Generalized additive models for medical research

Trevor Hastie and Robert Tibshirani

Stat Methods Med Res 1995 4: 187
DOI: 10.1177/096228029500400302

Statistical Methods in Medical Research 1995; **4**: 187–196

Generalized additive models for medical research

Trevor Hastie and **Robert Tibshirani*** Department of Statistics and Division of Biostatistics, Stanford University, Stanford, California, USA

This article reviews flexible statistical methods that are useful for characterizing the effect of potential prognostic factors on disease endpoints. Applications to survival models and binary outcome models are illustrated.

1 Introduction

In the statistical analysis of clinical trials and observational studies, the identification and adjustment for prognostic factors is an important component. Valid comparisons of different treatments requires the appropriate adjustment for relevant prognostic factors. The failure to consider important prognostic variables, particularly in observational studies, can lead to errors in estimating treatment differences. In addition, incorrect modelling of prognostic factors can result in the failure to identify nonlinear trends or threshold effects on survival. This article describes flexible statistical methods that may be used to identify and characterize the effect of potential prognostic factors on disease endpoints. These methods are called 'generalized additive models'.

Two of the most commonly used statistical models in medical research are the proportional hazards regression model for survival data and the logistic regression model for binary data. Both of these techniques (and many others) model the effects of prognostic factors x_j in terms of a linear predictor of the form $\sum x_j\beta_j$, where the β_j are parameters. The generalized additive model replaces $\sum x_j\beta_j$ with $\sum f_j(x_j)$ where f_j is an unspecified ('nonparametric') function. This function is estimated in a flexible manner using a scatterplot smoother. The estimated function $\hat{f}_j(x_j)$ can reveal possible nonlinearities in the effect of the x_j.

We illustrate how this approach is used to extend the proportional hazards model in Section 2. Section 3 gives some background on the methodology, while Section 4 illustrates the logistic regression model and its generalization. Some related developments are discussed in Section 6.

2 Example: the proportional hazards model

The proportional hazards model of Cox[1] is a popular tool for analysing censored failure time data. It is semiparametric; that is, the model does not make any distributional assumptions about the failure times, but does specify the form in which covariates (prognostic factors) affect the hazard rate of failure. It is most commonly used for right censored data but can also be adapted to left censored and truncated data. The model is expressed in terms of the hazard rate $h(t\,|\,x_{i1},\ldots x_{ip})$, defined as the

*On sabbatical leave from Department of Preventive Medicine and Biostatistics, and Department of Statistics, University of Toronto.

Address for correspondence: Robert Tibshirani, Department of Statistics, Sequoia Hall, Stanford University, Stanford, California 94305, USA.

probability that an individual with covariates $x_{i1}, \ldots x_{ip}$ fails in a short time interval after time t, given survival to time t. The (linear) proportional hazards model expresses the hazard rate as

$$h(t \mid x_{i1}, \ldots, x_{ip}) = h_0(t) \exp \sum_{j=1}^{p} \beta_j x_{ij} \tag{2.1}$$

where $h_0(t)$ is a baseline hazard function of arbitrary form. The quantities $x_{i1}, \ldots x_{ip}$ are fixed covariates, and $\beta_1, \ldots, \beta_p$ are unknown parameters. A nonmathematical description of this model is given in Tibshirani.[2]

While some prognostic factors (such as disease stage) may be linearly related to survival, the influence of other factors, such as patient characteristics and clinical laboratory values, may be more accurately described by a nonlinear relationship. The methods described here relax the linearity assumptions and allow smooth nonlinear functions of the covariates to be included into the log hazard ratio. A linear relationship remains a special case.

The *generalized additive proportional hazards model* has the form

$$h(t \mid x_{i1}, \ldots, x_{ip}) = h_0(t) \exp \sum_{j=1}^{p} f_j(x_{ij}) \tag{2.2}$$

Here the f_js are smooth functions. Rather than specify a specific form for each f_j, like linear or logarithmic, we estimate them in a flexible manner using *scatterplot smoothers*.

