Generalized Additive Models: Some Applications

TREVOR HASTIE and ROBERT TIBSHIRANI*

Generalized additive models have the form $\eta(\mathbf{x}) = \alpha + \sum f_j(x_j)$, where η might be the regression function in a multiple regression or the logistic transformation of the posterior probability $\Pr(y = 1 \mid \mathbf{x})$ in a logistic regression. In fact, these models generalize the whole family of generalized linear models $\eta(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$, where $\eta(\mathbf{x}) = g(\mu(\mathbf{x}))$ is some transformation of the regression function. We use the local scoring algorithm to estimate the functions $f_j(x_j)$ nonparametrically, using a scatterplot smoother as a building block. We demonstrate the models in two different analyses: a nonparametric analysis of covariance and a logistic regression. The procedure can be used as a diagnostic tool for identifying parametric transformations of the covariates in a standard linear analysis. A variety of inferential tools have been developed to aid the analyst in assessing the relevance and significance of the estimated functions: these include confidence curves, degrees of freedom estimates, and approximate hypothesis tests.

The local scoring algorithm is analogous to the iterative reweighted least squares algorithm for solving likelihood and nonlinear regression equations. At each iteration, an adjusted dependent variable is formed and an additive regression model is fit using the backfitting algorithm. The backfitting algorithm cycles through the variables and estimates each coordinated function by smoothing the partial residuals.

KEY WORDS: Generalized linear model; Smooth; Nonparametric regression; Logistic regression.

1. INTRODUCTION

There are a number of advancements and enhancements for the linear regression model. The analyst is equipped with an ever-growing toolbox of diagnostic checks for outliers and model deficiencies (see, e.g., Belsley, Kuh, and Welsch 1980; Cook and Weisberg 1982).

Residual and partial residual plots are used to detect departures from linearity and often suggest parametric fixes. An attractive alternative to this indirect approach is to model the regression function nonparametrically and let the data decide on the functional form. For a single covariate a scatterplot smoother estimates the regression function in a local fashion (e.g., Cleveland 1979; Wahba and Wold 1975). The analyst can use the smooth to suggest a parametric form for the term and then apply the appropriate transformation in the linear regression model. Alternatively, the analyst can make interpretations and predictions from the smooth itself. (In what follows, *smoother* refers to the tool used to "smooth" the scatterplot; *smooth* as a noun refers to the resulting curve.)

A number of avenues have been opened for generalizing the scatterplot smoother to multivariate regression. An immediate generalization is the multidimensional smoother. Friedman and Stuetzle (1981), among others, pointed out the dimensionality problems incurred when using such smoothers. Essentially all smoothers base their estimates on some (weighted) average of the neighboring observations. In high dimensions, one has to reach out further to find sufficient neighbors, and the estimate is no longer local. The same authors proposed the *projection pursuit regression* technique as an alternative to multidimensional smoothing. This model has the form

$$E(Y \mid \mathbf{x}) = \sum_{k=1}^{K} f_k(\boldsymbol{\alpha}_k^t \mathbf{x}) = \sum_{k=1}^{K} f_k(z_k) , \qquad (1)$$

where $\mathbf{x}$ is a p-dimensional covariate, the $\boldsymbol{\alpha}_k$'s are direction vectors onto which the data is projected, forming the scalar variables z_k, and each f_k is an arbitrary one-dimensional function. The number K of projections is a parameter of the procedure. Projection pursuit techniques have received a lot of attention lately; for a recent overview see Huber (1985). The *additive model*

$$E(Y \mid \mathbf{x}) = \alpha + \sum_{j=1}^{p} f_j(x_j), \qquad E[f_j(x_j)] = 0, \qquad (2)$$

is a special case of the projection pursuit model in which there are exactly p directions *fixed at the coordinate directions*. This model is less general than the projection pursuit model, but is more easily interpretable. As in linear regression, we can examine the effect of covariates one at a time, conditional on the presence of the other covariates. But here we model the effects in a general nonparametric way. Alternatively, one can use the estimated functions in (2) to suggest parametric transformations for the covariates. A more traditional approach is to use residual and partial residual plots for this purpose; in this case nonlinearities are detected in one of the covariates and all of the others are kept linear. See Landwehr (1986) and Denby (1986) for recent developments. Partial residual plots can fail if nonlinearities are present for more than one covariate; in (2) we can detect all of the nonlinearities simultaneously.

Nelder and Wedderburn (1972) and McCullagh and Nelder (1983) described in detail the class of *generalized linear models* of the form $g(\mu(\mathbf{x})) = \boldsymbol{\beta}'\mathbf{x}$, where $\mu(\mathbf{x}) = E(y \mid \mathbf{x})$, and the density of y is assumed to be in the exponential family. This model includes the normal linear regression model and the logistic regression model as special cases. The function $g(\cdot)$ is called the link function and is usually assumed known. The model is estimated using an iteratively reweighted least squares procedure. If appropriate distributional assumptions are made, this is ex-

* Trevor Hastie is a member of the research staff, Data Analysis and Statistics Group, AT&T Bell Laboratories, Murray Hill, NJ 07974. Robert Tibshirani is NSERC University Research Fellow and Assistant Professor, Department of Preventive Medicine and Biostatistics and Department of Statistics, University of Toronto, Canada M5S1A8. The first draft of this article was prepared while Hastie was employed as a biostatistician at the South African Medical Research Council. Tibshirani acknowledges the Natural Science and Engineering Research Council of Canada for its support. The authors thank David Steinberg for bringing to their attention the data in Section 3 and the Rothamsted Experimental Station for allowing them to use it, and Dennis Capatos for the use of his preliminary analysis of the data in Section 4. They also thank the editor, associate editor, and referees for their helpful comments.

© 1987 American Statistical Association
Journal of the American Statistical Association
June 1987, Vol. 82, No. 398, Applications

actly maximum likelihood estimation; otherwise the procedure is justified on the basis of quasi-likelihood.

In this article we illustrate the extension of the additive model to the exponential family. A *generalized additive model* (GAM) has the form

$$g(\mu(\mathbf{x})) = \alpha + \sum_{j=1}^{p} f_j(x_j), \qquad E[f_j(x_j)] = 0, \qquad (3)$$

and we estimate it using the *local scoring algorithm*. This is a natural generalization of the iterative least squares procedure for the linear problem. This article describes these models and illustrates their use on some real data applications. Further details can be found in Hastie and Tibshirani (1986a).

Probably the most popular generalized linear model (other than the normal model) is the linear logistic model. Pregibon (1981, 1982) and Landwehr, Pregibon, and Shoemaker (1984) generalized the ideas of regression diagnostics to linear logistic regression problems, where the response variable y is 0 or 1. Partial residual plots were used to detect nonlinearities in the model. In this case model (3) is

$$\operatorname{logit} p(\mathbf{x}) = \alpha + \sum_{j=1}^{p} f_j(x_j), \qquad E[f_j(x_j)] = 0, \qquad (4)$$

where $p(\mathbf{x}) = \Pr(y = 1 \mid \mathbf{x})$ and $\operatorname{logit} p(\mathbf{x}) = \log[p(\mathbf{x})/(1 - p(\mathbf{x}))]$; hence the log odds are modeled in an additive but nonparametric fashion.

The local scoring algorithm is asymptotically equivalent to local likelihood estimation, another method for estimating models of the form (3) (Tibshirani and Hastie 1987). Local scoring has the advantage of being considerably faster. O'Sullivan, Yandell, and Raynor (1986) proposed a method of modeling generalized linear models in a nonparametric way by using spline functions. This technique would yield estimates similar to ours in the case of a single covariate. To our knowledge, they have not used splines to estimate the model (3); instead they model $g(\mu(\mathbf{x}))$ directly by using high-dimensional splines, which are computationally cumbersome and difficult to interpret and display beyond two dimensions. See also Green and Yandell (1985) for models with a single smooth function. Models involving two-dimensional surfaces can be useful for detecting interactions; we would favor models of the form $g(\mu(\mathbf{x})) = \alpha + f_1(x_1) + f_2(x_2) + f_{34}(x_3, x_4)$, say, for this purpose (Hastie 1986). The local scoring procedure can also be used in nonexponential family models like the proportional hazards model (Hastie and Tibshirani 1986a).

We give two applications of the local scoring procedure: in Section 3 we perform a nonparametric analysis of covariance, and in Section 4 we analyze some coronary risk factors by using nonparametric logistic regression. The analyses were performed using GAIM, an interactive Fortran program [the name derives from GLIM (Baker and Nelder 1978), with the obvious modification]. A copy of the software is available on request from either author.

2. THE LOCAL SCORING ALGORITHM

The *scatterplot smoother* is at the heart of the procedures and will be described first. Next we describe the *backfitting algorithm* that uses the scatterplot smoother to estimate the functions f_j in the model (1). Finally, the *local scoring algorithm* applies the backfitting algorithm iteratively to fit the GAM (3).

2.1 The Scatterplot Smoother

The scatterplot smooth of a set of observations (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ can be thought of as an estimate of $E(y \mid X = x_i)$, the mean of y at $x = x_i$. We denote the smooth of the data $\mathbf{x}$, $\mathbf{y}$ at the point x_i by $S(y \mid x_i)$. A variety of smoothers exists in the literature. We use the running lines smoother (Cleveland 1979; Friedman and Stuetzle 1982), which has the form

$$S(y \mid x_i) = \hat{a}_i + \hat{b}_i x_i, \qquad (5)$$

where $\hat{a}_i$ and $\hat{b}_i$ are the estimated intercept and slope in the simple linear regression using only those pairs (x_j, y_j) in some neighborhood $N(i)$ of x_i. We use *symmetric nearest neighborhoods*, which have the form $N(i) = \{x_{i-k}, x_{i-k+1}, \ldots, x_i, \ldots, x_{i+k}\}$ and are truncated at the endpoints if $i - k < 1$ or $i + k > n$. Thus a fit in the middle of the sequence will be based on the point itself and its left and right k neighbors; the rightmost fit will be based on the point itself and its k left neighbors. Finally, we correct our version of the smoother so that average$(S(y \mid x_i))$ = average(y_i).

The parameter k determines the size of the neighborhood, and we refer to the number $(2k + 1)/n$ as the span. Large spans produce smooth curves high in bias and low in variance, whereas small spans have the opposite effect. Although the spans can be picked by cross-validation (Friedman and Stuetzle 1982; Stone 1974), we tend to use intuitively reasonable spans for data exploration, a typical value being in the range 10%–50%. Figure 3 (see p. 376) shows a smooth of some crop yield data, using a span of 10%.

