Generalized Additive Models*

Trevor Hastie †
and
Robert Tibshirani ‡
Department of Statistics and Division of Biostatistics
Stanford University

12th May, 1995

Regression models play an important role in many applied settings, providing prediction and classification rules, and data analytic tools for understanding the interactive behaviour of different variables.

Although attractively simple, the traditional linear model often fails in these situations: in real life effects are generally not linear. This article describes flexible statistical methods that may be used to identify and characterize nonlinear regression effects. These methods are called "generalized additive models".

For example, a commonly used statistical model in medical research is the logistic regression model for binary data. Here we relate the mean of the binary response $\mu = P(y = 1)$ to the predictors via a linear regression model and the *logit* link function:

$$\log\left[\frac{\mu}{1-\mu}\right] = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p \tag{1}$$

*to appear in Encyclopedia of Statistical Sciences

†Department of Statistics, Sequoia Hall, Stanford University, Stanford California 94305; trevor@playfair.stanford.edu

‡On sabbatical leave from Department of Preventive Medicine and Biostatistics, and Department of Statistics, University of Toronto; tibs@playfair.stanford.edu; tibs@utstat.toronto.edu

The *additive* logistic regression model replaces each linear term by a more general functional form

$$\log\left[\frac{\mu}{1-\mu}\right] = \alpha + f_1(x_1) + \ldots + f_p(x_p) \tag{2}$$

where each f_j is an unspecified ("non-parametric") function. While the non-parametric form for the functions f_j makes the model more flexible, the additivity is retained and allows us to interpret the model in much the same way as before.

The additive logistic regression model is an example of a generalized additive model. In general the mean μ of a response y is related to an additive function of the predictors via a *link* function g:

$$g(\mu) = \alpha + f_1(x_1) + \ldots + f_p(x_p) \tag{3}$$

Other classical link functions and associated generalized additive models are:

- $g(\mu) = \mu = \sum_j f_j(x_j)$, the additive extension of the ordinary linear model;

- $g(\mu) = \log(\mu) = \sum_j f_j(x_j)$, a log-additive model for count (Poisson) data.

All three of these arise from exponential family sampling models, which in addition include the gamma and negative-binomial distributions. These families generate the well known class of generalized linear models (McCullagh & Nelder 1989), which are all extended in the same way to generalized additive models.

The functions f_j are estimated in a flexible manner, using an algorithm whose basic building block is a scatterplot smoother. The estimated function $\hat{f}_j(x_j)$ can then reveal possible nonlinearities in the effect of the x_j. Not all of the functions f_j need be nonlinear. We can easily mix in linear and other parametric forms with the nonlinear terms, a necessity when some of the variables are discrete factors. The nonlinear terms are not restricted to main effects either; we can have nonlinear components in two or more variables, or separate curves for each level of a discrete factor. Thus each of the following would qualify:

- $g(\mu) = X^t\beta + \alpha_k + f(z)$—a *semiparametric* model, where X is a vector of predictors to be modeled linearly, α_k the effect for the kth level of a discrete factor, and the effect of predictor z is modelled nonparametrically;
- $g(\mu) = f(x) + g_k(z)$ where again k indexes the levels of a factor, and thus creates an interaction term for the effect of k and z;
- $g(\mu) = f(x) + g(z, w)$ where g is a nonparametric function in two variables.

Additive models can replace linear models in most settings where the latter is appropriate; here are some examples:

- transformation models—ACE algorithm: $g(Y) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \varepsilon$
- censored survival data and the Cox model: $\lambda(x, t) = \lambda_0(t)e^{f_1(x_1)+f_2(x_2)+\cdots+f_p(x_p)}$
- resistant additive models via tapered likelihoods
- additive decomposition of time series: $Y_t = S_t + T_t + \varepsilon_t$ where S_t is a seasonal component and T_t a trend;
- varying coefficient models: $\eta(x, t) = \alpha(t) + x_1\beta_1(t) + x_2\beta_2(t)$ where given t, the model is linear, but the coefficients change with t.

In all these cases and many not listed we are replacing the traditional parametric components by more flexible nonparametric functions.

1 Smoothing Methods and Generalized Additive Models

In this section we describe a modular algorithm for fitting additive models and their generalizations. The building block is the scatterplot smoother for fitting nonlinear effects in a flexible way.

