[Final Draft: 11/24/87]

A NEW ALGORITHM FOR MATCHED CASE-CONTROL STUDIES WITH APPLICATIONS TO ADDITIVE MODELS

Trevor Hastie

Daryl Pregibon

AT&T Bell Laboratories

Murray Hill, New Jersey 07974

Abstract

Logistic models are commonly used to analyze matched case-control data. The standard analysis requires the computation of conditional maximum likelihood estimates. The Newton-Raphson method is quite effective in this regard though the computations do not resemble iterative, diagonally weighted, least-squares. Whitehead(1980) proposed an unconditional poisson analysis which leads to correct estimates of the odds ratios and their standard errors by using the stratum indicator as a factor in the linear predictor. We propose a different and even simpler procedure that uses a diagonal approximation for the (non-diagonal) weight matrix of the conditional algorithm. As such it is in the class of 'delta' algorithms as described by Jørgenson (1984).

The primary purpose of the new algorithms is to exploit iterative weighted least-squares procedures for fitting general additive structure rather than simple linear structure. Thus, writing the standard model as $\text{logit}(p) = \alpha_k + x_1\beta_1 + x_2\beta_2 + \cdots + x_m\beta_m$, we propose the extension to $\text{logit}(p) = \alpha_k + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$ where α_k accounts for the matching variables and f_j is an arbitrary smooth function of the covariate x_j. We demonstrate the methodology on two sets of data.

An abbreviated version of this paper appeared in the proceedings of Compstat 1988, Physica-Verlag, Heidelberg.

1. Introduction.

The linear logistic regression model has become a standard tool of epidemiologists and biostatisticians for the analysis of matched case-control data. Breslow and Day (1980) give a clear and detailed description. We establish notation which follows that of Pregibon (1984), and give a very brief outline of the methodology.

Typically the response Y measures the presence (1) or absence (0) of some disease. The linear logistic regression model relates the incidence or prevalence of the disease to a set of possible *exposure* variables $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$ via

$$\operatorname{logit}(P(Y = 1 \mid \boldsymbol{x})) = \alpha + x_1\beta_1 + x_2\beta_2 + \cdots + x_m\beta_m \tag{1}$$

where $\operatorname{logit}(p) = \log(p/(1-p))$. An equivalent way of expressing model (1) is given by

$$\log(\pi(\boldsymbol{x})) = x_1\beta_1 + x_2\beta_2 + \cdots + x_m\beta_m \tag{2}$$

where the odds ratio π is defined as the odds of an individual with exposure $\boldsymbol{x}$ developing the disease relative to an individual with baseline exposure $\boldsymbol{0}$. The regression coefficients β_j are estimable regardless of whether the sampling was prospective or retrospective in nature.

Often certain exposure variables are known *a priori* to affect prevalence but of limited interest otherwise. Let M_k denote a partition of the space of these variables. In this case model (1) can be written

$$\operatorname{logit}(P(Y = 1 \mid \boldsymbol{x}, M_k)) = \alpha_k + x_1\beta_1 + x_2\beta_2 + \cdots + x_m\beta_m \tag{3}$$

where α_k is an intercept term specific to M_k. The parameters of interest, β_j, are assumed to be constant across matched sets. In terms of odds ratios we have

$$\log(\pi(\boldsymbol{x}, M_k)) = x_1\beta_1 + x_2\beta_2 + \cdots + x_m\beta_m, \tag{4}$$

which is identical to (2).

If we were free to design the experiment (with no cost constraints), we would construct cohorts of subjects sampled randomly from $\boldsymbol{x} \mid M_k$, which would then be followed up and the disease status Y recorded. With sufficient observations in each set, the nuisance parameters, α_k, and the parameters of interest, β_j, could be estimated by maximum likelihood.

Often such prospective studies are impractical (e.g. for low incidence diseases), expensive, and perhaps even unethical. An alternative is to sample retrospectively from the cases and controls.

For the kth of K cases with covariate vector $\boldsymbol{x}_{0k}$, form a pool of controls having the same values of the matching variables. Randomly select R_k of these matched controls with covariate vectors $\boldsymbol{x}_{rk}$, $r = 1, \ldots, R_k$. For notational simplicity, we assume further that $R_k = R \ \forall k$.

Using (3), the *conditional* probability that within a matched set, the assignment of the $R+1$ values $\boldsymbol{x}_{rk}$ to case and controls is as observed, is given by Breslow and Day (1980) as

$$\mu_{0k} = \frac{\exp(\boldsymbol{x}_{0k}^t \boldsymbol{\beta})}{\sum_{r=0}^{R} \left(\exp(\boldsymbol{x}_{rk}^t \boldsymbol{\beta})\right)}. \tag{5}$$

The full conditional likelihood is simply the product, $L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \mu_{0k}$, over matched sets; we typically work with its logarithm:

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left(\boldsymbol{x}_{0k}^t \boldsymbol{\beta} - \log(\sum_{r=0}^{R} \left(\exp(\boldsymbol{x}_{rk}^t \boldsymbol{\beta})\right) \right) \tag{6}$$

In this paper we discuss algorithms for maximizing (6). Our primary goal is to develop algorithms for estimating non-parametric extensions of (4); in particular we concentrate on estimating the arbitrary but smooth functions f_j in the additive model:

$$\log(\pi(\boldsymbol{x}, M_k)) = f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m). \tag{7}$$

Additive models of this kind are discussed in Hastie and Tibshirani (1986a-c) who motivate the *local scoring* algorithm for estimating the f_j as a non-parametric extension of iterative reweighted least squares (IRLS). They concentrated, however, on situations where the iterative weight matrix is diagonal as in the case of exponential family densities. In this case each observation has an associated (iterative) weight, and the local scoring algorithm estimates the functions by *smoothing* appropriate partial residuals using a weighted scatterplot smoother. Although the Newton-Raphson algorithm for maximizing (6) can be written as IRLS, we will see that the weight matrix is not diagonal. This hampers the non-parametric extension and alternative algorithms are needed.