An advantage of the running line smoother is speed of computation; the formulas for the slope and intercept in (5) can be updated easily from one neighborhood to the next, since at most two points change. This feature, and its intuitive nature, makes the running lines smoother an attractive choice as a primitive in more complex algorithms, such as ours, where repeated smooths are required. We emphasize, however, that in what follows the operator $S(y \mid x_i)$ can be replaced by any regression-type scatterplot smoother, such as smoothing splines (e.g., Silverman 1985; Wahba and Wold 1975), kernel smoothers, or even the composite nonlinear smoothers described in Mosteller and Tukey (1977).

Finally, we will be dealing with situations where the observations are reweighted from one iteration to the next. In this case, the span determines the total *weight* in each neighborhood, and weighted least squares is used to compute $\hat{a}_i$ and $\hat{b}_i$ in (5).

2.2 The Backfitting Algorithm

The backfitting algorithm estimates the functions f_j in the model $E(y \mid \mathbf{x}) = \alpha + \Sigma f_j(x_j)$, as follows:

```
Initialization  $\hat{f}_j(x) = 0 \; \forall x$ and $\forall j$, $\hat{\alpha} = \bar{y}$ .
    Cycle  $j = 1, 2, \ldots, p, 1, 2, \ldots, p, 1, 2, \ldots,$
        $r_{ij} = y_i - \hat{\alpha} - \sum_{\substack{k=1 \\ k \neq j}}^{p} \hat{f}_k(X_{ki}),$   $i = 1, \ldots, n,$
        $\hat{f}_j(x_{ji}) = S(r_j \mid x_{ji}),$   $i = 1, \ldots, n,$
    until  the functions $\hat{f}_j$ converge.
```

Notice that at each stage the algorithm smooths residuals against the next covariate; these residuals are obtained by removing the estimated functions or covariate effects of all of the other variables.

The backfitting algorithm has appeared several times: Mosteller and Tukey (1977) used a similar algorithm (median polish) in fitting additive effects in an analysis of variance (ANOVA) situation. Friedman and Stuetzle (1981) referred to this special case of projection pursuit regression as *projection selection*. Breiman and Friedman (1985) discussed certain details of the backfitting algorithm, which forms the "inner loop" of their "ACE" (alternating conditional expectations) algorithm.

Some of the properties of the algorithm are as follows:

1. If the $S(\cdot)$ refers to global univariate least squares fits, then backfitting converges to the multivariate least squares solution. Hastie, Tibshirani, and Buja (1987) showed that the backfitting algorithm is the Gauss–Seidel method for solving an appropriate system of normal equations. They discussed an improvment of the algorithm given here, and they proved convergence if the smoothers used are linear, symmetric, and *shrinking*. Here, linear means $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ for some $n \times n$ matrix $\mathbf{S}$; the symmetry and shrinking imply that the eigenvalues of $\mathbf{S}$ have absolute values no larger than 1. Cubic spline smoothers are shrinking smoothers, as are projections; running lines smoothers are not necessarily shrinking. They show that individual functions will not be unique if there is *cocurvity* in the data, the analog of colinearity.

2. The backfitting algorithm can also be used to fit *semiparametric* models (Green and Yandell 1985; Stone 1986) of the form

$$E(y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + f_{j+1}(x_{j+1}) + \cdots + f_p(x_p). \qquad (6)$$

Whenever a linear term is encountered in the backfitting loop, a simple linear fit is used. This also allows categorical covariates, as well as categorical-smooth interactions [see Hastie (1986) for details]. In these cases it is more efficient to lump all of the linear fits together and estimate all of the coefficients simultaneously in a multiple regression, as a single big step in the iterative process.

3. It is useful to study the distributional version of the model. Under mild regularity assumptions, a unique L^2 additive approximation to $E(Y \mid \mathbf{X})$ of the form $\alpha + \Sigma f_j(X_j)$ exists. The backfitting algorithm, with $S(y \mid x_j)$ replaced by the conditional expectation $E(Y \mid X_j)$, converges to it (this result is well known in the approximation literature, e.g., see Deutsch 1983). Breiman and Friedman (1985) proved that the data algorithm is consistent for the additive approximation. See also Stone (1986) for rates of convergence.

4. When cubic spline smoothers are used in the backfitting algorithm, it converges to a solution of the penalized least squares problem:

$$\text{minimize} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} f_j(x_{ji}) \right)^2 + \sum_{j=1}^{p} \lambda_j \int [f_j''(x_j)]^2 \, dx_j \qquad (7)$$

(Hastie et al. 1987). The solution to this problem can be written in closed form, but involves matrix multiplication and inverses; the backfitting algorithm approximates the solution to any level of accuracy in $O(n)$ computations.

5. In practice we have had no convergence problems with the running lines smoothers, which run between 5 to 7 times faster than spline smoothers. For large data sets (1,000 observations) and many variables (8), this can make the difference between waiting 15 seconds for a fit versus a more tedious wait of $1\frac{1}{2}$ minutes.

2.3 The Local Scoring Algorithm for Generalized Additive Models

We give the algorithm for the logistic regression model (3), since the example we study in Section 4 uses this model. We have a sample $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$, where $y_i = 0$ or 1 and $\mathbf{x}_i$ is a vector of p covariates.

```
Initialization  $\hat{f}_j^0(x_j) = 0 \; \forall \; x_j$ and $\forall \; j$, $\hat{\alpha} = \text{logit}(\bar{y})$.
    Loop  over outer iteration counter $m$.
        $\hat{\eta}^m(\mathbf{x}_i) = \hat{\alpha} + \sum_{j=1}^{p} \hat{f}_j^m(x_{ji}),$
        $\hat{p}_i = \text{logit}^{-1}(\hat{\eta}^m(\mathbf{x}_i)),$
            $= \exp(\hat{\eta}^m(\mathbf{x}_i))/[1 + \exp(\hat{\eta}^m(\mathbf{x}_i))],$
        $z_i = \hat{\eta}^m(\mathbf{x}_i) + (y_i - \hat{p}_i)/[\hat{p}_i(1 - \hat{p}_i)],$
        $w_i = \hat{p}_i(1 - \hat{p}_i),$   $i = 1, \ldots, n.$
    Obtain $\hat{f}_j^{(m+1)}, j = 1, \ldots, p$ by applying the backfitting
    algorithm to the sequence $\mathbf{z}$ with covariates $\mathbf{X}$ and ob-
    servation weights $\mathbf{w}$
    until the deviance $D(\mathbf{y}, \hat{\mathbf{p}}) = -2\Sigma \, [y_i \ln \hat{p}_i + (1 - y_i)$
    $\ln(1 - \hat{p}_i)]$ converges.
```

The functions in Figures 4–9 (Sec. 4) were fitted using the local scoring procedure for 0–1 data. All smooths used were running lines with a span of 50%.

We make the following observations:

1. If the backfitting algorithm is replaced by the overall weighted least squares fit, this algorithm is identical to the iteratively reweighted least squares (IRLS) algorithm for solving the maximum likelihood equations in linear logistic regression [see McCullagh and Nelder (1983) and Green (1984) for accounts of the IRLS algorithm].

2. Landwehr et al. (1984) proposed smoothing partial residuals from the linear fit to detect nonlinearities. The partial residual for variable j and observation i is $r_{ij} = \hat{b}_j x_{ij} + (y_i - \hat{p}_i^0)/\hat{p}_i^0(1 - \hat{p}_i^0)$, where p_i^0 is the fit from the linear model. One can see that this is exactly the first step of the backfitting procedure within the local scoring algorithm, if we start with the linear fit. The local scoring procedure goes on to estimate all of the nonlinearities simultaneously.

3. In the local scoring algorithm for the GAM (2) we have $z_i = \hat{\eta}^m(\mathbf{x}_i) + (y_i - \hat{\mu}_i)\delta\eta/\delta\hat{\mu}$, $w_i = [\delta\eta/\delta\hat{\mu}]^2 \hat{v}_i^{-1}$, where $\eta = g(\mu)$ and $\hat{v}_i$ is the estimated variance of y_i (Hastie and Tibshirani 1986a).

4. In the Gaussian case the local scoring algorithm is exactly the backfitting algorithm for the additive model.

5. If cubic smoothing splines are used, then one can show that the local scoring algorithm converges to the additive function that maximizes the penalized log-likelihood criterion:

$$\text{maximize} \sum_i l(y_i, \mu(\mathbf{x}_i)) - \sum_j \lambda_j \int [f_j''(x_j)]^2 \, dx_j \qquad (8)$$

(Hastie and Tibshirani 1986b). O'Sullivan et al. (1986) and Green and Yandell (1985) derived similar results for the case of one smooth function plus some linear terms; such models are a special case of (8).

3. EXAMPLE 1: NONPARAMETRIC ANALYSIS OF COVARIANCE

The data in Figure 1 form part of a study on mildew control at Rothamsted Experimental Station (Jenkyn, Bainbridge, Dyke, and Todd 1979). Four treatments, 0 (none), 1, 2, and R, represent specific tridemorph spray frequencies. The response is the yield of grain in tons per hectare. The plots were arranged in one row, in nine blocks of four, with an extra plot at each end to achieve balance. This arrangement guaranteed that each treatment had as adjacent treatments every other treatment except itself [see Jenkyn et al. (1979) and Draper and Guttman (1980) for more details]. A number of analyses have appeared in