Suppose that we have a scatterplot of points (x_i, y_i) like that shown in figure 1. Here y is a response or outcome variable, and x is a predictor. We wish to fit a smooth curve $f(x)$ that summarizes the dependence of y

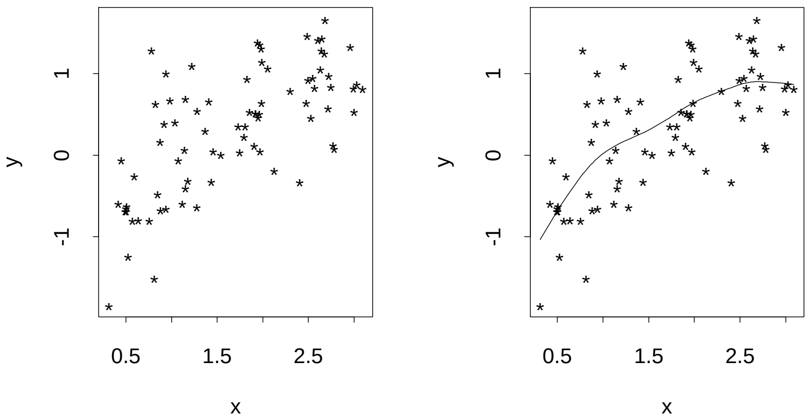

Figure 1: *Left panel shows a fictitious scatterplot of an outcome measure y plotted against a predictor x. In the right panel, a cubic smoothing spline has been added to describe the trend of y on x.*

on x. More formally, we want to fit the model $y = f(x) + \varepsilon$ where $f(x)$ is specified in a flexible way. If we were to find the curve that simply minimizes $\sum(y_i - f(x_i))^2$, the result would be an interpolating curve that would not be smooth at all.

The cubic smoothing spline overcomes this by imposing smoothness directly on $f(x)$. We seek the function $f(x)$ that minimizes

$$\sum(y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt \tag{4}$$

Notice that $\int f''(x)^2$ measures the "wiggliness" of the function f: linear fs have $\int f''(x)^2 = 0$, while non-linear fs produce values bigger than zero. λ is a non-negative smoothing parameter that must be chosen by the data analyst. It governs the tradeoff between the goodness of fit to the data (as measured by $\sum(y_i - f(x_i))^2$) and wiggliness of the function. Larger values of λ force f to be smoother. In fact the interpolating curve corresponds to $\lambda = 0$ at one extreme, and the straight line fit is the limit as $\lambda \to \infty$.

For any value of λ, the solution to (4) is a cubic spline, i.e., a piecewise cubic polynomial with pieces joined at the unique observed values of x in the dataset. Fast and stable numerical procedures are available for computation of the fitted curve. The right panel of figure 1 shows a cubic spline fit to the data.

What value of λ did we use in figure 1? In fact it is not convenient to express the desired smoothness of f in terms of λ, as the meaning of λ depends on the units of the prognostic factor x. Instead, it is possible to define an "effective number of parameters" or "degrees of freedom" of a cubic spline smoother, and then use a numerical search to determine the value of λ to yield this number. In figure 1 we chose the effective number of parameters to be 5. Roughly speaking, this means that the complexity of the curve is about the same as a polynomial regression of degrees 4. However, the cubic spline smoother "spreads out" its parameters in a more even manner, and hence is much more flexible than a polynomial regression. Note that the degrees of freedom of a smoother need not be an integer.

The above discussion tells how to fit a curve to a single predictor. With multiple predictors, if x_{ij} denotes the value of the jth predictor for the ith observation, we fit the additive model

$$\hat{y}_i = \sum_j f_j(x_{ij}) + \varepsilon \tag{5}$$

where (for simplicity) we have absorbed the constant into one of the functions. A criterion like (4) can be specified for this problem:

$$\sum_i (y_i - \sum_j f_j(x_{ij}))^2 + \sum_j \lambda_j \int f_j^{''}(t_j)^2 dt_j \tag{6}$$

and a simple iterative procedure exists for optimizing it and hence estimating the f_js. We apply a cubic spline smoother to the outcome $y_i - \sum_{j \neq k} \hat{f}_j(x_{ij})$ as a function of x_{ik}, for each predictor in turn. The process is continued until the estimates $\hat{f}_j$ stabilize. This procedure is known as "backfitting" and the resulting fit is analogous to a multiple regression for linear models. More formally this procedure can be shown (Buja, Hastie & Tibshirani 1989) to be a Gauss-Seidel algorithm for solving the following set of *estimating equations*:

$$\begin{aligned}
f_1(x_1) &= S_1(y - \bullet - f_2(x_2) - \cdots - f_p(x_p)) \\
f_2(x_2) &= S_2(y - f_1(x_1) - \bullet - \cdots - f_p(x_p)) \\
&\vdots \\
f_p(x_p) &= S_p(y - f_1(x_1) - f_2(x_2) - \cdots - \bullet)
\end{aligned}$$

where S_j is a smoothing spline operator for smoothing against the jth variable, and the $\bullet$s highlight the missing term in each row.