Most of the ideas developed here carry over directly to the proportional hazards regression model for censored survival data (Cox, 1972). Here one models the hazard function, $\lambda(\boldsymbol{x}, T) = \lambda_0(T) \exp(\boldsymbol{x}^t \boldsymbol{\beta})$, where T is the time to death, $\lambda_0(T)$ is the baseline hazard function, and the multiplier $\exp(\boldsymbol{x}^t \boldsymbol{\beta})$ measures the effect of covariates $\boldsymbol{x}$ on the hazard. The logarithm of the baseline hazard at the kth observed death time, $\log(\lambda_0(T_k))$, plays a role similar to α_k in model

(7). These so-called nuisance parameters are eliminated in the log partial likelihood

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left(\boldsymbol{x}_{0k}^t \boldsymbol{\beta} - \log(\sum_{r \in R_k} \exp(\boldsymbol{x}_{rk}^t \boldsymbol{\beta}) \right), \tag{8}$$

where R_k, the risk set at death time T_k, plays a role similar to a matched set.

2. Motivation.

In this section we introduce an example to motivate the methodology we propose. Our purpose is not to provide a definitive analysis of the data. Rather we show that our methods lead easily and naturally to those models already suggested and also new structure not previously discussed.

Breslow et al (1978) analyze data from a study of oesophageal cancer in Singapore. Each case was matched with four hospital controls based on sex, race, and age (within five years). The data considered here refer to 80 male cases and exposure variables dialect group (DG: coded 0 or 1), consumption of distilled liquor (LIQ: coded 0 or 1), number of beverages drunk at burning hot temperatures (BEV: scored 0,1,2,3), and number of cigarettes smoked daily (SMO: cigarettes/day).

These authors considered the linear model

$$\log(\pi) = \mathrm{DG}\beta_1 + \mathrm{LIQ}\beta_2 + \mathrm{BEV}\beta_3 + \mathrm{SMO}\beta_4, \tag{9}$$

demonstrating how one would make inferences concerning the β's.

Age was used in the matching procedure, but since matching was imperfect one could assess whether an adjustment to the parameters of interest was possible by augmenting (9) with the term AGE. Thus one could entertain the linear model

$$\log(\pi) = \mathrm{DG}\beta_1 + \mathrm{LIQ}\beta_2 + \mathrm{BEV}\beta_3 + \mathrm{SMO}\beta_4 + \mathrm{AGE}\beta_5. \tag{10}$$

Upon fitting (10) one finds that the AGE contribution is not significant. In contrast consider the additive formulation of (10)

$$\log(\pi) = \mathrm{DG}\beta_1 + \mathrm{LIQ}\beta_2 + \mathrm{BEV}\beta_3 + f_{\mathrm{SMO}}(\mathrm{SMO}) + f_{\mathrm{AGE}}(\mathrm{AGE}). \tag{11}$$

Figure 1 displays the non-parametric fitted function for AGE which is non-monotone in nature. The function for SMO (not shown) is reasonably linear and quite flat. A fit of the linear model

$$\log(\pi) = \mathrm{DG}\beta_1 + \mathrm{LIQ}\beta_2 + \mathrm{BEV}\beta_3 + \mathrm{SMO}\beta_4 + (\mathrm{AGE} - 62.5)^2\beta_5 \tag{12}$$

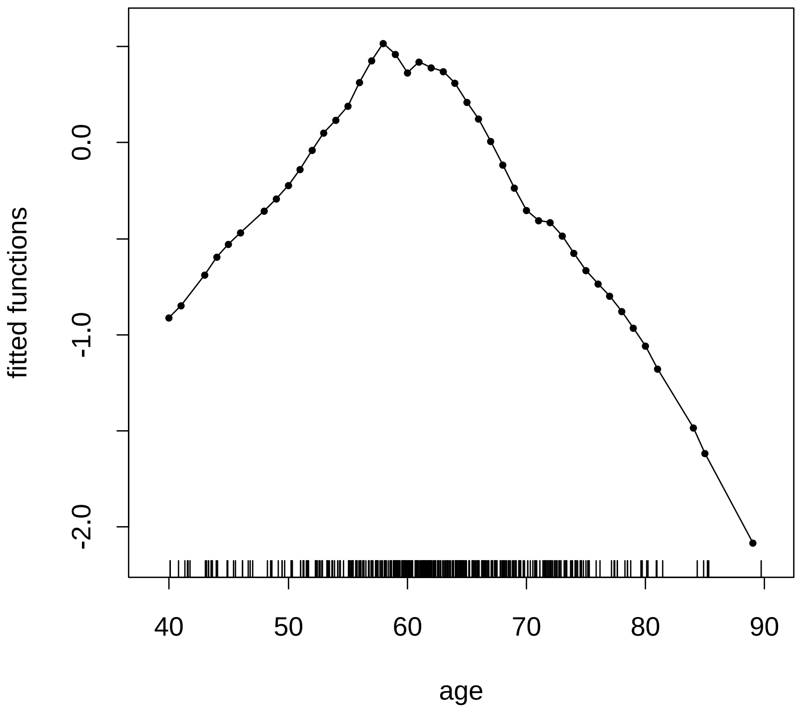

Figure 1. The fitted function for AGE in model (11). The vertical bars at the base of the figure represent frequency of AGE values. The plot suggests a non-monotone functional dependence of the log-odds ratio on the matching variable AGE.

confirms that the non-monotone dependence on AGE is significant.

The implications of this finding can be severe since the typical method of detecting interactions between matching and exposure variables consists of adding terms such as MATCHING$\times$EXPOSURE. Now for any exposure variable highly correlated with AGE, AGE$\times$EXPOSURE will be highly correlated with AGE^2. Thus, without the AGE^2 term in the model, the imperfect matching on AGE can lead to mistakingly identifying the interaction AGE$\times$EXPOSURE as significant. In the present data none of the exposure variables were highly correlated with AGE so this was not a problem. Our point is that in other cases it can be and that additive modeling can prevent anomalous findings.