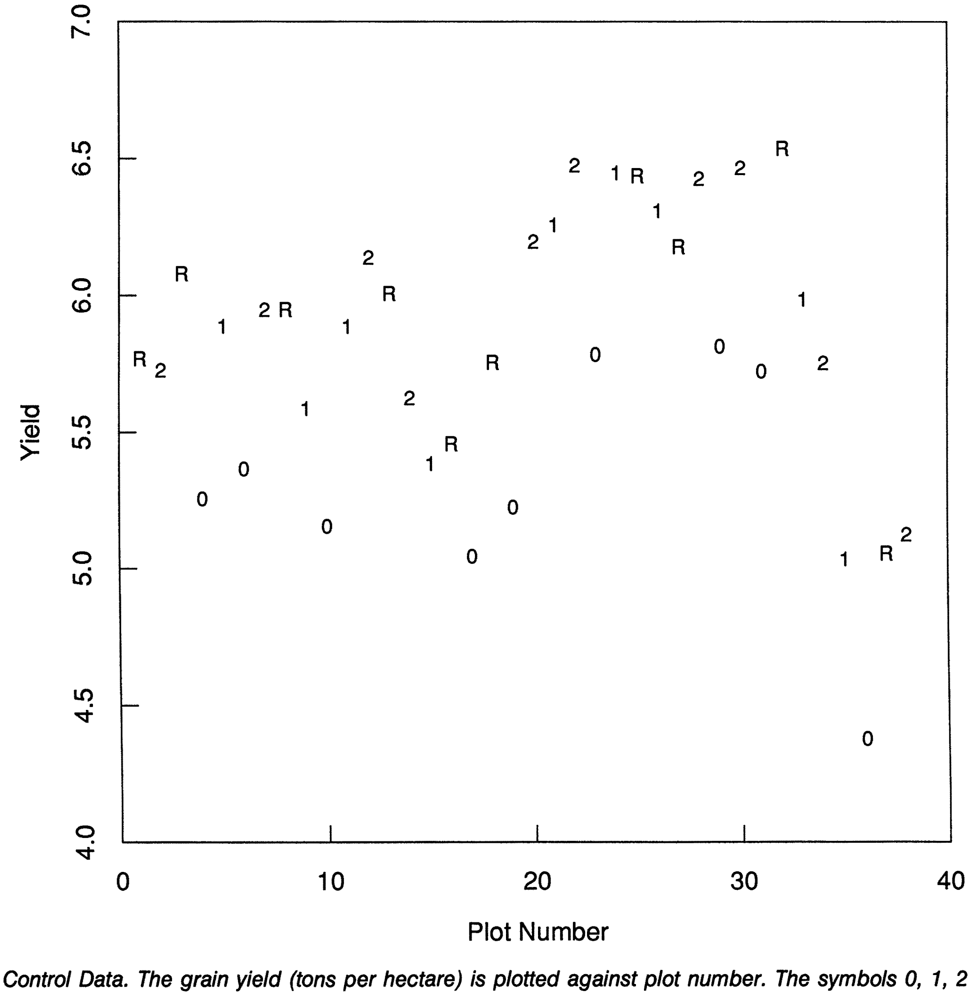


Figure 1. Mildew Control Data. The grain yield (tons per hectare) is plotted against plot number. The symbols 0, 1, 2, R correspond to the treatments used in the plots.

Table 1. ANOVA

Effect	*Sum of squares*	*Mean square*	*df*	*F*	*p*
	(without covariance adjustment)				
0 versus any treatment	2.64	2.64	1	13.2	<.01
Separate treatments	.09	.05	2	.20	Not significant
Error	6.81	.20	34		
Corrected total	9.52	.26	37		
	(with covariance adjustment)				
Nonlinear plot effect	6.31	.49	13	40.4	<.001
0 versus treatment	2.86	2.86	1	237.9	<.001
Separate treatments	.12	.06	2	5.5	<.05
Error	.24	.01	21		
Error (linear adjustment)	6.79	.21	33		
Corrected total	9.52	.26	37		

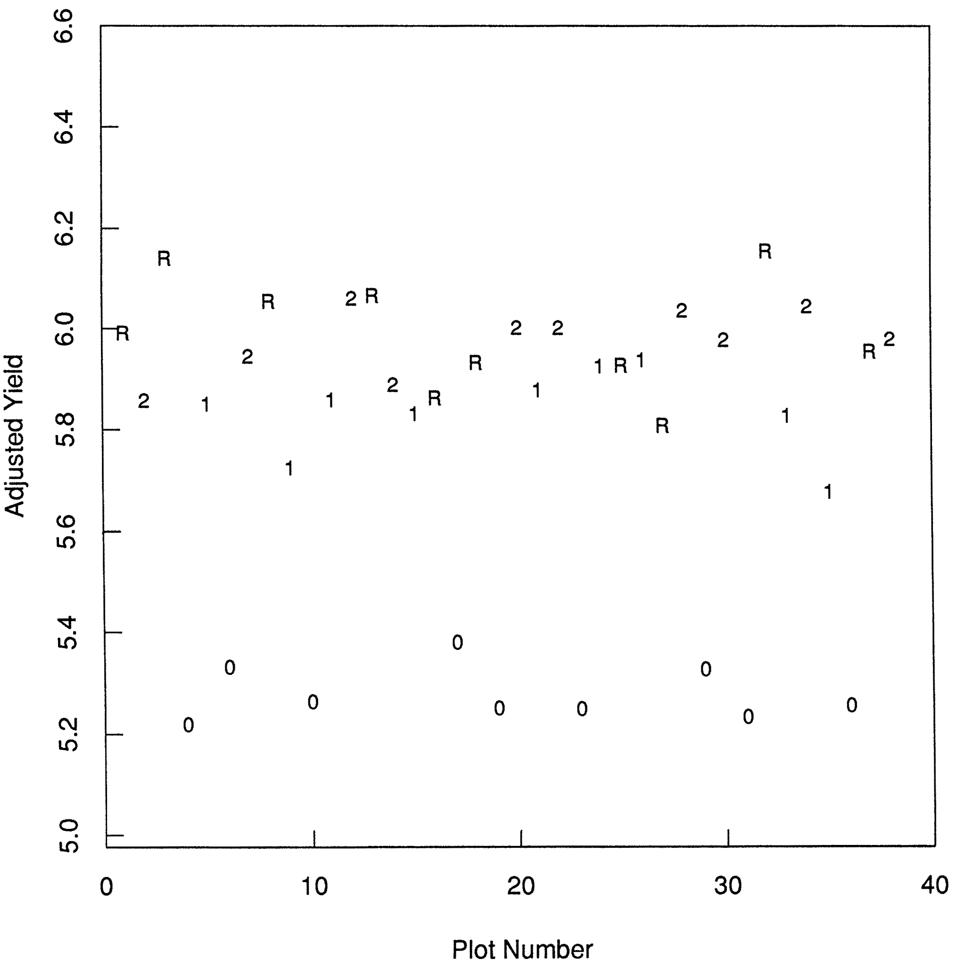
NOTE: 0 versus any treatment lumps all of the treatments together; separate treatments allows for four separate treatment effects.

the literature; we present ours and then follow it with a discussion of some of the others.

The main idea of any analysis of the treatment differences is to take into account the blocking effect and the effects of neighboring plots in this one-dimensional layout. Figure 1 suggests that there is a nonlinear plot effect, with a small amount of variation about it due to treatments. We, therefore, fit the analysis of covariance model

$$Y_i = \beta_0 + \beta_1 T_1(i) + \beta_2 T_2(i) + \beta_R T_R(i) + f(i) + \varepsilon_i, \tag{9}$$

where $T_j(i)$ are dummy variables such that $T_j(i) = 1$ if plot i received treatment j, else 0. The term $f(\mathrm{i})$ is a general



Adjusted Yield by Plot Number. The data are as in Figure 1, except that the estimated plot effect has been removed. The no-roup is clearly separated from the rest. The treatment effects are the averages of these adjusted yields.

function of the plot number itself. The model was fitted using the backfitting procedure, and the first few steps can be viewed as follows:

1. The treatment means are calculated, and the data is centered about them.

2. These residuals are smoothed against the plot number. In our case we used a small span of 10%, which amounts to using three observations in any neighborhood: the point in question and its two neighbors. This is also in accordance with some of the other methods to be discussed. The resulting smooth is the first estimate of f, which is then subtracted from the original (uncentered) data.

3. The treatment means of the *covariate adjusted* yield are then calculated. The data is once again centered, and the process is continued.

Table 1 summarizes the results. The first part of the table gives the usual ANOVA results for a one-way layout, ignoring the plot effect. Even in this case, the treatment effect is significant. We can see this by inspection, since the no-treatment group (0) in Figure 1 has markedly lower yields. The differences among treatments 1, 2, and R, however, are not significant. We need to use the analysis of covariance to sharpen these contrasts. The second part of Table 1 is the nonparametric analysis of covariance, in which we compare the adjusted treatment effects. The error term is reduced by a factor of 28 from 6.81 (no adjustment) to .24. As would be expected for these data, the error after linear covariate adjustment (6.79) hardly differs from that for no adjustment. There is a significant difference among the three treatments; Figure 2 shows the plot-adjusted yields, and it seems evident that treatment 1 lies below 2 and R. Figure 3 gives the estimated function f, together with the partial residuals obtained by subtracting the treatment effects from the yield.

The function f is in fact the smooth of these residuals. Also plotted are two curves corresponding to ±2 standard deviations from the fitted values. We describe the estimation of these curves later, as well as the *degrees of freedom* of the fitted smooth, which is estimated to be 13. The coefficient $\hat{\beta}_0$ is estimated to be 5.28, and the three treatment effects together with their estimated covariance matrix are given by

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} .55 \\ .70 \\ .71 \end{bmatrix}, \quad \widehat{\text{cov}}(\hat{\beta}) = \begin{bmatrix} .0031 & & \\ .0016 & .0031 & \\ .0015 & .0016 & .0030 \end{bmatrix}.$$

A crude test of treatment differences shows that Treat-ments 2 and R are both significantly different from Treatment 1, but not from each other.

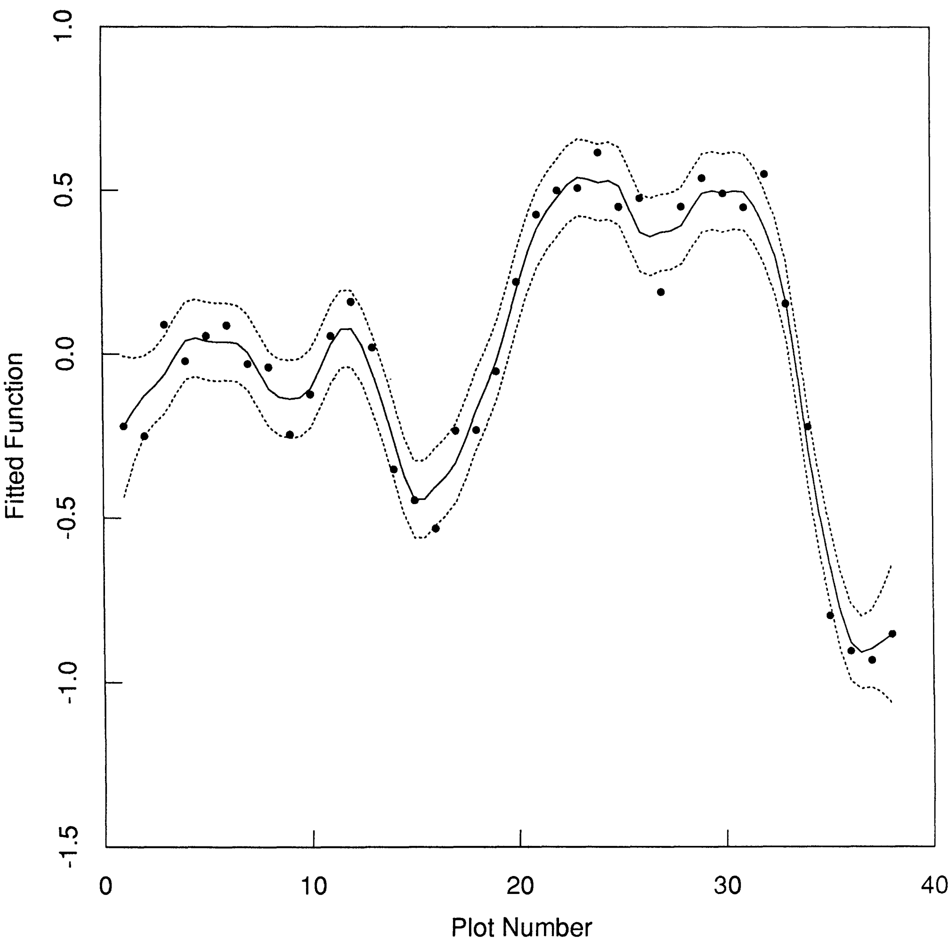


Figure 3. Estimated Plot Effect. The points plotted are the partial residuals for yield, having removed the treatment effects. The curve f(plot) is the "smooth" of these residuals. The dotted curves represent f(plot) ± 2SD[f(plot)].