This same algorithm can accommodate other fitting methods in exactly the same way, by specifying appropriate operators S_j:

- other univariate regression smoothers such as local polynomial regression and kernel methods;
- linear regression operators yielding polynomial fits, piecewise constant fits, parametric spline fits, series and Fourier fits;
- more complicated operators such as surface smoothers for 2nd or higher order interactions or periodic smoothers for seasonal effects.

If we interpret the the elements $f_j(x_j)$, y, etc as vectors corresponding to the n samples, then the S_j will be $n \times n$ operator matrices like the *hat* matrices in linear regression (but not necessarily projections). The *df* for the jth term discussed earlier are intuitively defined as $df_j = \text{tr}(S_j)$ by analogy with linear regression, and this definition can be given a more rigorous justification.

For the logistic regression model and other generalized additive models, the appropriate criterion is a penalized log likelihood or a penalized log partial-likelihood. To maximize it, the backfitting procedure is used in conjunction with a maximum likelihood or maximum partial likelihood algorithm. The usual Newton-Raphson routine for maximizing log-likelihoods in these models can be cast in a IRLS (iteratively reweighted least squares) form. This involves a repeated weighted linear regression of a constructed response variable on the covariates: each regression yields a new value of the parameter estimates which give a new constructed variable, and the process is iterated. In the generalized additive model, the weighted linear regression is simply replaced by a weighted backfitting algorithm. We describe the algorithm in more detail for logistic regression below, and in more generality in chapter 6 of Hastie & Tibshirani (1990).

2 Example: Additive Logistic Regression

Probably the most widely used model in medical research is the logistic model for binary data. In this model the outcome y_i is 0 or 1, with 1 indicating an event (like death or relapse of a disease) and 0 indicating no event. We wish to model $p(y_i|x_{i1}, x_{i2}, \ldots x_{ip})$, the probability of an event given prognostic factors $x_{i1}, x_{i2}, \ldots x_{ip}$. The linear logistic model assumes that the log-odds are linear:

$$\log \frac{p(y_i|x_{i1}, \ldots x_{ip})}{1 - p(y_i|x_{i1}, \ldots x_{ip})} = \beta_0 + x_{i1}\beta_1 + \ldots x_{ip}\beta_p \tag{7}$$

The generalized additive logistic model assumes instead that

$$\log \frac{p(y_i|x_{i1}, \ldots x_{ip})}{1 - p(y_i|x_{i1}, \ldots x_{ip})} = \beta_0 + f_1(x_{i1}) + \ldots f_p(x_{ip}) \tag{8}$$

The functions $f_1, f_2, \ldots f_p$ are estimated by an algorithm like the one described earlier—backfitting within Newton-Raphson:

- Compute starting values: β_0^{old} and f_j^{old} and $\eta^{old} = \beta_0^{old} + \sum_j f_j^{old}(x_j)$ e.g. using linear logistic regression
- Iterate
  - construct an adjusted dependent variable

    $$z_i = \eta_i^{old} + \frac{(y_i - p_i^{old})}{p_i^{old}(1 - p_i^{old})}$$

  - construct weights $w_i = p_i^{old}(1 - p_i^{old})$
  - compute $\eta^{new} = A_w z$, the weighted additive model fit to z.
- Stop when the functions don't change.

He we study an example on the survival of children after cardiac surgery for heart defects, taken from Williams, Rebeyka, Tibshirani, Coles, Lightfoot, Freedom & Trusler (1990). The data was collected during for the period 1983-1988. A pre-operation warm-blood cardioplegia procedure, thought to improve chances for survival, was introduced in February 1988. This was not used on all of the children after February 1988, only on those for which it was thought appropriate and only by surgeons who chose to use the new procedure. The main question is whether the introduction of the warming procedure improved survival; the importance of risk factors age, weight and diagnostic category is also of interest.