3. Algorithms for the linear model.

3.1. The standard algorithm.

Differentiating (6) the score for $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} \boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{k=1}^{K} \sum_{r=0}^{R} (y_{rk}\boldsymbol{x}_{rk} - \mu_{rk}\boldsymbol{x}_{rk}) \\ &= \boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{\mu}) \end{aligned} \tag{13}$$

where $\boldsymbol{\mu} = \{\mu_{rk}\}$, and $\boldsymbol{y} = \{y_{rk}\}$ indicates cases (1) or controls (0). It is also convenient to define $\boldsymbol{\mu}_k$, the sub-vector of these probabilities (which sum to 1) for each matched set, and $\boldsymbol{U} = \text{diag}(\boldsymbol{\mu})$.

Similarly the information matrix can be written

$$\begin{aligned} \mathcal{I} &= \sum_{k=1}^{K} \left[\sum_{r=0}^{R} \mu_{rk}\boldsymbol{x}_{rk}\boldsymbol{x}_{rk}^t - \left(\sum_{r=0}^{R} \mu_{rk}\boldsymbol{x}_{rk} \right) \left(\sum_{r=0}^{R} \mu_{rk}\boldsymbol{x}_{rk} \right)^t \right], \\ &= \boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X} \end{aligned} \tag{14}$$

where $\boldsymbol{W}$ is a $N \times N$ block diagonal matrix with kth block $\boldsymbol{W}_k = \boldsymbol{U}_k - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^t$.

The Newton-Raphson update (eg Breslow and Day, 1980) can be expressed in IRLS form as

$$\boldsymbol{\beta}^{new} = \left(\boldsymbol{X}^t\boldsymbol{W}^{old}\boldsymbol{X} \right)^{-1} \boldsymbol{X}^t\boldsymbol{W}^{old} \left(\boldsymbol{X}\boldsymbol{\beta}^{old} + (\boldsymbol{W}^o ld)^-(\boldsymbol{y} - \boldsymbol{\mu}^{old}) \right), \tag{15}$$

where $\boldsymbol{\mu}^{old}$ and $\boldsymbol{W}^{old}$ denote $\boldsymbol{\mu}$ and $\boldsymbol{W}$ evaluated at $\boldsymbol{\beta}^{old}$, and $\boldsymbol{W}^-$ denotes a generalized inverse of $\boldsymbol{W}$ and is equivalent to $\text{diag}(\boldsymbol{\mu})^{-1}$. See Green (1984) for a discussion of IRLS algorithms.

3.2. Implementation of the Newton-Raphson algorithm.

Although the Newton-Raphson update (15) is quite straightforward, it is not entirely convenient because the weight matrix $\boldsymbol{W}$ is not diagonal. Thus one could not fit the model directly in GLIM. Matched *pairs* are a notable exception (Holford et al, 1978); standard logistic regression analysis of the 'response' $y \equiv 1$ and 'exposures' $\boldsymbol{X} = \boldsymbol{X}_{case} - \boldsymbol{X}_{control}$, gives the correct estimates of $\boldsymbol{\beta}$ (no intercept). But since this trick does not generalize to more general matching, and does not help us in the non-parametric setting, we do not consider it further.

One can exploit the special structure of the weight matrix $\boldsymbol{W}$ to derive a diagonally weighted version of the conditional algorithm (15) (Pregibon, 1982). We write the matrix $\boldsymbol{W}_k = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{\mu}_k^t)^t\boldsymbol{U}_k(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{\mu}_k^t)$, where $\boldsymbol{1}$ is a column of R+1 ones. Now if $\boldsymbol{X}_k$ denotes the matrix of exposure variables for the kth matched set, $\tilde{\boldsymbol{X}}_k = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{\mu}_k^t)\boldsymbol{X}_k$ is the corresponding centered version; the centering is done by removing the μ-weighted average for each exposure variable. Thus $\boldsymbol{X}_k^t\boldsymbol{W}_k\boldsymbol{X}_k = \tilde{\boldsymbol{X}}_k^t\boldsymbol{U}_k\tilde{\boldsymbol{X}}_k$, and if $\tilde{\boldsymbol{X}}$ denotes the entire matrix of exposure variables with each block centered by its block μ-weighted average, then $\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X} = \tilde{\boldsymbol{X}}^t\boldsymbol{U}\tilde{\boldsymbol{X}}$. Noting that $\boldsymbol{X}^t(\boldsymbol{y}-\boldsymbol{\mu}) = \tilde{\boldsymbol{X}}^t(\boldsymbol{y}-\boldsymbol{\mu})$, substitution into (15) yields

$$\boldsymbol{\beta}^{new} = \left(\tilde{\boldsymbol{X}}^t\boldsymbol{U}^{old}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^t\boldsymbol{U}^{old}\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta}^{old} + (\boldsymbol{U}^{old})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}^{old})\right). \tag{16}$$

For clarity we describe the procedure in algorithmic form.

Newton-Raphson Algorithm

Initialize: Set $\boldsymbol{\beta} = \boldsymbol{0}$.

Cycle: Compute $\mu_{rk} = \exp(\boldsymbol{x}_{rk}^t\boldsymbol{\beta})/\sum_{r=0}^{R}\exp(\boldsymbol{x}_{rk}^t\boldsymbol{\beta})$.

Center $\boldsymbol{x}_{rk}$ by its μ-weighted average, i.e. $\tilde{\boldsymbol{x}}_{rk} = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{\mu}_k^t)\boldsymbol{x}_{rk}$.

Compute $z_{rk} = \tilde{\boldsymbol{x}}_{rk}^t\boldsymbol{\beta} + (y_{rk} - \mu_{rk})/\mu_{rk}$.

Compute the new $\boldsymbol{\beta}$ by regressing $\boldsymbol{z}$ on $\tilde{\boldsymbol{X}}$ with weights $\boldsymbol{\mu}$.

Until: the deviance $D = -2\sum_k \log\mu_{0k}$ converges.