These data have appeared in the literature before, as well as the analyses of data with similar designs:

1. A method proposed by Papadakis (1937) first corrects the yield for treatment effects. These residuals are then averaged over adjacent units, and the resulting *smooth* is used as a covariate in an analysis of covariance. This method can be seen as a variant of the backfitting procedure. The first three steps are identical to ours here. Then instead of smoothing the partial residuals against plot a second time, he effectively regressed it linearly and continued with the backfitting algorithm until convergence. In retrospect his intuitive procedure was one of the first estimation techniques for nonparametric additive models.

2. Atkinson (1969) compared the Papadakis procedure to a method proposed by Williams (1952). Here the errors are assumed to follow a first-order autogressive series, and the terms in the model are estimated by maximum likelihood. Atkinson shows that under this model assumption, the Papadakis estimates are very similar to the maximum likelihood estimates (MLE's). It follows from the previous paragraph that our estimates will be similar to the MLE's as well.

3. Draper and Guttman (1980) analyzed these data using a number of variations of the linear model. In addition to a standard block design, they fitted the model $\mathbf{y} = \mathbf{G}\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. Here $\mathbf{Z}$ represents a design matrix for the block effects and $\mathbf{X}$ a design for the treatment effects, and $\mathbf{G}$ creates an overlap of neighboring treatments. Their best fitting model produced a residual sum of squares (RSS) of .63 using 13 parameters (and based on 36 observations). They also reported on the analyses of Jenkyn et al. (1979), who fitted a complicated model that included four Fourier terms in the plot number. Their RSS was .25 using 24 parameters; our final fit has an RSS of .24 using 17 "parameters" and two extra observations. See also Draper and Faraggi (1985) for a synthesis of the various methods.

4. Steinberg (1984, personal communication), along the lines of Papadakis, fitted a model similar to (9) by using a cubic smoothing spline in the smoothing step. In demonstrating that these procedures are backfitting algorithms, we have shown that they are in fact also estimating an additive model as in (9).

5. Green (1985) and Denby (1986) considered a penalized least squares criterion for this problem, analogous to (7). They derived explicit solutions for $\boldsymbol{\beta}$ and $\hat{\mathbf{f}}$, namely $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t(\mathbf{I} - \mathbf{S})\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{I} - \mathbf{S})\mathbf{y}$ and $\mathbf{f} = \mathbf{S}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. Here

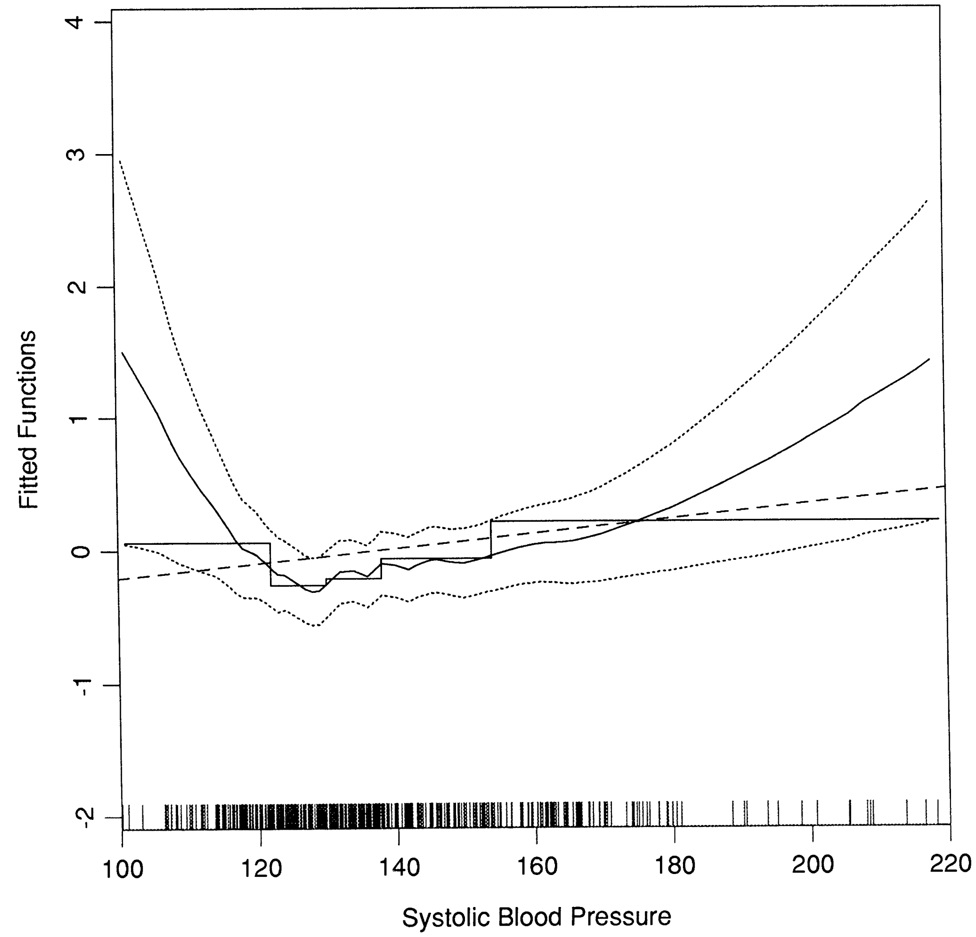


Figure 4. The estimated contribution $\hat{f}_1(x_1)$ of SBP to logit $(\hat{p}(\mathbf{x}))$ is the solid bold curve. The two dotted curves are $\hat{f}_1(x_1) \pm 2SD[\hat{f}_1(x_1)]$. The step function gives estimates of the terms β_{1k} in the model (11); the dashed straight line is the usual linear logistic regression estimate. The vertical bars at the base of the plot represent the marginal distribution of SBP.

X denotes the entire design matrix and **S** is a cubic spline smoother matrix. These equations hold for any linear smoother **S** and thus we could have avoided backfitting in this problem. (This simplification occurs because there is only one nonparametric term in the model.) Green and Denby also related these ideas to generalized least squares.

Inference

We have produced *degrees of freedom* (df) and *confidence curves* for our estimated functions. We give some motivation here and refer the reader to Hastie et al. (1987) for further details of the df calculations.

The fixed span smoothers we use are linear. This means that the vector of fitted values $\hat{\mathbf{f}}$ can be written $\hat{\mathbf{f}} = \mathbf{Sy}$, where $\mathbf{S}$ is an $n \times n$ smoother matrix not depending on $\mathbf{y}$. If we assume that the conditional variance of $\mathbf{y}$ is $\sigma^2\mathbf{I}$, then $\mathrm{cov}(\hat{\mathbf{f}}) = \sigma^2\mathbf{SS}^t$. From this and an estimate of σ^2 (based on the RSS) we obtain the variances of each term $\hat{f}(x_i)$. The estimates produced by the backfitting procedure are also linear in $\mathbf{y}$. Thus $\hat{\mathbf{f}}_1 = \mathbf{G}_1\mathbf{y}$, $\hat{\mathbf{f}}_2 = \mathbf{G}_2\mathbf{y}$, and so on, and $\hat{\boldsymbol{\eta}} = \hat{\alpha}\mathbf{1} + \hat{\mathbf{f}}_1 + \hat{\mathbf{f}}_2 + \cdots + \hat{\mathbf{f}}_p = \mathbf{G}_{\eta}\mathbf{y}$, where $\hat{\boldsymbol{\eta}}$ is the vector of fitted values, and $\mathbf{G}_i$ and $\mathbf{G}_\eta$ are matrix compounds of the individual smoothing matrices $\mathbf{S}_i$. So we can readily derive variance estimates for each function estimated by the backfitting algorithm if we know the relevant matrices. In practice we obtain these matrices by applying the algorithm to a sequence of response vectors $\mathbf{e}_1 = (1, 0, 0, \ldots, 0)^t$, $\mathbf{e}_2 = (0, 1, 0, \ldots, 0)^t, \ldots, \mathbf{e}_n = (0, 0, \ldots, 1)^t$. It is clear that the matrix concatenation of the outputs give the relevant matrices; for example, if $\mathbf{g}_{1i} = \mathbf{G}_1\mathbf{e}_i$ is the estimated function for covariate 1 using the vector $\mathbf{e}_i$ in the backfitting algorithm, then $[\mathbf{g}_{11} : \mathbf{g}_{12} : \cdots : \mathbf{g}_{1n}] = \mathbf{G}_1[\mathbf{e}_1 : \mathbf{e}_2 : \cdots : \mathbf{e}_n) = \mathbf{G}_1\mathbf{I} = \mathbf{G}_1$. The approximate confidence curves for $\hat{\mathbf{f}}_j$ are thus given by $\hat{\mathbf{f}}_j \pm 2\mathrm{SD}(\hat{\mathbf{f}}_j)$, where $\mathrm{SD}(\hat{\mathbf{f}}_j) = \hat{\sigma}\,\mathrm{diag}(\mathbf{G}_j\mathbf{G}_j^t)^{1/2}$. These computations take $O(n^2)$ operations for the GAM; for a single nonparametric function (plus linear effects in the other covariates) one can get away with $O(n)$ operations. The same is true if spline smoothers are used. Note that the smoothers are linear only if the spans are chosen a priori.