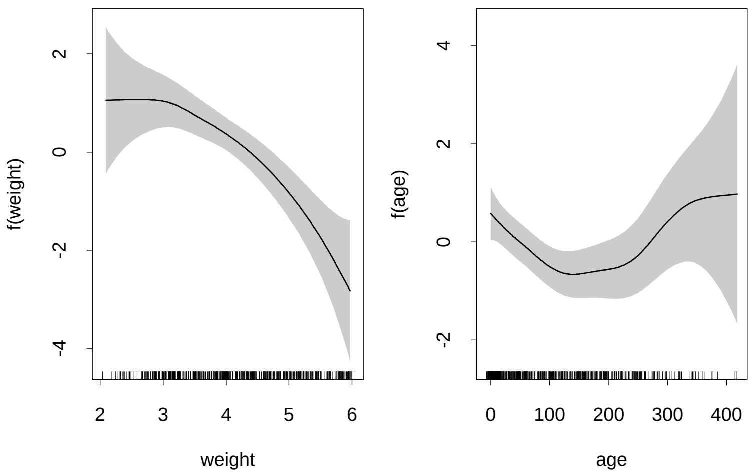

Figure 2: *Estimated functions for* weight *and* age *for warm cardioplegia data. The shaded region represents twice the pointwise asymptotic standard errors of the estimated curve.*

If the warming procedure was given in a randomized manner, we could simply focus on the post-February 1988 data and compare the survival of those who received the new procedure to those who did not. However allocation was not random so we can only try to assess the effectiveness of the warming procedure as it was applied. For this analysis, we use all of the data (1983–1988). To adjust for changes that might have occurred over the five-year period, we include the date of the operation as a covariate. However operation date is strongly confounded with the warming operation and thus a general nonparametric fit for date of operation might unduly remove some of the effect attributable to the warming procedure. To avoid this, we allow only a linear effect for operation date. Hence we must assume that any time trend is either a consistently increasing or decreasing trend.

We fit a generalized additive logistic model to the binary response *death,* with smooth terms for *age* and *weight,* a linear term for *operation date,* a categorical variable for *diagnosis,* and a binary variable for the *warming* operation. All the smooth terms are fitted with 4 degrees of freedom.

The resulting curves for *age* and *weight* are shown in figure 2. As one would expect, the highest risk is for the lighter babies, with a decreasing risk over 3 kg. Somewhat surprisingly, there seems to be a low risk age around