An alternative derivation of the above algorithm is based on a suggestion of Whitehead (1980) for the proportional hazards model. Here one treats the case-control indicator y_{rk} as Poisson with mean $\nu_{rk} = \exp(\alpha_k + \boldsymbol{x}_{rk}^t\boldsymbol{\beta})$.* IRLS for the Poisson amounts to repeatedly regressing $z_{rk} = \alpha_k + \boldsymbol{x}_{rk}^t\boldsymbol{\beta} + (y_{rk} - \nu_{rk})/\nu_{rk}$ onto the columns of $(\boldsymbol{A} : \boldsymbol{X})$ using weights ν_{rk}, where A is the $N \times K$ design matrix for the α's. But since for fixed $\boldsymbol{\beta}$, $\nu_{rk} = \mu_{rk}$, we can achieve the multiple regression by first *adjusting* $\boldsymbol{X}$ for the columns of $\boldsymbol{A}$, yielding $\tilde{\boldsymbol{X}}$, and then simply regressing $\boldsymbol{z}$ on $\tilde{\boldsymbol{X}}$. This is in fact the Newton-Raphson update (16).

3.3. Delta Method algorithms.

The above algorithm requires duplicate storage for $\tilde{\boldsymbol{X}}$. At the cost of slightly slower convergence this can be eliminated. It derives from the conditional algorithm described by (15) where we replace the block matrix $\boldsymbol{W}$ by its diagonal $\overline{\boldsymbol{W}} = \mathrm{diag}(\mu_{rk}(1-\mu_{rk}))$. The algorithm results in the IRLS step

$$\boldsymbol{\beta}^{new} = \left(\boldsymbol{X}^t\overline{\boldsymbol{W}}^{old}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\overline{\boldsymbol{W}}^{old}\left(\boldsymbol{X}\boldsymbol{\beta}^{old} + (\overline{\boldsymbol{W}}^{old})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}^{old})\right). \tag{17}$$

Jørgenson (1984) gave the name "delta" to modified Newton algorithms of this kind, where the weight matrix is replaced by an approximation. His suggestion in the case of conditional problems

* This model is not to be confused with the original logit model (7); it is simply a device for maximizing the conditional likelihood. This works since the maximum of the 'Poisson' log likelihood function for β is identical, apart from constants, to that of the log conditional likelihood function, $l(\boldsymbol{\beta})$.

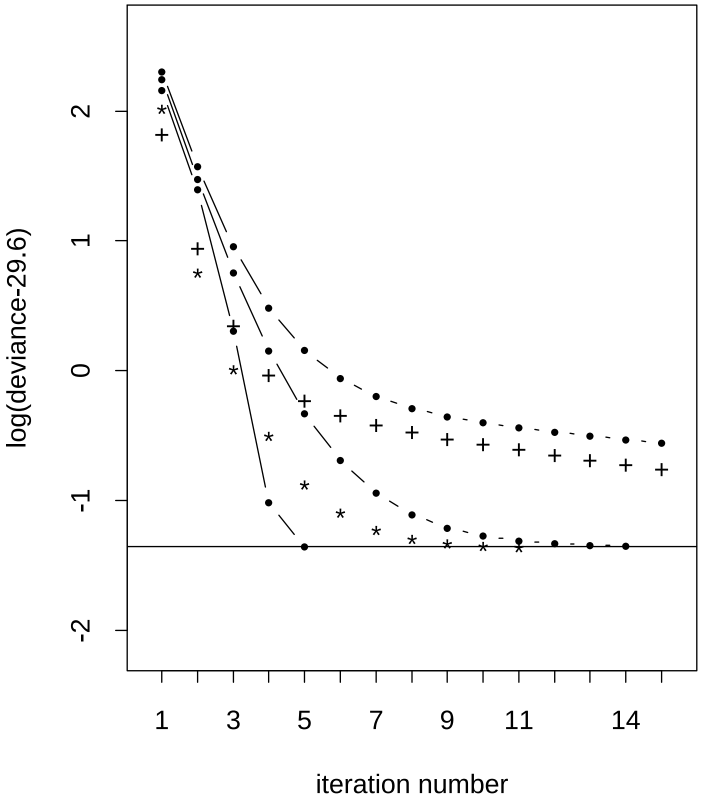

Figure 2. Convergence patterns in fitting the linear model to the depression data (7 variables). The steepest curve is the Newton-Raphson algorithm, the flattest curve is the delta algorithm, and the middle curve the delta algorithm with intercept. The $+$'s below the top curve show the effect of step length optimization for that curve, as do the $*$'s below the middle curve

such as this implies using the 'Poisson' weights μ_{rk} on the diagonal, but we found empirically that with these weights the convergence was extremely slow.

We demonstrate the convergence of the different algorithms on data from a study of possible physiological causes of depression (Rubin et al, 1987). The exposure variables represent concentrations of 8 hormones for 40 depressed patients and their matched controls. Figure 2 shows, on the log scale, the convergence of the deviance for the Newton-Raphson algorithm (15), the delta algorithm just described, and one still to be described. Initially the delta algorithm shows a convergence pattern similar to the Newton-Raphson, but then flattens off and converges linearly. We can improve the delta algorithm (17) by augmenting $\boldsymbol{X}$ with a column of ones, i.e. adding the grand mean to the linear predictor. (This constant cancels in the definition of μ_{rk}.) The resulting algorithm is still in the class of delta algorithms since, $\boldsymbol{X}^{*t}\overline{\boldsymbol{W}}\boldsymbol{X}^{*} = \boldsymbol{X}^{t}\overline{\overline{\boldsymbol{W}}}\boldsymbol{X}$, where $\boldsymbol{X}^{*} = (\boldsymbol{1} : \boldsymbol{X})$ and $\overline{\overline{\boldsymbol{W}}} = \overline{\boldsymbol{W}} - \overline{\boldsymbol{W}}\boldsymbol{1}(\boldsymbol{1}^{t}\overline{\boldsymbol{W}}\boldsymbol{1})^{-1}\boldsymbol{1}^{t}\overline{\boldsymbol{W}}$. The constant term effectively puts values on the off-diagonal of the weight matrix which attempts to adjust for all the matched sets simultaneously. As can be seen in the figure, the convergence for this augmented diagonal algorithm is improved considerably.