Details of this procedure are given later; first we illustrate the model on data from a clinical trial for the treatment of node-positive breast cancer. A more detailed analysis of this example is given in Hastie, Sleeper and Tibshirani.[3] This clinical trial is based on 260 postmenopausal women and is described in detail in Taylor *et al.*[4]

The median follow-up time as of November 1991 was 6.86 years. By that date, 143 (54%) deaths and 176 (67%) disease recurrences have been observed. All cases were 69 years or younger (mean ±1 SD, 57.4 ± 4.7 years) and underwent mastectomy and axillary node dissection within eight weeks of randomization; no postoperative radiation therapy was allowed. Patients were randomized to one of three treatment arms: (1) observation only; (2) CMFP (a drug combination of cyclophosphamide, methotrexate, fluorouracil, and prednisone); or (3) CMFPT (CMFP plus tamoxifen).

The endpoint is disease-free survival, and there are 260 cases. The following variables were considered as potential prognostic factors for disease-free survival:

1) the presence or absence of tumour necrosis;
2) tumour size;
3) number of nodes examined;
4) age of the patient;
5) body mass index (kg/m^2);
6) number of days from surgery to randomization.

With the exception of tumour necrosis, all of these variables may be modelled using the smoothing methods described above. Although the number of nodes examined is not a continuous variable, it is ordinal and ranges from 1 to 48.

Patients were randomized within four strata, defined by ER (oestrogen receptor) status positive or negative, and number of positive nodes ≤ 3 or >3. We used the appropriate partial likelihood for stratified data in our analysis.[5]

We fit a generalized additive model of the form (2.2), which included all six potential prognostic factors as well as treatment effects. A flexible function fit was used for each of the five continuous covariates, and indicator variables were used for the tumour necrosis and the treatment groups. The three arms of treatment are represented in the model in order to determine their relative impact on disease recurrence while the effect of prognostic factors is modelled simultaneously. A multivariate model will help to correct imbalances among the treatment groups with respect to important factors. For example, in this trial, women receiving CMFPT were somewhat more likely to have smaller tumours and evidence of tumour necrosis, and had fewer nodes examined than women on the Observation-only and CMFP arms.

Two of the quantitative covariates were positively associated with the risk of disease recurrence: tumour size ($p < 0.01$) and the number of days between surgery and randomization $p < 0.05$). Furthermore, the graphs of the estimated risk function ($\hat{f}$) for each of these factors was quite linear. The presence of tumour necrosis was also identified as an important risk factor ($p < 0.05$).

The fitted curves for number of nodes examined, age and body mass index are shown in Figure 1. A threshold effect is suggested with respect to the number of nodes examined: risk decreases from one to 16 nodes, and then remains fairly constant. The log hazard ratio as a function of age is roughly an inverted U-shaped function: risk is similar for women aged 50 to 60 years (with a slight dip around 56 years); women below 50 and above 60 years are at decreased risk. The body mass index fit reveals no change in risk until approximately 32 kg/m^2; thereafter, risk decreases. However, the standard error curves around the decreasing portion of the curve are very wide, reflecting the very small number of women with high BMI. The decreasing risk in this region may be the result of this small group of patients having unusually long disease-free survival times.

3 Smoothing methods and generalized additive models

In this section we give some background on the methodology that was used in the previous example, and indicate the way in which it is applied to other models.

The building block of the generalized additive model algorithm is the scatterplot smoother. We will first describe scatterplot smoothing in a simple setting, and then indicate how it is used in models like the proportional hazards model

Suppose that we have a scatterplot of points (x_i, y_i) like that shown in Figure 2. Here y is a response or outcome variable, and x is a prognostic factor. We wish to fit a smooth curve $f(x)$ that summarizes the dependence of y on x. 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 spline smoother imposes smoothness on $f(x)$. We seek the function $f(x)$ that minimizes

$$\sum(y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx \tag{3.1}$$

Notice that $\int f''(x)^2$ measures the 'wiggliness' of the function f: linear fs have $\int f''(x)^2 = 0$, while nonlinear fs produce values bigger than zero. λ is a nonnegative smoothing parameter that must be chosen by the data analyst. It governs the tradeoff between the goodness of fit to the data and (as measured by $\sum(y_i - f(x_i))^2$) and wiggliness of the function. Larger values of λ force f to be smoother.

For any value of λ, the solution to (3.1) 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 2 shows a cubic spline fit to the data.