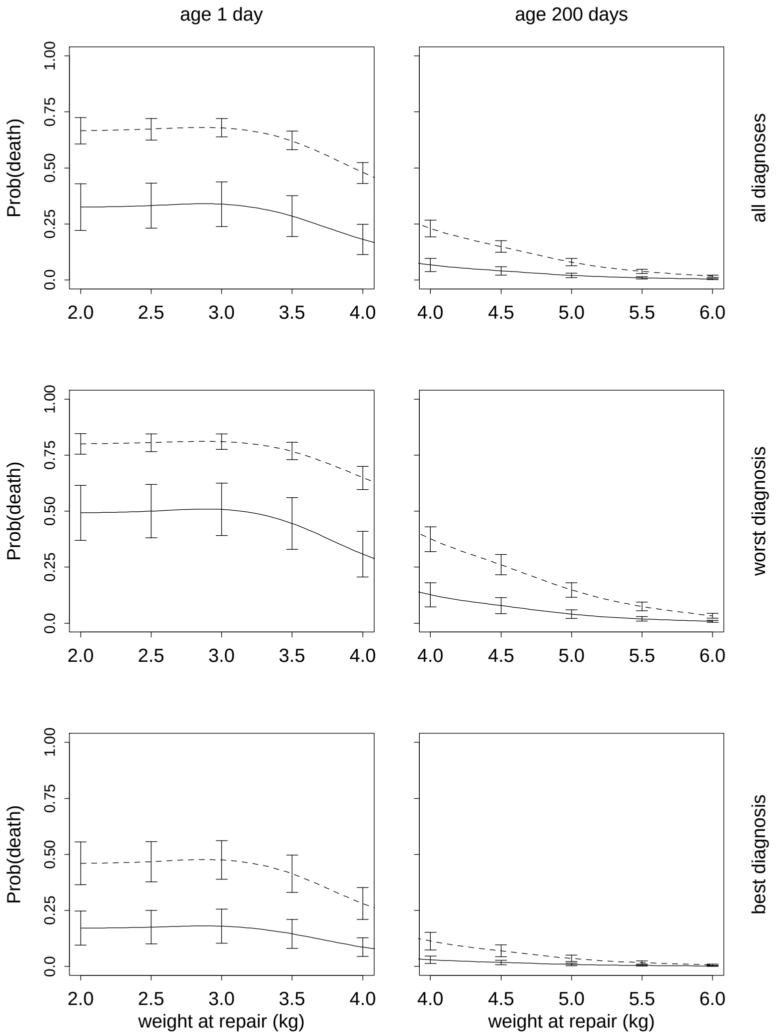

Figure 3: *Estimated probabilities for warm cardioplegia data, conditioned on two ages (columns), and three diagnostic classes (rows). Broken line is standard treatment, solid line is warm cardioplegia. Bars indicate* $\pm$ *one standard error.*

200 days, with higher risk for younger and older children. Note that the numerical algorithm is not able to achieve exactly 4 degrees of freedom for the age and weight terms, but 3.80 and 3.86 degrees of freedom respectively.

Table 1: *Results of a generalized model fit to warm cardioplegia data. The model was fit using the authors' software package GAIM, and the output is verbatim.*

```
Null deviance (-2 log likelihood ratio) = 590.97
Model deviance= 453.18

 variable    df     coef     std-err    p-value  nonlinear p-value
 -------    ----   ------    -----       ------    -----
  intcpt     1      2.43      0.893       2.72        --
  age       3.80  -.002      0.002     -.9856      0.005
  weight    3.86  -.9367     0.2031     -4.612     0.144
  diag1     1.00   1.37      0.481       2.85        --
  diag2     1.00  0.009      0.371       0.230       --
  diag3     1.00  -1.33      0.382      -3.47        --
  diag4     1.00  -1.51      0.402      -3.75        --
  diag5     1.00 -0.499      0.466      -1.07        --
  treatment 1.00  1.430      0.450       3.18        --
  operdate  1.00 -0.799E-04  0.188E-03  -0.425       --
            -----
            15.7
```

In the table each line gives the fit summary for the factor listed in the right column, *diag1* – *diag5* are the 5 indicator variables for the 6 diagnosis categories, and *df* is the degrees of freedom used for that variable. For ease of interpretation, the estimated curve for each variable is decomposed into a linear component and the remaining non-linear component (the linear component is essentially a weighted least squares fit of the fitted curve on the predictor, while the non-linear part is the residual). Other columns are *coef, std-err* and *p-value,* the estimated coefficient, standard error and normal score respectively for the linear component of the factor, while *nonlinear p-value* is the p-value for a test of nonlinearity of the effect. Note however that the effects of the other factors (e.g. *treatment)* are fully adjusted for the other factors, not just for their linear parts.

We see that warming procedure is strongly significant, with an estimated coefficient of 1.43 and a standard-error of 0.45, indicating a survival benefit. There are strong differences in the diagnosis categories, while the estimated effect of operation date is not large.

Since a logistic regression is additive on the logit scale but not on the probability scale, a plot of the fitted probabilities is often informative. Figure 3 shows the fitted probabilities broken down by age and diagnosis, and is a concise summary of the findings of this study. The beneficial effect of the treatment at the lower weights is evident. As with all nonrandomized studies, the results here should be interpreted with caution. In particular, one must ensure that the children were not chosen for the warming operation based on their prognosis. To investigate this, we perform a second analysis in which a dummy variable (say period), corresponding to before versus after February 1988, is inserted in place of the dummy variable for the warming operation. The purpose of this is to investigate whether the overall treatment strategy improved after February 1988. If this turns out not to be the case, it will imply that warming was used only for patients with a good prognosis, who would have survived anyway. A linear adjustment for operation date is included as before. The results are qualitatively very similar to the first analysis: age and weight are significant, with effects similar to those in Fig. 2; diagnosis is significant, while operation date (linear effect) is not. Period is highly significant, with a coefficient of -1.12 and a standard-error of 0.33. Hence there seems to be a significant overall improvement in survival after February 1988. For more details, see Williams et al. (1990).