We define the df of a fit $\hat{\boldsymbol{\eta}}$ by drawing analogies to the linear regression model. Suppose that $y_i = g(\mathbf{x}_i) + \varepsilon_i$ with ε_i iid $(0, \sigma^2)$, $\hat{\eta}_i$ estimates $g(\mathbf{x}_i)$, and RSS is the residual sum of squares about $\hat{\boldsymbol{\eta}}$. Then one can show that $E(\mathrm{RSS}) = (n - \mathrm{tr}[2\mathbf{G}_\eta - \mathbf{G}_\eta{}^t\mathbf{G}_\eta])\sigma^2$ + positive bias term. We define df $= \mathrm{tr}[2\mathbf{G}_\eta - \mathbf{G}_\eta{}^t\mathbf{G}_\eta]$; for the linear model $\mathbf{G}_\eta$ is the "hat" matrix $\mathbf{H}$ and df equals p, the rank of the design matrix. Cleveland (1979) used this definition, calling it the "effective number of parameters." This definition is especially useful when comparing fits, for then the change in df gives the expected change in the RSS if we assume that the biases are the same. [See Cleveland (1979), Cleveland and Devlin (1986), and Hastie et al. (1987) for further

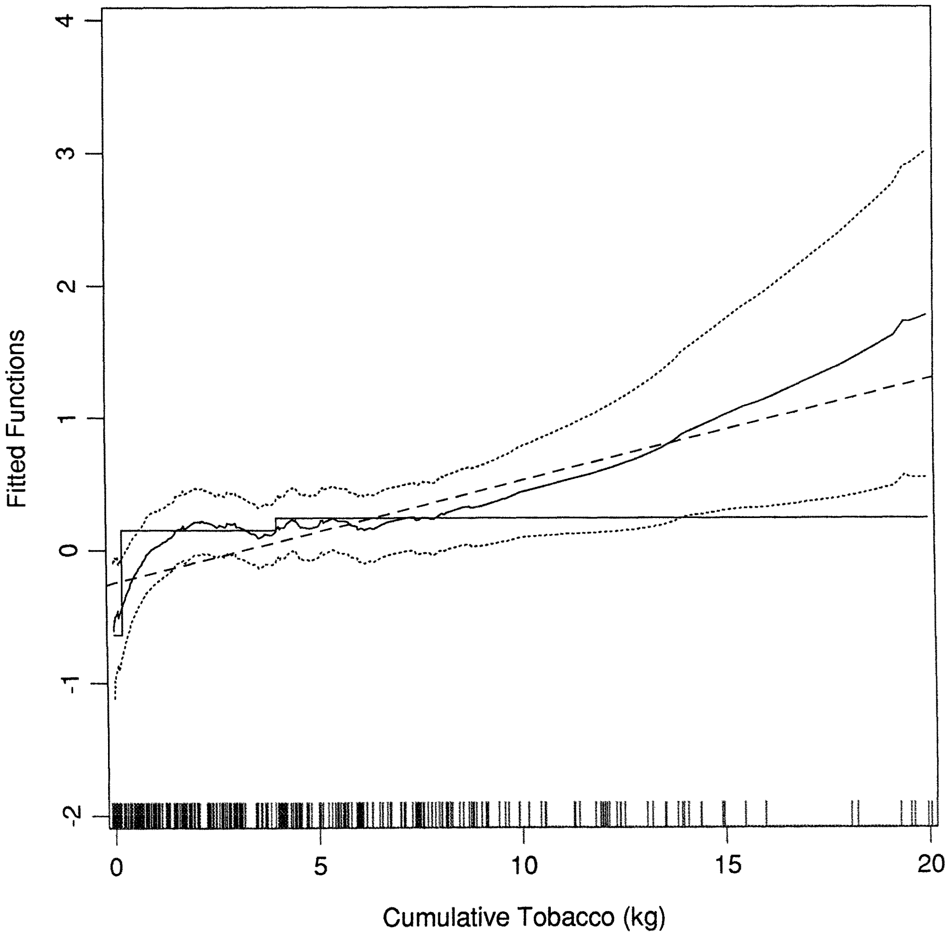


Figure 5. The estimated contribution $\hat{f}_2(x_2)$ of cumulative tobacco to logit(p(x)). (See Fig. 4 for notation.)

details on this and the approximate χ^2 distribution of the RSS.] In practice it appears that we can approximate df by the sum of the df_i of each individual smoother matrix (Hastie et al. 1987). This is preferable in the case of the running line smoother, for which it can be shown that $\text{df}_i = \text{tr}(\mathbf{S}_i)$; these are $O(n)$ to compute, whereas df is $O(n^2)$.

4. EXAMPLE 2: NONPARAMETRIC LOGISTIC REGRESSION

The data in this example are a subset of the Coronary Risk Factor Study (CORIS) baseline survey, carried out in three rural areas of the Western Cape, South Africa. This part of the study aims to identify and establish the intensity of ischemic heart disease (IHD) risk factors in this high incidence region. The baseline study is to be followed by a two-level intervention program (Rossouw et al. 1983). We analyze the data for the 3,357 white males between the ages 15 and 64 and concentrate on risk factors for myocardial infarction (MI). The overall prevalence of MI by 1979 was 5.13% for this group.

Let the binary variable y_i denote the presence or absence of MI for observation i. Initially a large number of possible risk factors and functions thereof were considered; by using stepwise logistic regression techniques these were later reduced to a set of possibly significant factors. We used this set as the starting point in our analysis. Since these people represent the entire population of a particular area, we view our analysis as an easily interpreted model of the risk pattern for this area. Tests and confidence regions will simply be indicators of where our model succeeds or falls short in describing this reality.

The set of risk factors are as follows:

1. *Systolic blood pressure* (SBP).
2. *Cumulative tobacco* (in kg) attempts to measure the total tobacco consumed in the subject's lifetime and is simply the average per day multiplied by the period of use.
3. *Cholesterol ratio* (ChR) is defined as (Total − HDL)/HDL, where Total is the overall serum cholesterol measurement, HDL is the amount of high-density lipoprotein, and LDL is the amount of low-density lipoprotein (Total − HDL = LDL + other lipoproteins). It is fairly well accepted that the "good" HDL tends to counteract the "bad" LDL, and so this ratio becomes relevant.
4. *Type A* is a measure of psychosocial stress, as measured by the self-administered Bortner Scale.
5. *Age.*
6. *Total energy* is a measure of the total energy expended in leisure-time and occupational activities.
7. *Family history* is a 0–1 variable, where a 1 indicates that a family member of the subject has had heart disease.

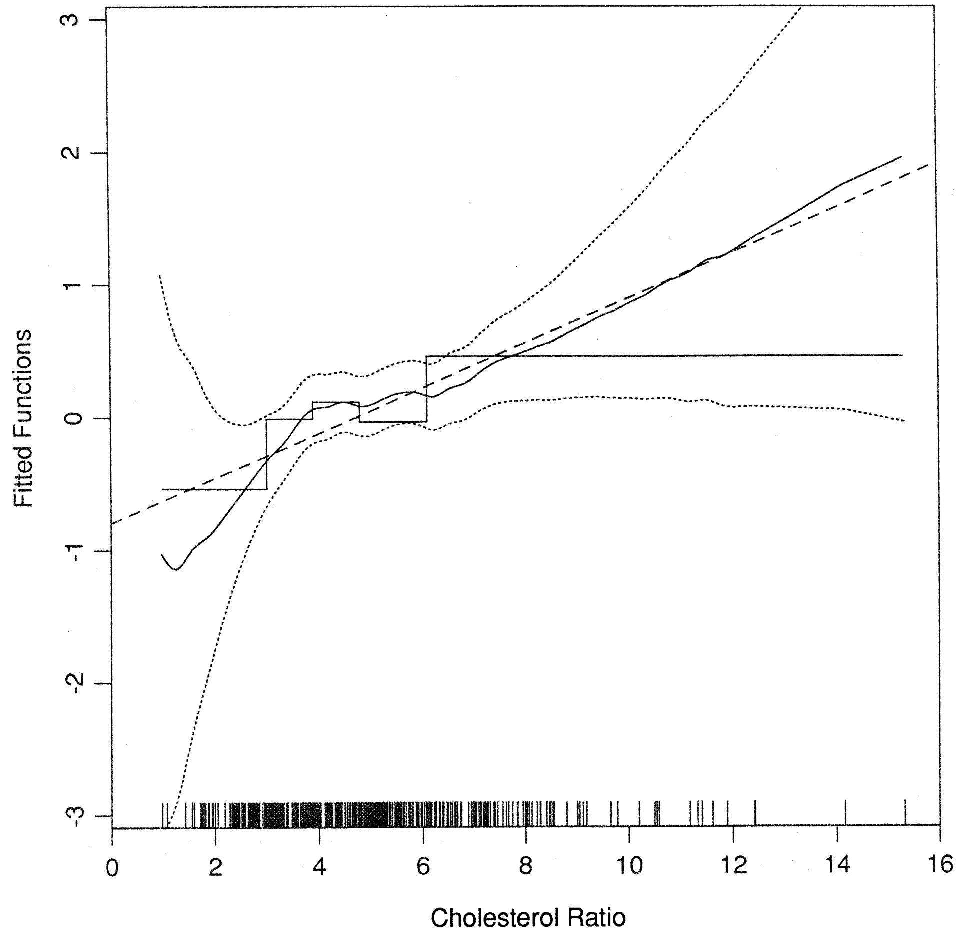


*Figure 6. The Estimated Contribution $\hat{f}_3(x_3)$ of Cholesterol Ratio to logit(p(**x**)). (See Fig. 4 for notation.)*

In our analysis we used the 162 cases and sampled roughly double the number (303) from the "controls." As pointed out by Anderson (1972), such sampling still allows us to model the contribution of each risk factor to the log-odds of prevalence consistently; the overall prevalence (the constant term) is, however, distorted.

We used the local scoring procedure to estimate the model

$$\log[p(\mathbf{x})/(1 - p(\mathbf{x}))] = \alpha + \sum f_j(x_j), \tag{10}$$

where x_j ($j = 1, \ldots, 7$) denote the seven covariates. A span of .5 was used for all the covariates except family history, where we fit a constant. The fitted functions are given in Figures 4–9, together with the ±2SD curves. The estimated coefficient for family history is 1.01. The analysis of deviance (ANODEV), shown in Table 2, summarizes the contribution of each variable to the fitted model. Each entry in the table corresponds to the increase in the deviance as a result of the exclusion of that term from the full model. All of the variables appear to be important. We tested for nonlinearity in each of the variables by forcing their terms, one at a time, to be linear. (Details are not given in Table 2.) SBP, cumulative tobacco, and Type A were significantly nonlinear.