What value of λ did we use in Figure 2? 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 2 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

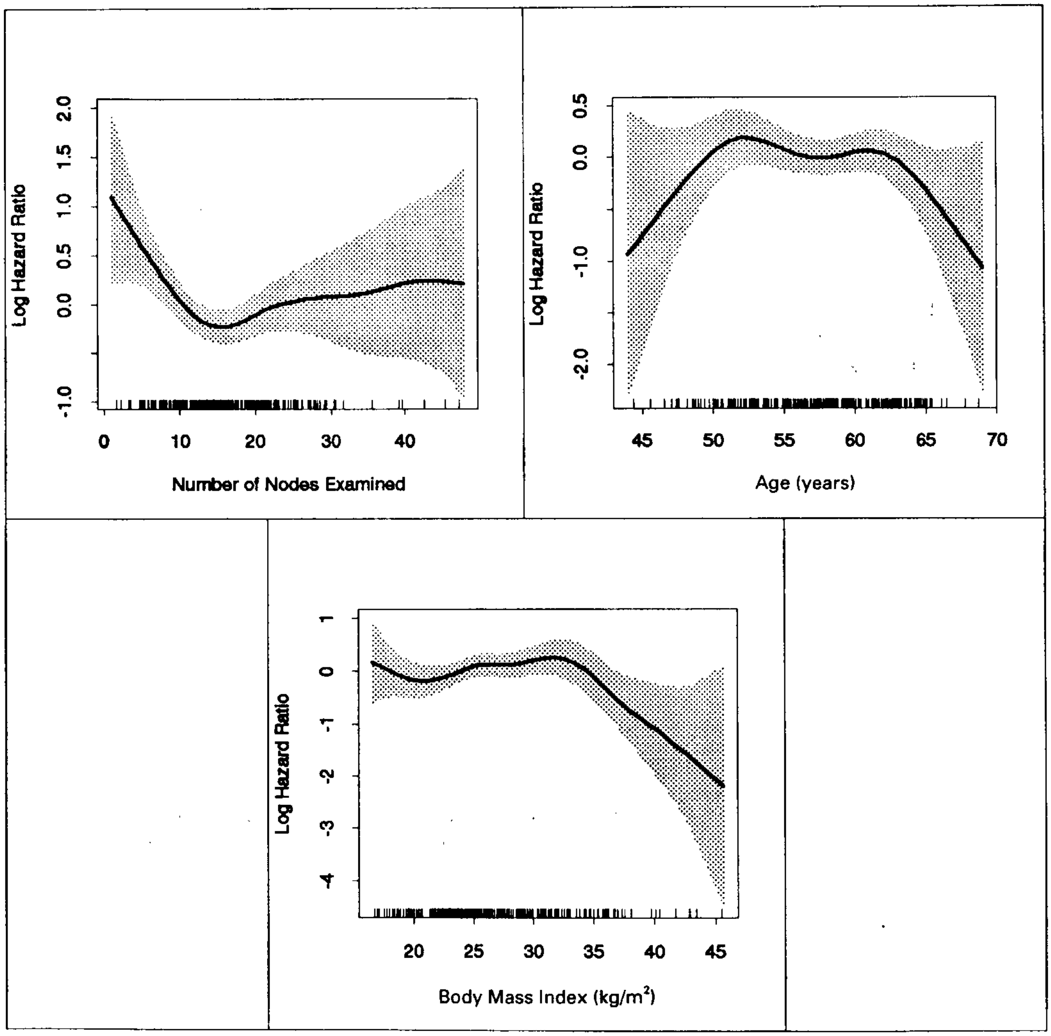

Figure 1 Function plots for three of the terms in the additive proportional hazards model. Each curve is centred to have average zero over the range of the data. The broken curves indicate approximate pointwise 95% confidence intervals, and the vertical bars at the base of the plots represent a frequency plot of the predictor variable.

