likelihood emphasis suggests concentrating on the deviance Δ, defined in § 6. Restricting attention to one or more particular components of θ, define

$$\Delta(F) = 2\left\{\sup_{\tau} L(\theta(\hat{\beta}, \hat{\xi}) + F\tau) - L(\theta(\hat{\beta}, \hat{\xi}))\right\}, \tag{8.1}$$

twice the maximum increase in log likelihood attained by freeing θ from its dependence on β and ξ in the directions spanned by columns of F. If F is nonsingular, $\Delta(F) = \Delta$. Note that $\Delta(F) \leqslant \Delta^{\dagger}(F)$; the sole difference between the two quantities lies in the inclusion or exclusion of the corresponding components of θ from the *fitting* of the model. Also note that the maximum penalized likelihood ratio statistic lies between $\Delta(F)$ and $\Delta^{\dagger}(F)$. Choosing a single coordinate direction $e^{(i)}$ for F, we obtain the raw deviance and discrepancy $\Delta(e^{(i)})$ and $\Delta^{\dagger}(e^{(i)})$ respectively, which we abbreviate as Δ_i and $\Delta_i^{\dagger}$. The raw deviances have been customarily used to define residuals in generalized linear models; see discussion of paper by Green (1984). Finally, we denote the signed square roots by $z_i = \operatorname{sign}(\tau_{\max})\sqrt{\Delta_i}$ and $z_i^{\dagger} = \operatorname{sign}(\tau_{\max})\sqrt{\Delta_i^{\dagger}}$, where $\tau_{\max}$ denotes the value of τ in the maximization of (8.1) and in (7.1) respectively.

These concepts tie in well with other treatments of residuals. In the case of normal linear regression (with known variances = 1, say, for simplicity), z_i and $z_i^{\dagger}$ are just the ordinary and predicted residuals, respectively, of Cook & Weisberg (1982, Ch. 2). These are known to be correlated, and improperly standardized for variance. Cook & Weisberg point out that z_i and $z_i^{\dagger}$ respectively under- and over-emphasize discrepancies for high-leverage data points. This difficulty will persist in the more general cases. There is a very detailed treatment of various definitions of residuals for generalized linear models in the paper by Pierce & Schafer (1986), which advocates the use of deviance-based residuals for most purposes.

When y is distributed normally with expectation $\theta = D\beta$ and known nondiagonal variance matrix V we find

$$z_i = \{V^{-1}(y-\theta)\}_i \{(V^{-1})_{ii}\}^{-\frac{1}{2}},$$

$$z_i^{\dagger} = \{V^{-1}(y-\theta)\}_i \{(V^{-1})_{ii}\}^{\frac{1}{2}} \{(V^{-1} - V^{-1}D(D^{\mathrm{T}}D)^{-1}D^{\mathrm{T}}V^{-1})_{ii}\}^{-1},$$

which are in fact y_i standardized by its expectation and variance assuming the parameters

to be equal to their estimates with and without y_i, respectively, *conditional* on the other components of y.

In contrast to Jørgensen (1984), we believe that the use of unconditional moments in this standardization is inappropriate for dependent observations.

For these examples, the linearization leading to (7.2) and by a simpler argument to $\Delta(F) = \hat{u}^{\mathrm{T}}F(F^{\mathrm{T}}AF)^{-1}F^{\mathrm{T}}\hat{u}$ involves no approximation. In general, when the likelihood is not quadratic, replacing Δ_i and $\Delta_i^{\dagger}$ by these approximations leads to different residuals z_i and $z_i^{\dagger}$. The former have been called 'score residuals' (Jørgensen, 1983). But as pointed out by Green (1984), these score residuals are not appropriate when the quadratic approximation is badly wrong: for example, they are not monotonic functions of the observations in linear regression with a prescribed error density that is not log concave.

9 Nuisance parameters

The approach to penalized likelihood estimation described here can handle with no difficulty certain types of nuisance parameter entering the probability model in addition to the predictors θ. Suppose, following Jørgensen (1983) that $L = L(y; \theta, \kappa) = c(y; \kappa) + \sigma(\kappa)t(y; \theta)$, where σ, which we might term the precision parameter, is a scalar function of the possibly vector-valued nuisance parameter κ. See also Green (1984). This is in a sense the ultimate generalization of the property of generalized linear models in which the scale parameter factors out from the fitting procedure and is estimated at convergence from the deviance. Examples include the variance in the normal distribution, the index in the gamma distribution, and also the extra parameter often allowed as a modification of the binomial or Poisson distributions to allow for 'over-dispersion'.

The maximum penalized likelihood estimates of β and ξ now satisfy $\sigma D^{\mathrm{T}}\hat{u} = 0$, $\sigma E^{\mathrm{T}}\hat{u} = \lambda K\hat{\xi}$. Fisher scoring is no longer available necessarily, because the expectation of $t(y; \theta)$ will in general involve κ but, if we write $A = -\partial^2 t/\partial\theta\theta^{\mathrm{T}}$, then neglecting the second derivatives of θ with respect to β and ξ, the 'linearization method' of Jørgensen (1983), we obtain the approximate Newton–Raphson iteration:

$$\begin{bmatrix} D^{\mathrm{T}}AD & D^{\mathrm{T}}AE \\ E^{\mathrm{T}}AD & E^{\mathrm{T}}AE + \sigma^{-1}\lambda K \end{bmatrix} \begin{bmatrix} \beta^* \\ \xi^* \end{bmatrix} = (D\ E)^{\mathrm{T}}AY,$$

demonstrating that β and ξ can be estimated without paying any attention to the nuisance parameter κ except that the value of the tuning constant λ is now effectively measured with respect to the unknown precision σ, a consequence that is not likely to be of any serious concern.