The plots also contain for comparison the MLE's for the linear terms in the model $\log[p(\mathbf{x})/(1 - p(\mathbf{x}))] = \beta_0 + \boldsymbol{\beta}'\mathbf{x}$. These are represented by dashed lines. The deviance for this model is 463.4 with 8 parameters or df in the estimate. The estimated additive model has a deviance of 440.7 for 17.7 df, a drop of 22.7 for 9.7 df. Figures 4–6 also contain step functions. Each continuous variable is divided into a suitable manner of categories. In our case we chose the quintiles; the variable cumulative tobacco was broken into three categories since 40% of the observations recorded a value of 0.

A separate constant is estimated for each category in the model

$$\begin{aligned}\log p(\mathbf{x})/(1 - p(\mathbf{x})) &= \beta + \beta_{11}I_{11}(x_1) + \beta_{12}I_{12}(x_1) + \cdots + \beta_{15}I_{15}(x_1)\\ &\quad + \beta_{21}I_{21}(x_2) + \cdots + \beta_{75}I_{75}(x_7),\end{aligned} \tag{11}$$

where I_{jk} is an indicator variable for category k of variable j. The step function model has a deviance of 472.9 for 24 df, a worse fit than the linear model with three times the number of parameters! This is not too surprising, since we notice that a number of the nonparametric fits appear fairly linear, and the constants model will pick up bias in these cases. We note, in addition, that it is not as easy to spot nonlinearities by using step functions as it is with the nonparametric curves. It is also clear that category choice can play an important role in the constants model. We have omitted the step functions in Figures 7–9, since they are not very illuminating and tend to clutter the picture. A row of dots is plotted at the base of each curve; these represent the occurrence of data points, although the frequency is not represented. We note that the large widths of the confidence curves at the upper range in Figures 4 and 5, for example, are mainly due to the sparseness and outlying nature of the observations in that region.

We now provide some interpretations and further analysis of the various effects.

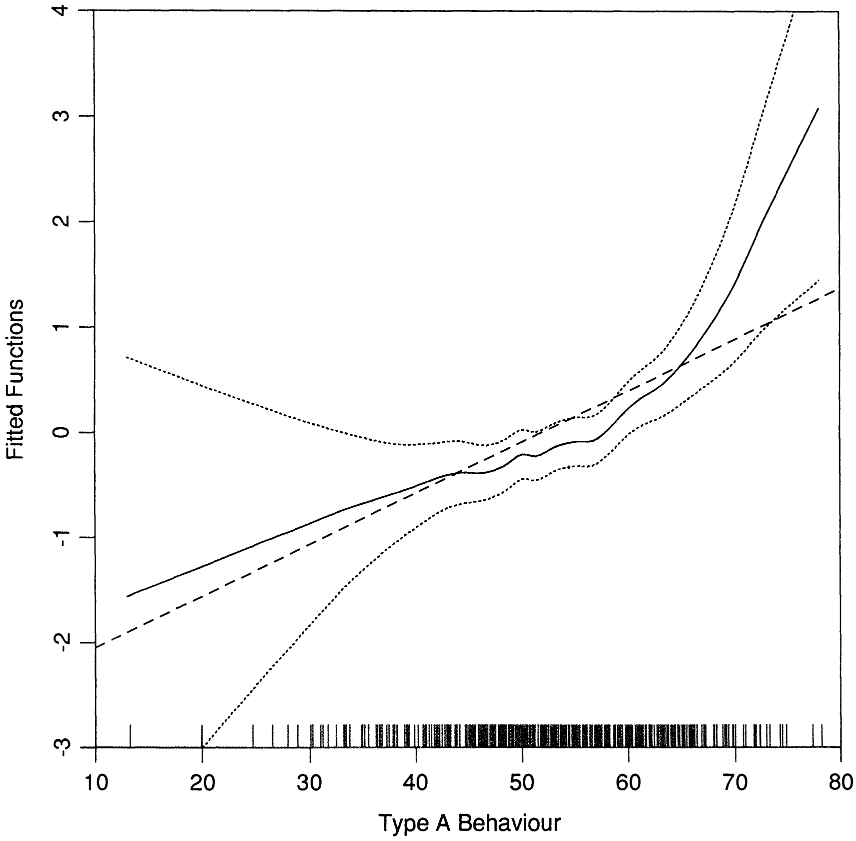


*Figure 7. The Estimated Contribution $\hat{f}_4(x_4)$ of Type A Behavior to logit(p(**x**)). (See Fig. 4 for notation.)*

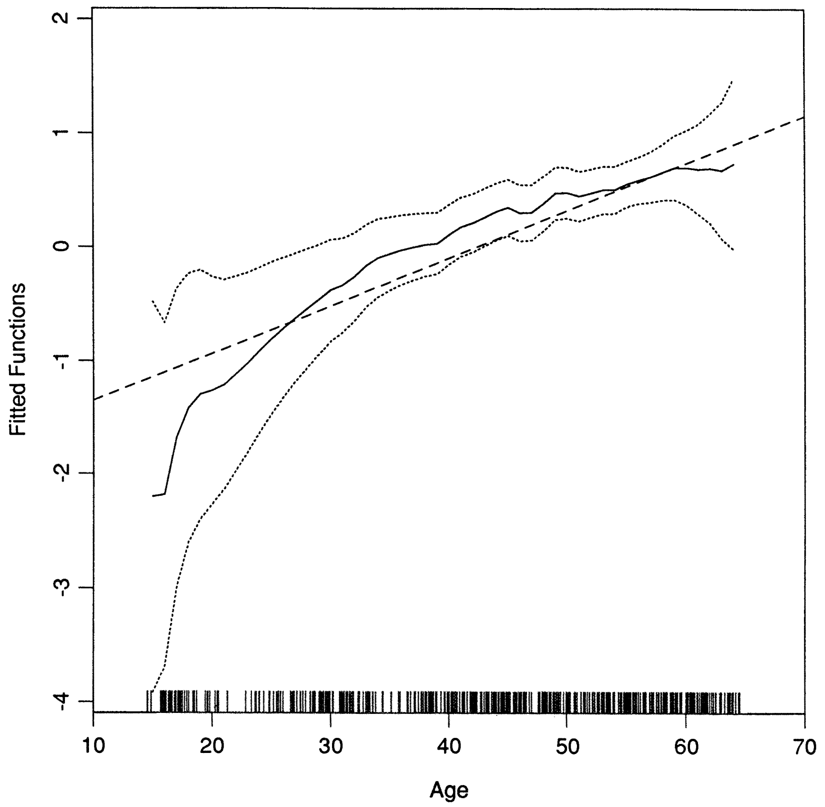


Figure 8. The Estimated Contribution $\hat{f}_5(x_5)$ of Age to logit(p(x)). (See Fig. 4 for notation.)

4.1 Systolic Blood Pressure

Exclusion of the linear term for systolic blood pressure (SBP) from the linear model causes the deviance to increase by 1.8 to 465.2. The deviance for the additive model, on the other hand, increases by 11.4 for 2.8 df w is excluded. Thus the linear model is unable to significant effect for SBP. This is not surprising Figure 4 we see that the nonparametric estimate is U-shaped.

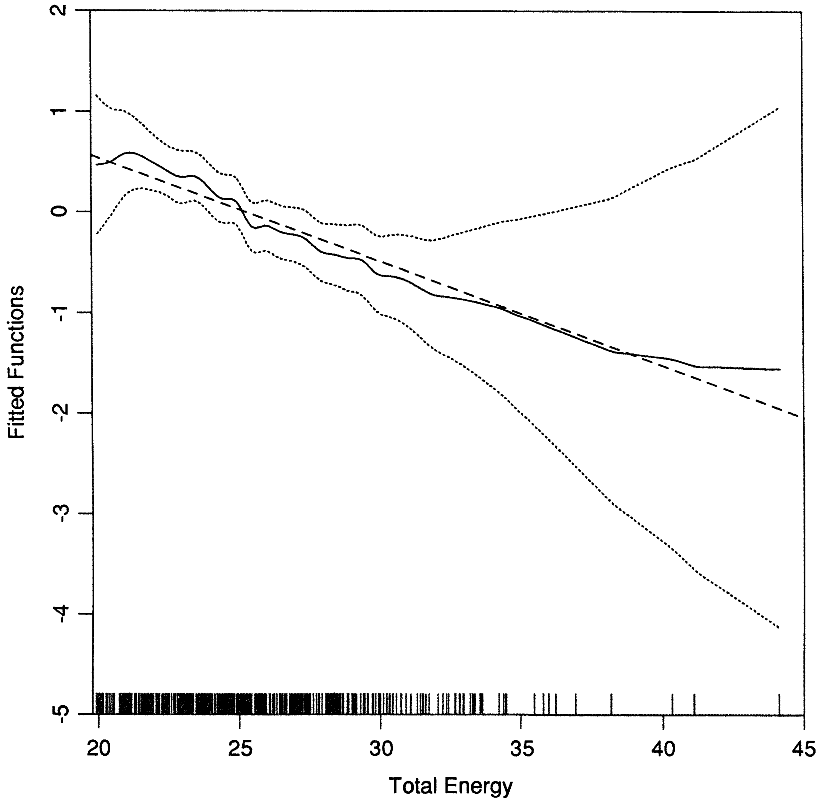


Figure 9. The Estimated Contribution $\hat{f}_6(x_6)$ of Total Energy to logit(p(x)). (See Fig. 4 for notation.)

Table 2. ANODEV

Additive predictor logit p(**x**)	df	Residual deviance	Effect deviance	Effect df
Linear in all 7 covariates	8	463.4		
Excluding SBP	7	465.2	1.8	1
Step functions for all 7 covariates	24	472.9		
Full additive model in 7 covariates	17.7	440.7		
Excluding SBP	14.9	452.3	11.6	2.8
Excluding cumulative tobacco	15.2	448.2	7.5	2.5
Excluding ChR	14.9	453.0	12.3	2.8
Excluding Type A	15.0	460.1	19.4	2.7
Excluding age	15.2	455.4	14.7	2.5
Excluding total energy	15.1	456.9	16.2	2.6
Excluding family history	16.7	460.5	19.8	1

Figure 10 is a partial residual plot for SBP (one unduly large residual point was removed). Our partial residuals are a natural extension of those of Landwehr et al. (1984) as defined in Section 2: $r_{ij} = \hat{f}_j(x_{ij}) + (y_i - \hat{p}(\mathbf{x}_i))/[\hat{p}(\mathbf{x}_i)(1 - \hat{p}(\mathbf{x}_i))]$. Here $\hat{p}(\mathbf{x}_i)$ refers to a model with all of the terms in, including SBP. The point removed from Figure 10 has $\hat{p}(\mathbf{x}_i) = .015$ and $y_i = 1$, resulting in a partial residual of about 65 for all of the covariates! As in linear logistic regression, our model is not fully robust against such outliers; we say fully, since to first order, the term affects only those fits whose window it shares in the smoothing algorithm.