3 Discussion

The nonlinear modeling procedures described here are useful for two reasons. First, they help to prevent model misspecification, which can lead to incorrect conclusions regarding treatment efficacy. Second, they provide information about the relationship between the predictors and the response that is not revealed by the use of standard modeling techniques. Linearity always remains a special case, and thus simple linear relationships can be easily confirmed with flexible modeling of predictor effects. Recently neural network models have become popular for flexible nonparametric regression modelling (Ripley 1994, for example). Although an interesting and general class of nonparametric models, they tend to be too heavy a hammer for many data analysis problems for several reasons:

- it is difficult to untangle the role of individual variables, while this goal is at the heart of additive models.
- neural networks tend to be most successful with very large data sets where many observations are available for fitting complex nonlinear interactions; additive models can get by with far fewer observations since they explicitly focus on lower order interactions.
- the fitting of neural networks models requires some experience, since multiple local minima are standard, and delicate regularization is required.

The most comprehensive source for generalized additive models is the text of that name by Hastie and Tibshirani (1990), from which this example was taken. Different applications of this work in medical problems are discussed in Hastie, Botha & Schnitzler (1989) and Hastie & Herman (1990). Green & Silverman (1994) discuss penalization and spline models in a variety of settings. Wahba (1990) is a good source for the mathematical background of spline models. Efron & Tibshirani (1991) give an exposition of modern developments in statistics (including generalized additive models), for a non-mathematical audience.

There has been some recent related work in this area. Kooperberg, Stone & Truong (1993) describe a different method for flexible hazard modelling, as well as for other regression models, using fixed knot regression splines. Friedman (1991) proposed a generalization of additive modelling that finds interactions among prognostic factors. Of particular interest in the proportional hazards setting is the *varying coefficient* model of Hastie & Tibshirani (1995), in which the parameter effects can change with other factors such as time. The model has the form

$$h(t|x_{i1},\ldots,x_{ip}) = h_0(t)\exp\sum_{j=1}^{p}\beta_j(t)x_{ij} \tag{9}$$

The parameter functions $\beta_j(t)$ are estimated by scatterplot smoothers in a similar fashion to the methods described earlier. This gives a useful way of modelling departures from the proportional hazards assumption by estimating the way in which the parameters β_j change with time.

Software for fitting generalized additive models is available as part of the S/S-PLUS statistical language (Becker, Chambers & Wilks 1988, Chambers & Hastie 1991), in a Fortran program called `gamfit` available at statlib (in `general/gamfit` at the ftp site lib.stat.cmu.edu) and also in the GAIM package for MS-DOS computers (information available from the authors).

4 Acknowledgments

The second author was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Becker, R., Chambers, J. & Wilks, A. (1988), *The New S Language*, Wadsworth International Group.

Buja, A., Hastie, T. & Tibshirani, R. (1989), 'Linear smoothers and additive models (with discussion)', *Annals of Statistics* **17**, 453–555.

Chambers, J. & Hastie, T. (1991), *Statistical Models in S*, Wadsworth/Brooks Cole, Pacific Grove, California.

Efron, B. & Tibshirani, R. (1991), 'Statistical analsysis in the computer age', *Science*.

Friedman, J. (1991), 'Multivariate adaptive regression splines (with discussion)', *Annals of Statistics* **19**(1), 1–141.

Green, P. & Silverman, B. (1994), *Nonparametric regression and generalized linear models: a roughness peanlty approach*, Chapman and Hall.

Hastie, T. & Herman, A. (1990), 'An analysis of gestational age, neonatal size and neonatal death using nonparametric logistic regression', *Journal of Clinical Epidemiology* **43**, 1179–90.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall.

Hastie, T. & Tibshirani, R. (1995), 'Discriminant analysis by mixture estimation', *J. Royal Statist. Soc. (Series B).* to appear.

Hastie, T., Botha, J. & Schnitzler, C. (1989), 'Regression with an ordered categorical response', *Statistics in Medicine* **43**, 884–889.

Kooperberg, C., Stone, C. & Truong, Y. (1993), Hazard regression, Technical report, Dept of statistics, Univ. of Cal. Berkeley.

McCullagh, P. & Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall.

Ripley, B. (1994), 'Neural networks and related methods for classification', *J. Royal Statist. Soc. (Series B)* pp. 409–456. (with discussion).

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia.

Williams, W., Rebeyka, I., Tibshirani, R., Coles, J., Lightfoot, N., Freedom, R. & Trusler, G. (1990), 'Warm induction cardioplegia in the infant: a technique to avoid rapid cooling myocardial contracture', *J. Thorac. and Cardio. Surg* **100**, 896–901.