Acknowledgements

I am indebted to Tom Leonard for useful discussions on a Bayesian interpretation of the decoupling in §§ 7 and 8, and to Brian Yandell for suggestions and subsequent collaboration on computing methods. The referees' comments on an earlier version are also much appreciated.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.* **68,** 337–404.

Bowman, A.W., Hall, P. & Titterington, D.M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika* **71,** 341–351.

Buckley, M.J. & Eagleson, G.K. (1986). Distribution theory for local estimates of variance. Unpublished manuscript.

Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Cox, D.D. (1984). Multivariate smoothing spline functions. *SIAM J. Numer. Anal.* **21,** 789–813.

Cox, D.D. & O'Sullivan, F. (1983). Asymptotic analysis of the roots of penalised likelihood equations. Technical Report, Dept. of Statistics, University of Wisconsin–Madison.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31,** 377–403.
Demmler, A. & Reinsch, C. (1975). Oscillation matrices with spline smoothing. *Numer. Math.* **24,** 375–382.
Engle, R.F., Granger, C.W.J., Rice, J. & Weiss, A. (1986). Non-parametric estimates of the relation between weather and electricity demand. *J. Am. Statist. Assoc.* **81,** 310–320.
Good, I.J. & Gaskins, R.A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58,** 255–277.
Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *J. R. Statist. Soc.* B **46,** 149–192.
Green, P.J. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika* **72,** 527–537.
Green, P.J., Jennison, C. & Seheult, A.H. (1983). Discussion of paper by G.N. Wilkinson et al., *J. R. Statist. Soc.* B **45,** 193–195.
Green, P.J., Jennison, C. & Seheult, A.H. (1985). Analysis of field experiments by least squares smoothing. *J. R. Statist. Soc.* B **47,** 299–315.
Green, P.J. & Yandell, B.S. (1985). Semi-parametric generalized linear models. In *Lecture Notes in Statistics,* **32,** pp. 44–55. Berlin: Springer.
Heckman, N.E. (1986). Spline smoothing in a partly linear model. *J. R. Statist. Soc.* B **48,** 244–248.
Jørgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70,** 19–28.
Jørgensen, B. (1984). Discussion of paper by P.J. Green. *J. R. Statist. Soc.* B **46,** 171–172.
Jørgensen, B. (1986). Small dispersion asymptotics. Odense University preprint 1986/5.
Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *Ann. Statist.* **22,** 79–86.
Leonard, T. (1982). An empirical Bayesian approach to the smooth estimation of unknown functions. Tech. Summary Report 2339, MRC, University of Wisconsin–Madison.
McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B **42,** 109–142.
McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11,** 59–67.
McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models. *Int. Statist. Rev.* **53,** 61–68.
Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A **135,** 370–384.
O'Sullivan, F. (1983). The analysis of some penalized likelihood schemes. Ph.D. Thesis, Technical Report No. 726, Department of Statistics, University of Wisconsin–Madison.
O'Sullivan, F., Yandell, B.S. & Raynor, W.J. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Am. Statist. Assoc.* **81,** 96–103.
Pierce, D.A. & Schafer, D.W. (1986). Residuals in generalized linear models. *J. Am. Statist. Assoc.* **81,** 977–986.
Rice, J. (1986). Convergence rates for partially splined models. *Statist. Prob. Letters* **4,** 203–208.
Silverman, B.W. (1985a). Penalized maximum likelihood estimation. In *Encyclopedia of Statistical Sciences,* **6,** Ed. S. Kotz and N.L. Johnson, pp. 664–667. New York: Wiley.
Silverman, B.W. (1985b). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc.* B **47,** 1–52.
Speckman, P. (1982). The asymptotic integrated mean square error for smoothing noisy data by splines. Unpublished manuscript.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B **36,** 111–147.
Wahba, G. (1977). A survey of some smoothing problems, and the method of generalized cross-validation for solving them. In *Applications of Statistics,* Ed. P.R. Krishnaiah, pp. 507–523. Amsterdam: North-Holland.
Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc.* B **40,** 364–372.
Wahba, G. (1984). Partial spline models for the semi-parametric estimation of functions of several variables. In *Statistical Analysis of Time Series,* pp. 319–329. Tokyo: Institute of Statistical Mathematics.
Wahba, G. (1985). Discussion of paper by P.J. Huber. *Ann. Statist.* **13,** 518–521.
Yandell, B.S. & Green, P.J. (1986). Who owns phones: diagnosis using semi-parametric generalized linear modes. Unpublished report.

Résumé

On examine ici l'estimation par la vraisemblance pénalisée dans le contexte des problèmes généraux de régression, caractérisés comme des modèles avec des fonctions composites de vraisemblance. On accentue la situation fréquente quand on trouve un modèle paramétrique comme utile sauf pour la nonhomogénéité à l'égard de quelques variables supplémentaires. Une formulation de dimension finie est adoptée avec une base convenable de fonctions. Des définitions appropriées de la déviation, des dégrés de liberté, et de résidu sont examinées, et la méthode de validation croisée pour un choix du paramètre d'ajustement est discutée. Des approximations quadratiques sont présentées pour toutes les statistiques nécessaires.

[*Received May* 1985, *revised June* 1987]