The Newton-Raphson algorithm typically converges quadratically given a reasonable starting guess, although convergence is never guaranteed without step size optimization. In practice this is seldom a problem. Delta algorithms can be shown to be ascent methods so they will also converge with step size optimization (Jørgenson, 1984). One way of viewing the step size is instead of using $\boldsymbol{\eta}^{new}$ as the new linear predictor, one uses $(1-w)\boldsymbol{\eta}^{old} + w\boldsymbol{\eta}^{new}$ for some $w > 0$. We implemented an *Armijo-Goldstein* step search routine (Gill, Murray and Wright, 1980, section 4.3) which safeguards against over-stepping. Figure 2 shows the improvement due to step size optimization to both our delta algorithms. In particular, the delta remains inferior to its augmented version (for these examples).

4. Estimation of the additive model.

In this section we derive an algorithm for estimating the components of the additive model (7). We follow the penalized likelihood approach and present it in a simplified form, referring to the already large literature for details. For penalized likelihood techniques close in spirit to this work see Hastie and Tibshirani (1986c) and O'Sullivan (1986 a&b).

4.1. Spline estimation for a single function.

Consider a single exposure variable x. Without loss of generality we assume there are no ties in the x_{rk}. The logarithm of the penalized conditional likelihood for the model $\log(\pi(x)) = f(x)$ is

$$l(f) = \sum_{k=1}^{K}\left[f(x_{0k}) - \log\left(\sum_{r=0}^{R}\exp(f(x_{rk}))\right)\right] - \frac{1}{2}\lambda\int[f''(s)]^2 ds. \tag{18}$$

The criterion has two components: the likelihood component measures fidelity of the function to the data, and the integrated squared second derivative component measures its smoothness. The smoothness penalty λ controls this balance, and has to be supplied.

When working with problems such as these there are some standard approaches that can make the task much easier. One first establishes that a solution exists, and that it is a cubic spline. In fact, existence is sufficient, since given any solution with fitted values $f(x_{rk})$, the interpolating cubic spline does better on the penalty (Reinsch, 1967). The elegant results of O'Sullivan (1983) show that if the likelihood is convex, existence (uniqueness) is guaranteed as long as a solution exists (is unique) for the likelihood part of (18) over the space of *linear* functions (i.e. over the null space of the penalty). This is the case here (see also O'Sullivan, Yandell and Raynor, 1986, O'Sullivan, 1986a). So we know the solution exists and is in fact a piecewise cubic polynomial between the unique values of x_{rk}, and linear beyond the endpoints.

Next we pick a suitable basis for representing such functions. The standard computational representation is usually in terms of B-splines (de Boor, 1978): $f(s) = \sum_i b_i(s)\theta_i$, where the b_i are the B-spline basis functions and the θ_i are the parameters to be estimated. Instead we use the same bases as Green and Yandell (1985), which essentially uses as basis the delta function at each fitted value, and so the parameters to be estimated are the $N = K \times (R+1)$ fitted values themselves $\boldsymbol{f} = \{f_{rk}\}$. This representation is convenient when only the fitted values are of concern, and leads to slightly simpler algebra. It is worth noting that even for problems that do not yield cubic splines as the optimum, one can simply impose cubic spline structure using either of these bases—see O'Sullivan, 1986b. Using the delta function basis, the integrated squared second derivative component simplifies to $-\frac{1}{2}\lambda\boldsymbol{f}^t\boldsymbol{K}\boldsymbol{f}$ where $\boldsymbol{K}$ is a $N \times N$ *penalty* matrix. Differentiating the penalized log-likelihood (18) with respect to the N components f_{rk} of $\boldsymbol{f}$ leads to the score function

$$\boldsymbol{S}(\boldsymbol{f}) = \boldsymbol{y} - \boldsymbol{\mu} - \lambda\boldsymbol{K}\boldsymbol{f} \tag{19}$$

where $\mu_{rk} = \exp(f_{rk})/\sum_{r=0}^{R}\exp(f_{rk})$ as in (13). The information matrix is $\mathcal{I} = \boldsymbol{W} + \lambda\boldsymbol{K}$, where $\boldsymbol{W}$ is the same as in (14).

The Newton-Raphson update for $\boldsymbol{f}$ is

$$\begin{aligned}\boldsymbol{f}^{new} &= (\boldsymbol{W}^{old} + \lambda\boldsymbol{K})^{-1}\boldsymbol{W}^{old}\left[\boldsymbol{f}^{old} + \boldsymbol{W}^{old-}(\boldsymbol{y} - \boldsymbol{\mu}^{old})\right] \\ &= \boldsymbol{S}_W\boldsymbol{z},\end{aligned} \tag{20}$$

where $\boldsymbol{z}$ is the adjusted dependent variable. The matrix $\boldsymbol{S}_W = (\boldsymbol{W} + \lambda\boldsymbol{K})^{-1}\boldsymbol{W}$ resembles a cubic spline smoother matrix. Indeed, this linear matrix operator produces a vector of fitted values that lie on a cubic spline. In most non-parametric regression contexts $\boldsymbol{W}$ is diagonal and $\boldsymbol{K}$ has special banded structure which allows one to apply the smoother in $\mathrm{O}(N)$ operations. This is not the case here, even though $\boldsymbol{W}$ is block diagonal and thus banded itself. $\boldsymbol{K}$ is banded if the rows are ordered with $\boldsymbol{x}$, but then this ordering destroys the block diagonal structure of $\boldsymbol{W}$. Thus $\boldsymbol{W} + \lambda\boldsymbol{K}$ is a full matrix and expensive to invert ($\mathrm{O}(N^3)$ operations). O'Sullivan (1986a) encountered exactly the same problem with the proportional hazards model and solved the system in O(N) operations using a specialized preconditioned conjugate gradient technique which required 2 or 3 iterations per inversion.

Our solution is to replace $\boldsymbol{W}$ by its diagonal $\overline{\boldsymbol{W}}$ as in section 3. Note that we do not use $\overline{\overline{\boldsymbol{W}}}$ since the smoother itself fits an intercept. The update formula for $\boldsymbol{f}$ is

$$\begin{aligned} \boldsymbol{f}^{new} &= (\overline{\boldsymbol{W}} + \lambda \boldsymbol{K})^{-1} \overline{\boldsymbol{W}} \left[\boldsymbol{f}^{old} + \overline{\boldsymbol{W}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}^{old}) \right] \\ &= \boldsymbol{S}_{\overline{W}} \boldsymbol{z}, \end{aligned} \tag{21}$$

where $\boldsymbol{S}_{\overline{W}}$ is a diagonally weighted cubic spline smoother.