A robust version of the local scoring algorithm would be possible, using some of the ideas outlined in Green (1984); we have not yet implemented such a version.

It was discovered that 91 of the 465 people in this subset of the data were on treatment for high blood pressure. A possible explanation for the U-shaped curve is that people

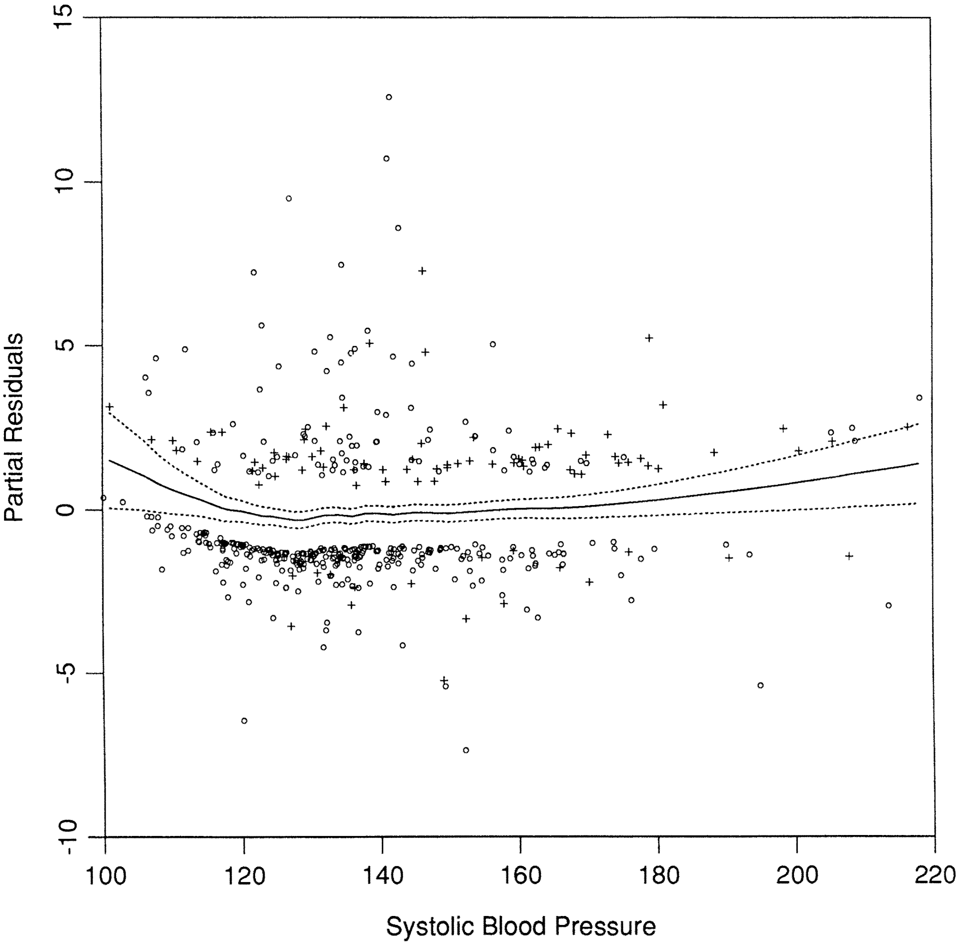


Figure 10. Partial Residual Plot for SBP. The solid curve is the estimated function, as in Figure 4, with the dashed curves the approximate confidence curves. The +'s refer to individuals on treatment for high blood pressure, and the O's refer to those not on treatment.

who have an MI subsequently go on treatment for high blood pressure, and their pressure drops. We have coded the treatment variable into the partial residual plots (+ is treatment, O no treatment), and note that the +'s tend to be above the curve. More conclusive evidence is provided in Figure 11, where we fit a separate SBP curve for the two groups. As expected, the no-treatment SBP curve is increasing; the treatment curve adds the contamination that results in the U shape. The fit is shown on the probability scale (as opposed to logit scale), since for this demonstration we have not adjusted for the other covariates. This also demonstrates the ability of the local scoring procedure in uncovering the regression in a scatterplot with little visual information. One would naturally include separate functions for the two groups in any further analysis of the joint behavior of the covariates.

4.2 Cholesterol Ratio

As in the foregoing examples, we can code any information in the partial residual plots. Figure 12 is a plot for cholesterol ratio (ChR), with the variable family history (+ = yes) encoded. There is a predominance of O's above the curve, and +'s below. Figure 13 shows a marginal fit for the two groups, and the interaction is clearly demonstrated. The original model attempted to capture the different sloped curves by two parallel curves with different intercepts.

4.3 Cumulative Tobacco

This curve (Fig. 5) shows a sharp increase in the logit of prevalence from people who have never smoked to people who have smoked at all; from then on the prevalence increases gradually with the amount smoked.

The other three curves behave as expected. The age curve shows a slight flattening out at higher ages. It has been suggested that younger people tend to survive the MI and are thus still alive to report the ordeal. The curve for energy supports the popular belief that exercise reduces the risk of a heart attack.

Note that in the systolic blood pressure and cholesterol analyses, interactions were detected by graphical means. We could just as well have modeled interaction terms by creating a new variable that is the product of existing variables, then estimating a smooth for it. For systolic blood pressure this amounts to fitting a separate smooth for each of the high blood pressure no-treatment/treatment groups.

4.4 Inference

The approximate 95% confidence bands plotted with the functions in this section are calculated in a fashion

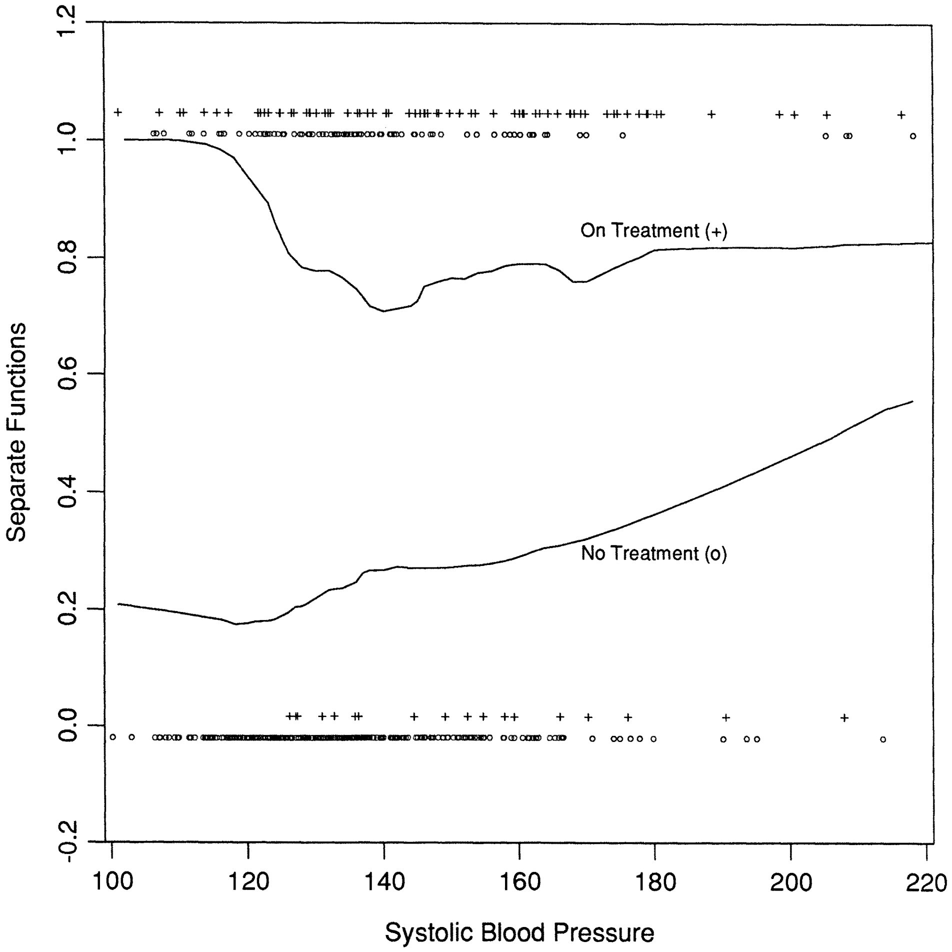


Figure 11. Separate Fits for SBP for Those on (+) and not on (O) Treatment for High Blood Pressure. No other terms are used in this model, so the functions can be plotted on the probability scale.

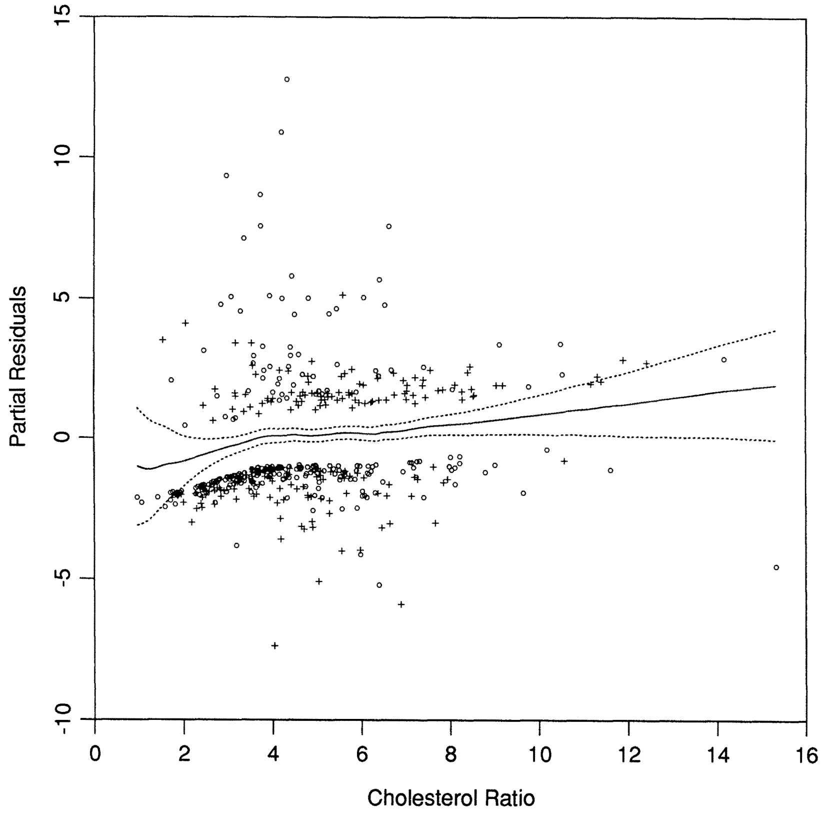


Figure 12. Partial Residual Plot for Cholesterol Ratio. The solid curve is the estimated function, as in Figure 6, with the dashed curves the approximate confidence curves. The + 's refer to individuals with family history of heart disease, and the O's refer to those without.