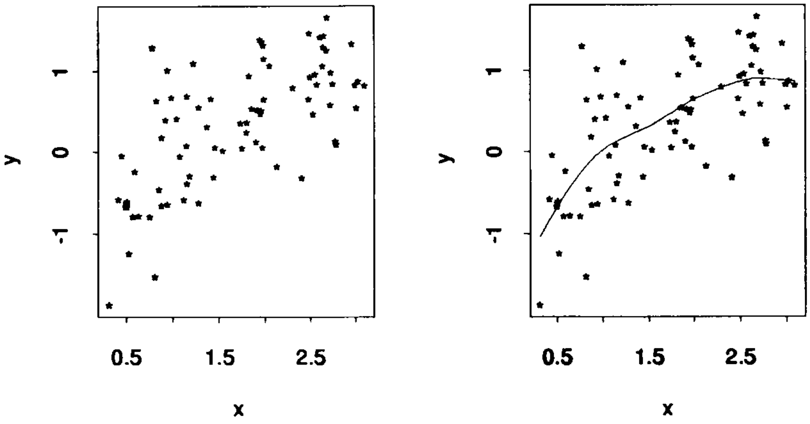

Figure 2 Left panel shows a fictitious scatterplot of an outcome measure y plotted against a prognostic factor x. In the right panel, a scatterplot smooth has been added to describe the trend of y on x.

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 prognostic factor. With multiple prognostic factors, if x_{ij} denotes the value of the jth prognostic factor the ith observation, we fit the additive model

$$\hat{y}_i \approx \sum_j f_j(x_{ij}) \tag{3.2}$$

A criterion like (3.1) can be specified for this problem, and a simple iterative procedure exists for 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 prognostic factor in turn. The process continues 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.

For models such as the proportional hazards 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 an 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. Details can be found in Chapter 6 of Hastie and Tibshirani.[6]

4 The generalized additive logistic model

Generalized additive models can be used in virtually any setting where linear models are used. The basic idea is to replace $\sum x_{ij}\beta_j$, the linear component of the model with an additive component $\sum f_j(x_{ij})$. Along with the proportional hazards model described earlier, 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:

$$\log \frac{p(y_i | x_{i1}, \ldots x_{ip})}{1 - p(y_i | x_{i1}, \ldots x_{ip})} = \beta_0 + x_{i1}\beta + \ldots x_{ip}\beta_p \tag{4.1}$$

The generalized additive logistic model assumes instead that

$$\log \frac{p(y_i | x_{i1}, \ldots x_{ip})}{1 - p(y_i | x_{i1}, \ldots x_{ip})} = \beta_0 + f_1(x_{i1}) + \ldots f_p(x_{ip}) \tag{4.2}$$

The functions $f_1, f_2, \ldots f_p$ are estimated by an algorithm like the one described earlier.

To illustrate this, we describe a second example on the survival of children after cardiac surgery for heart defects, taken from Williams *et al.*[7] The data was collected during the period 1983–1988. A preoperation 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.

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 with 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 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. 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.

The resulting curves for age and weight are shown in Figure 3. 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 200 days, with higher risk for younger and older children.

In the table each line gives the fit summary for the factor listed in the right column. `diag1` – `diag5` are the five indicator variables for the six diagnosis categories. `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 nonlinear component (the linear component is essentially a weighted least squares fit

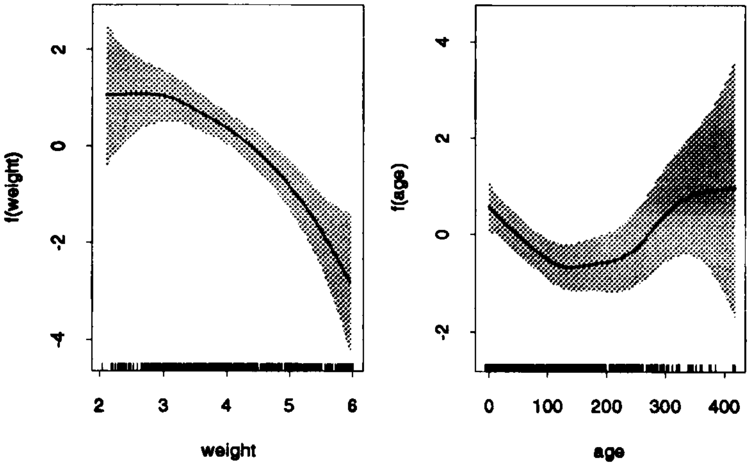

Figure 3 Estimated functions for weight and age for warm cardioplegia data. The shaded region represents twice the pointwise asymptotic standard errors of the estimated curve.

of the fitted curve on the predictor, while the nonlinear part is the residual). `coef`, `std-err` and `p-value` are the estimated coefficient, standard error and normal score for the linear component of the factor. `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 4 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