This algorithm works well in practice; figure 3(a) shows a typical convergence pattern using this algorithm to estimate a single function for the depression data. Figure 3(b) shows the fitted function it produced (broken curve).

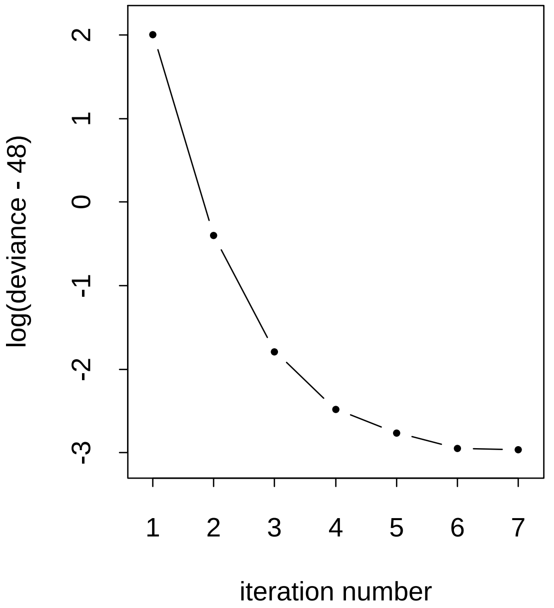

Figure 3a. The convergence pattern of the deviance for fitting the spline function displayed in figure 3(b).

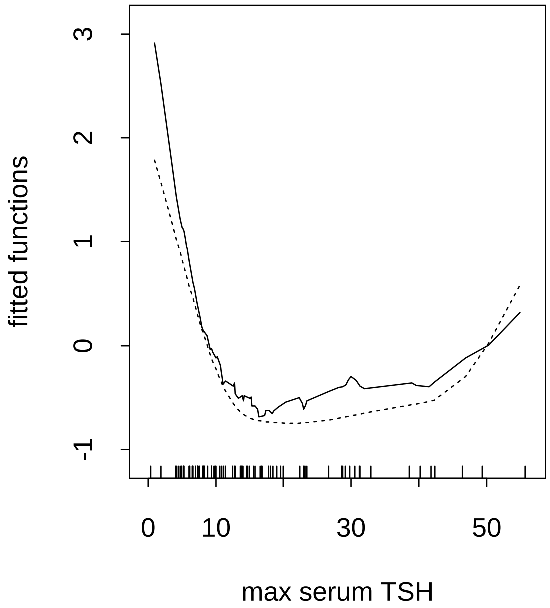

Figure 3b. The fitted function for one of the variables in the depression data set. The dashed curve was estimated using the spline smoother outlined above; the solid curve used the same algorithm with a running lines smoother. Both smoothers were standardized to do about the same amount of smoothing.

4.2. The additive model algorithm.

The additive model has the form

$$\mu_{rk} = \frac{\exp\left(\eta(\boldsymbol{x}_{rk})\right)}{\sum_{r=0}^{R} \exp\left(\eta(\boldsymbol{x}_{rk})\right)} \quad r = 0, \cdots, R,\ k = 1, \cdots, K, \tag{22}$$

where the *additive predictor* is $\log(\pi(\boldsymbol{x}_{rk})) = \eta(\boldsymbol{x}_{rk}) = \sum_{j=1}^{m} f_j(x_{rkj})$. The conditional penalized likelihood now has a penalty for each of the functions:

$$l(\boldsymbol{f}_1, \cdots, \boldsymbol{f}_m) = \sum_{k=1}^{K} \log(\mu_{0k}) - \frac{1}{2} \sum_{j=1}^{m} \lambda_j \boldsymbol{f_j}^t \boldsymbol{K} \boldsymbol{f}_j. \tag{23}$$

The derivation is similar to the univariate case and a bit of algebra leads to the set of *normal equations* which need to be solved for the vectors of new fitted functions in a Newton-Raphson step:

$$\begin{pmatrix} I & \boldsymbol{S}_1 & \boldsymbol{S}_1 & \cdots & \boldsymbol{S}_1 \\ \boldsymbol{S}_2 & I & \boldsymbol{S}_2 & \cdots & \boldsymbol{S}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_m & \boldsymbol{S}_m & \boldsymbol{S}_m & \cdots & I \end{pmatrix} \begin{pmatrix} \boldsymbol{f}_1^{new} \\ \boldsymbol{f}_2^{new} \\ \vdots \\ \boldsymbol{f}_m^{new} \end{pmatrix} = \begin{pmatrix} \boldsymbol{S}_1 \boldsymbol{z} \\ \boldsymbol{S}_2 \boldsymbol{z} \\ \vdots \\ \boldsymbol{S}_m \boldsymbol{z} \end{pmatrix}, \tag{24}$$

where the $\boldsymbol{S}_j$ are $N \times N$ spline smoother matrices of the form $\boldsymbol{S}_W$ for smoothing against variable x_j, and $\boldsymbol{z}$ is once again the adjusted dependent variable $\boldsymbol{z} = \boldsymbol{\eta}^{old} + (\boldsymbol{W}^{old})^{-}(\boldsymbol{y} - \boldsymbol{\mu}^{old})$.

Except in some special cases the system (24) is prohibitively expensive to solve directly. The Gauss-Seidel iterative procedure makes exceptional good use of its structure and has intuitive appeal. Also termed the *backfitting algorithm*, it proceeds as follows:

The Backfitting Algorithm

$$\begin{aligned} &\textbf{Initialize:} && \boldsymbol{f}_j = \boldsymbol{f}_j^0, j = 1, 2, \ldots m, \\ &\textbf{Cycle:} && j = 1, 2, \ldots m, 1, 2, \ldots \\ & && \boldsymbol{f}_j = S_j(\boldsymbol{z} - \sum_{k \neq j} \boldsymbol{f}_k) \\ &\textbf{Until:} && \text{the individual functions don't change.} \\ &\textbf{Finalize:} && \text{set } \boldsymbol{f}_j^{new} = \boldsymbol{f}_j,\ j = 1, 2, \ldots m. \end{aligned} \tag{25}$$

Buja et al (1987) prove that the backfitting algorithm always converges for spline smoothers. They also discuss more sophisticated versions of the same algorithm.