similar to those in Section 3. In fact, the final backfitting loop in the local scoring algorithm is applied to each of the dummy response vectors $\mathbf{e}_j$. This is nearly the same as in Section 3, except each observation i will have a weight w_i calculated from $\hat{p}(\mathbf{x}_i)$, the final fitted probability. For linear logistic regression, the estimated asymptotic co-

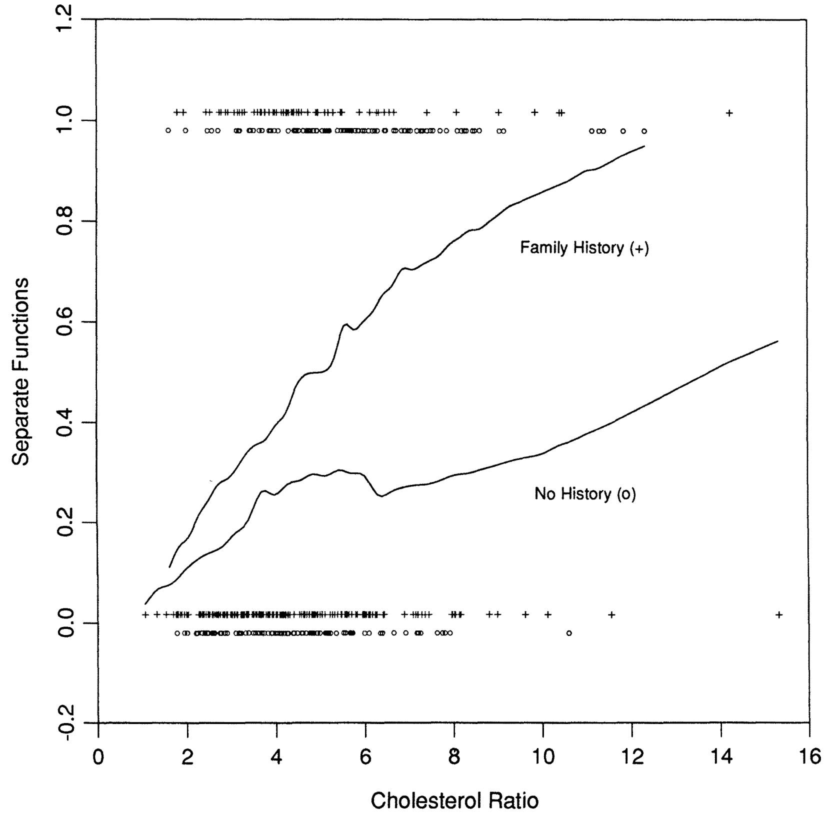


Figure 13. Separate Fits for Cholesterol Ratio for Those With (+) and Without (O) Family History of Heart Disease. No other terms are used in this model, so the functions can be plotted on the probability scale.

variance matrix of the parameter estimates $\hat{\boldsymbol{\beta}}$ is given by $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, where $\mathbf{W} = \operatorname{diag}(w_1, \ldots, w_n)$. Similarly, the estimated covariance matrix of the fitted values $\hat{\boldsymbol{\eta}}$ is given by $\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'$. We can intuitively motivate these well-known results. Assuming that the model is adequate and the usual maximum likelihood regularity conditions are satisfied, the adjusted dependent variable at the final iteration is asymptotically equivalent to

$$z = \operatorname{logit}[p(\mathbf{x})] + (y - p(\mathbf{x}))/[p(\mathbf{x})(1 - p(\mathbf{x}))], \quad (12)$$

where $p(\mathbf{x})$ is the true probability at $\mathbf{x}$. Now given $\mathbf{x}$, z has mean $\operatorname{logit}[p(\mathbf{x})]$ and variance $1/[p(\mathbf{x})(1 - p(\mathbf{x}))]$. The parameter estimates are obtained by a weighted linear regression of z on $\mathbf{x}$. The foregoing covariances are simply the appropriate expressions for such a weighted regression.

Since the final step of the local scoring procedure is a weighted additive model fit, the generalization of the results in Section 3 carry through. Suppose that $\mathbf{S}$ is the smoother matrix that results in some weighted estimate $\hat{\mathbf{f}}$ ($\hat{\mathbf{f}}$ might be the fitted values or the values for one of the nonparametric functions). If $\mathbf{W}^{-1}$ is the diagonal matrix of estimated variances of z, and thus $\mathbf{W}$ the weights used in the weighted fit, then $\widehat{\operatorname{cov}}(\hat{\mathbf{f}}) = \mathbf{S}\mathbf{W}^{-1}\mathbf{S}^t$. Degrees of freedom is defined as before and is based on the weighted sum of squared residuals for z (which is a first-order approximation to the deviance): $\mathrm{df} = \operatorname{tr}(2\mathbf{S} - \mathbf{S}^t\mathbf{W}\mathbf{S}\mathbf{W}^{-1})$. If $\mathbf{S}$ is the weighted running lines smoother, this reduces to $\operatorname{tr}(\mathbf{S})$.

5. DISCUSSION

GAM's provide a flexible method for identifying nonlinear covariate effects in a variety of modeling situations; notably the very situations in which it has become popular to use the generalized linear models. The additive models can be used in a data-analytic fashion to understand the effect of covariates and to test hypotheses about effects. A more conservative approach is to allow the nonparametric functions to suggest parametric transformations and then proceed with the usual linear analysis on the transformed variables.

A certain amount of theory already exists for these models, notably results on existence and uniqueness of best additive models in a theoretical setting, convergence of the algorithm in this setting, and consistency.

We can estimate the df of the terms in the model and calculate approximate standard deviation curves for the fitted functions. This theory is still developing, with a variety of relevant questions remaining to be answered. One such example is the effect of dependence between two covariates on the fitting algorithm, the standard deviations, and the df of the fit. Partial progress has been made in this direction.

We have illustrated the procedures on two types of problems here; Hastie and Tibshirani (1986a) analyzed survival data and used an extension of the ideas to estimate, in addition, a nonparametric link function. Recent extensions include two-dimensional surface interaction terms in the models (Hastie 1986) and the additive proportional odds model for ordered categorical data (Hastie and Tibshirani, in press).

[*Received August 1985. Revised October 1986.*]

REFERENCES

Anderson, J. A. (1972), "Separate Sample Logistic Regression," *Biometrika*, 59, 19–35.

Atkinson, A. C. (1969), "The Use of Residuals as a Concomitant Variable," *Biometrika*, 56, 33–41.

Baker, R. J., and Nelder, J. A. (1978), *The GLIM System, Release 3*, distributed by National Algorithms Group, Oxford, U.K.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics*, New York: John Wiley.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation" (with discussion), *Journal of the American Statistical Association*, 80, 580–619.

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

Cleveland, W. S., and Devlin, S. J. (1986), "Locally Weighted Regressions: An Approach to Regression Analysis by Local Fitting," AT&T Bell Laboratories Statistical Research Report.

Cook, R. P., and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman & Hall.

Denby, L. (1986), "Smooth Regression Functions," AT&T Bell Laboratories Statistical Research Report 26.

Deutsch, F. (1983), "Von Neumann's Alternating Method: The Rate of Convergence," in *Approximation Theory IV*, eds. C. K. Chui, L. L. Schumaker, and J. D. Ward, New York: Academic Press, pp. 427–434.

Draper, N. E., and Faraggi, D. (1985), "Role of the Papadakis Estimator in One- and Two-Dimensional Field Trials," *Biometrika*, 72, 223–226.

Draper, N. E., and Guttman, I. (1980), "Incorporating Overlap Effects From Neighboring Units Into Response Surface Models," *Applied Statistics*, 29, 128–134.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

——— (1982), "Smoothing of Scatterplots," technical report (Orion 3), Stanford University, Statistics Dept.

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

——— (1985), "Linear Models for Field Trials, Smoothing, and Cross-Validation," *Biometrika*, 72, 527–538.

Green, P. J., and Yandell, B. (1985), "Semi-Parametric Generalized Linear Models," in *Proceedings of the GLIM 1985 Conference*, Springer-Verlag Lecture Notes in Statistics, 32.

Hastie, T. (1986), "Generalized Additive Models: A GAIM Analyst's Toolbox," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 41–47.

Hastie, T., and Tibshirani, R. (1986a), "Generalized Additive Models" (with discussion), *Statistical Science*, 1, 297–318.

——— (1986b), "Generalized Additive Models, Penalized Likelihood and Cubic Splines," submitted for publication.

Hastie, T., and Tibshirani, R. (in press), "Nonparametric Logistic and Proportional Odds Regression," *Applied Statistics*.

Hastie, T., Tibshirani, R., and Buja, A. (1987), "Linear Smoothers and Additive Models," AT&T Bell Laboratories Statistical Research Report.

Huber, P. J. (1985), "Projection Pursuit" (with discussion), *The Annals of Statistics*, 13, 435–475.

Jenkyn, J., Bainbridge, A., Dyke, G., and Todd, A. (1979), "An Investigation Into Interplot Interactions, in Experiments With Mildew on Barley, Using Balanced Designs," *Annals of Applied Biology*, 92, 11–28.

Landwehr, J. M. (1986), "Using Residual Plots to Detect Nonlinearity in Multiple Regression," AT&T Bell Laboratories Statistics Research Report, 15.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. (1984), "Graphical Methods for Assessing Logistic Regression Models" (with discussion), *Journal of the American Statistical Association*, 79, 61–63.

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

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Nelder, J. A., and Wedderburn, R. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Ser. A, 135, 370–384.

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.

Papadakis, J. (1937), "Méthode statistique pour des expériences sur champ," *Bulletin Institute Plantes á Salonique*, 23.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics*, 9, 705–724.

——— (1982), "Resistant Fits for Some Commonly Used Logistic Models With Medical Applications," *Biometrics*, 38, 485–498.

Rossouw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P., and Ferreira, J. (1983), "Coronary Risk Factor Screening in Three Rural Communities," *South African Medical Journal*, 64, 430–436.

Silverman, B. W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 47, 1–52.

Stone, C. J. (1986), "The Dimensionality Reduction Principal for Generalized Additive Models," *The Annals of Statistics*, 14, 590–606.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

Wahba, G., and Wold, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics, Part A—Theory and Methods*, 4, 1–7.

Williams, R. (1952), "Experimental Designs for Serially Correlated Observations," *Biometrika*, 39, 151–167.