Our implementation uses the diagonal approximation of $\boldsymbol{W}$ by $\overline{\boldsymbol{W}}$ in order that the system be solved in $O(N)$ operations. In addition it splits the function f into linear and non-linear components, $f(x) = x\beta + g(x)$, resulting in the additive predictor, $\eta_f(\boldsymbol{x}_{rk}) = \eta_L(\boldsymbol{x}_{rk}) + \eta_g(\boldsymbol{x}_{rk})$. There are several advantages in doing this:

- The linear component can be computed by projection. Thus the general orientation of the fitted functions (their slopes) is determined without Gauss-Seidel iteration.
- We know how to fit the linear component efficiently without approximating the weight matrix.
- As a by-product we get the linear component of the fit, and can use it to judge the amount of non-linearity.

The Split Backfitting Algorithm

0: Initialize Fit the *linear* model $\eta_L(\boldsymbol{x}_{rk}) = \boldsymbol{x}_{rk}^t\boldsymbol{\beta}$, using either the Newton-Raphson or one of the delta algorithms described in section 3. Set the non-linear component $\eta_g(\cdot) = 0$. Compute $\mu_{rk} = \exp(\eta(\boldsymbol{x}_{rk}))/\sum_k \exp(\eta(\boldsymbol{x}_{rk}))$, and deviance.

1: Additive step. Compute weights $\overline{w}_{rk} = \mu_{rk}(1-\mu_{rk})$ and adjusted dependent variates

$$z_{rk} = \eta_g(\boldsymbol{x}_{rk}) + \frac{y_{rk} - \mu_{rk}}{\mu_{rk}(1-\mu_{rk})}.$$

Fit a weighted additive model $\eta_g(\boldsymbol{x}_{rk}) = \sum_{j=1}^m g_j(\boldsymbol{x}_{rkj})$ to the z_{rk} using the weighted backfitting algorithm. Compute μ_{rk}.

2: Linear step. Compute weights $w_{rk} = \mu_{rk}$ and adjusted dependent variates

$$z_{rk} = \eta_L(\boldsymbol{x}_{rk}) + \frac{y_{rk} - \mu_{rk}}{\mu_{rk}}.$$

Compute $\tilde{\boldsymbol{x}}_{rk} = (\boldsymbol{I} - \boldsymbol{1}\boldsymbol{\mu}_k^t)\boldsymbol{x}_{rk}$ and fit the linear model $\eta_L(\boldsymbol{x}_{rk}) = \tilde{\boldsymbol{x}}_{rk}^t\boldsymbol{\beta}$ to the z_{rk} by weighted least squares. Compute μ_{rk} and the deviance.

3: Until: Repeat steps 1 and 2 until the fitted functions and coefficients do not change.

4: Finalize. Compute $f_j(x_j) = \beta_j x_j + g_j(x_j)$, $j = 1, \ldots, m$.

5. Discussion.

The system of equations (24) is appropriate for a wide variety of regression estimators, not only smoothing splines. Some common examples are:

- If we wish to replace f_j by the linear term $\beta_j x_j$, then we replace $\boldsymbol{S}_j$ in (24) by the least squares "hat" matrix $\boldsymbol{H}_j$. Groups of linear terms can be collected together and one "$\boldsymbol{H}$" matrix used to represent them.

- If variable x_j is categorical with J levels, then we replace $\boldsymbol{S}_j$ by the operator that calculates the category means.
- Other smoothers, not only splines, can be represented by $\boldsymbol{S}$.

In practice this means that each variable can be associated with a particular smoothing operator. Each time the variable has a turn in the backfitting algorithm, the smoothing operator is applied to the partial residuals resulting in a fitted function for the variable. This procedure can be further generalized to include associating a single smoothing operator with more than one variable (e.g. surface smoothing).

The point of view that the spline smoother can be replaced with any reasonable non-parametric regression estimator, suggests that any of the commonly used scatterplot smoothers can be used provided they can be modified to incorporate observation weights. Amongst others, this class includes "locally weighted running lines" (Cleveland, 1979), kernel smoothers (e.g.Watson, 1964), and "supersmoother" (Friedman and Stuetzle, 1982). Figure 3(b) shows the fit produced using a weighted *running lines* smoother in place of $\boldsymbol{S}_{\overline{W}}$; the shapes are very similar (the curves are location free).

Both smoothers in figure 3 have a smoothing parameter that needs to be specified. There exist a number of automatic techniques for making this selection in simple non-parametric regression, such as generalized and ordinary cross-validation. These techniques become far more complicated in iterative and multiple regression situations, as well as computationally intensive. The approach we recommend, although admittedly somewhat ad hoc, is based on the "degrees of freedom" (DF) of the smoother. One of the suggestions of Buja et al (1987) in this regard is to estimate the number of parameters used up by the smooth by $DF = trace(\boldsymbol{S})$. The quantity DF is monotone decreasing in the smoothing parameter λ. When λ approaches infinity, the smooth approaches a line and DF approaches 2; similarly $DF(0) = N$. We have found that values for DF around 4 (the value used in figure 3) produce curves useful for exploring functional form in regression-like models.

Thus our algorithm draws only on basic tools available in most statistical packages, namely weighted linear least squares and weighted scatterplot smoothing. Thus no additional specialized software is required although high quality graphical output is desirable.

Our example illustrates that additive models are a valuable tool in model formulation and interpretation of results. As much of the model formulation stage of an analysis is informal and exploratory in nature, we resist the temptation to promote formal theories of estimation and testing for our methods. Rather we feel that a greater contribution would be to detail how we use these tools in practice, together with other tools already in use, such as outlier detection techniques, regression diagnostics, goodness-of-fit, etc.

We see no alternative starting point other than 'looking at the data'. In particular, the univariate and bivariate distributions of the covariates and/or any stratification variables should be examined. The purposes are at least threefold

- to detect gross outliers in the univariate distributions (e.g. age ≤ 0)
- to detect inconsistencies in related covariates (e.g. length of residence in Singapore $\geq$ age)
- to identify variables with unusual distributions (e.g. spikes at 0)

The idea is to know ahead of time, i.e. prior to modeling, how the data behave so that obvious blunders can be avoided and that the techniques can be tuned to match particular features of the data. Most techniques we use tend to be graphical in nature, so that when we say 'look at the data', we mean it literally. When dealing with a moderate number of covariates, the 'scatter plot' matrix (Chambers et al, 1985), provides a useful summary of the data as regards the features of interest at this stage.

After editing the data and perhaps determining an initial parametric form for each covariate, we then recommend fitting the 'classical' (linear) model. At this stage, the emphasis is not on the fitted model so much, but rather on the residuals, i.e. what was not fitted by the model. Thus a thorough diagnostic analysis is recommended at this stage (Pregibon, 1984). The outcome is typically the identification of subsets of cases/controls which either are not fitted well by the model or are unduly influential. These are the symptoms; as for the cure many possibilities present themselves:

- leaving out troublesome observations
- re-expressing to enhance linearity
- re-parameterization to capture model inadequacies

The problem in choosing one of these alternative 'cures' is that the diagnostics do not give much insight into which might be best. This is where we believe additive modeling can usefully fit in, especially as regards the last two possibilities.

Currently we employ a backward selection strategy, deleting unimportant terms sequentially. Only after the model has settled do we look at the fitted functions and see what they suggest. Without this strategy there is a tendency to interpret insignificant but exotic functions whose presence may also affect the important functions.

We feel that the additive model should be treated as a supplement to, rather than a substitute for, the classical linear model. The non-linear contributions of covariates graphically displayed by the function plots not only indicate problems with the linear model but also indicate how to ameliorate them. Thus parametric re-expressions and innovative re-parametrizations suggest themselves rather than being left to the ingenuity of the modeler.

References

Baker, R.J., and Nelder, J.A., (1978), *The GLIM System Release 3*, Oxford: Numerical Algorithms Group.

Breslow, N.E. and Day, N.E. (1980). *The Analysis of Case-Control Studies*, IARC Scientific Publications, Lyon.

Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.L., and Sabai, C. (1978). "Estimation of Multiple Relative Risk Functions in Matched Case-Control Studies", *American Journal of Epidemiology*, **4**, 299-307.

Buja, A., Hastie, T., and Tibshirani, R. (1987). "Linear Smoothers and the Additive Model", submitted for publication.

Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P. (1985). *Graphical Methods for Data Analysis*, Wadsworth, California.

Cleveland, W.S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots", *Journal of the American Statistical Association* , **74**, 829-836.

Cox, D.R. (1972). "Regression Models and Life Tables (with discussion)", *Journal of the Royal Statistical Society, B*, **34**, 187-202.

De Boor, C. (1978), *A Practical Guide to Splines.* Springer-Verlag, New York.

Friedman, J.H. and Stuetzle, W. (1982). "Smoothing of Scatterplots", *Department of Statistics Technical Report, Orion 3*, Stanford University.

Gill, P., Murray, W., and Wright, M. (1980). *Practical Optimization*, Academic Press, New York.

Green, P.J. (1984). " Iteratively Reweighted Least Squares for Maximum Likelihood Estimation and some Robust and Resistant Alternatives (with discussion)", *Journal of the Royal Statistical Society, B*, **46**, 149-192.

Green, P. and Yandell, B. (1985). "Semi-Parametric Generalized Linear Models", *Proceedings 2nd International GLIM Conference, Lancaster*, Springer-Verlag lecture notes in Statistics #32, Berlin, Heidelberg.

Hastie, T. and Tibshirani, R. (1986a). "Generalized Additive Models (with discussion)", *Statistical Science*, **1**, No 3., 297-318.

Hastie, T. and Tibshirani, R. (1986b). "Generalized Additive Models, Cubic Splines and Penalized Likelihood", *Division of Biostatistics Technical Report*, University of Toronto.

Hastie, T. and Tibshirani, R. (1986c). "A Non-Parametric Extension of the Proportional Hazards Model", submitted for publication.

Holford, T.R., White, C. and Kelsey, J.L. (1978). "Multivariate Analysis for Matched Case-Control Studies", *American Journal of Epidemiology*, **107**, 245-256.

Jørgenson, B (1984), "The Delta Algorithm and GLIM", *International Statistical Review*, **52**, 283-300.

O'Sullivan, F. (1983). *The Analysis of some Penalized Likelihood Estimation Schemes*, Statistics Department Technical Report #726, University of Wisconsin, Madison.

O'Sullivan, F. (1986a). "Estimation of Densities and Hazards by the Method of Penalized Likelihood", *Department of Statistics Technical Report #58*, University of California, Berkeley.

O'Sullivan, F. (1986b). "Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators", *SIAM Journal of Statistical Computing*, to appear.

O'Sullivan, F., Yandell, B. and Raynor, W. (1986). "Automatic Smoothing of Regression Functions in Generalized Linear Models", *Journal of the American Statistical Association* , **81**, 96-103.

Pregibon, D. (1982). "Score Tests in GLIM", *Proceedings of the International Conference on Generalized Linear Models, London*, Springer-Verlag lecture notes in Statistics #14, Berlin, Heidelberg.

Pregibon, D. (1984). "Data Analytic Methods for Matched Case-Control Studies", *Biometrics*, **40**, 639-651.

Reinsch, C. (1967), "Smoothing by Spline Functions", *Numerische Mathematik*, **10**, 177-183.

Rubin, R., Hastie, T., Pregibon, D., and Wheeler, N. (1987). "Nueroendocrine Aspects of Primary Endogenous Depression-V. Methodology for Matched Controls", *In Preparation.*

Watson, G.S. (1964), "Smooth Regression Analysis", *Sankya Series A*, **26**, 359-372.

Whitehead, J. (1980). "Fitting Cox's Regression Model to Survival Data using GLIM", *Applied Statistics*, **29**, 268-275.