Table 1 Results of generalized model to fit warm cardioplegia data

Variable	df	coef	std-err	*p* value	Nonlinear *p* value
intcpt	1.00	2.43	0.893	2.72	–
age	3.80	−0.002	0.002	−0.9856	0.005
weight	3.86	−0.9367	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				

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

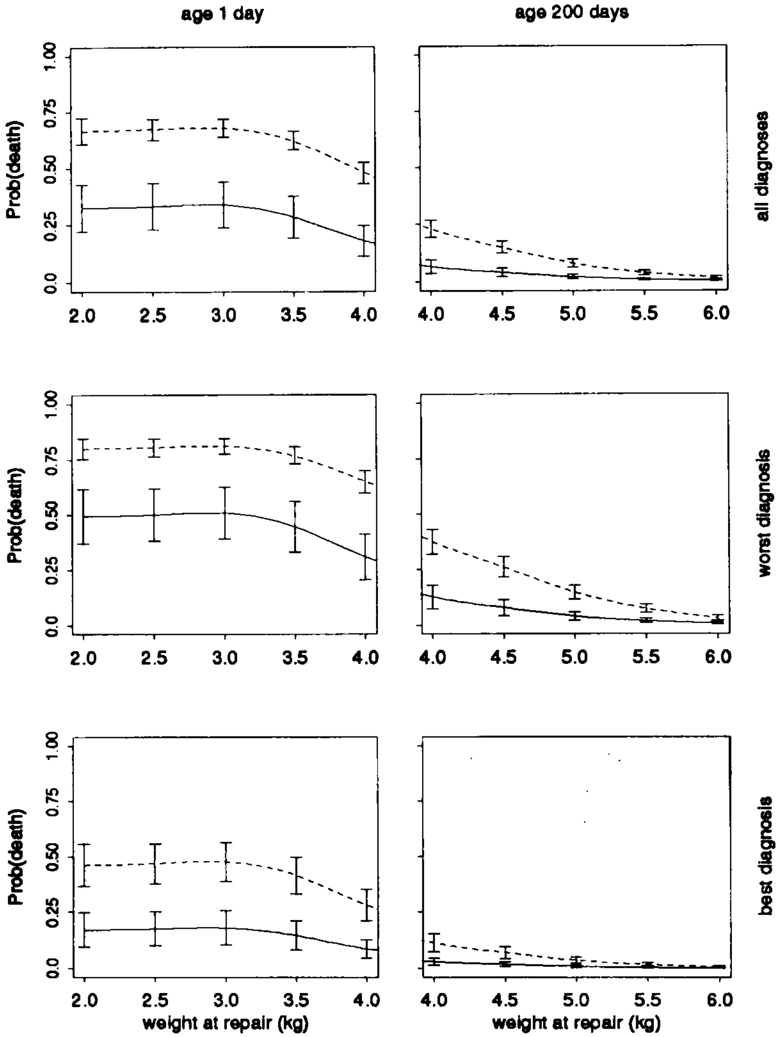

Figure 4 Estimated probabilities for warm cardioplegia data, conditioned on two ages (columns), and three diagnostic classes (rows). Broken line is standard treatment, solid line is warm cardioplegia. Bars indicate ± one standard error.

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 Figure 3; diagnosis is significant, while operation date (linear effect) is not. Period is highly 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.*[7]

5 Discussion

The nonlinear modelling procedures described here are useful for two reasons. First, they help to prevent model misspecification, which can lead to incorrect conclusions regarding treatment efficacy. Secondly, they provide information about the relationship between prognostic factors and disease risk that is not revealed by the use of standard modelling techniques. Linearity always remains a special case, and thus simple linear relationships can be easily confirmed with flexible modelling of covariate effects.

The most comprehensive source for generalized additive models is the text of that name by Hastie and Tibshirani,[6] from which the second example was taken. The proportional hazards example was taken from the more comprehensive analysis in Hastie *et al.*,[3] where the regression spline approach to additive modelling is also described. Different applications of this work in medical problems are discussed in Hastie, Botha and Schnitzler[8] and Hastie and Herman.[9] Green and Silverman[10] discuss penalization and spline models in a variety of settings. Wahba[11] is a good source for the mathematical background of spline models.

Efron and Tibshirani[12] give an exposition of modern developments in statistics (including generalized additive models), for a nonmathematical audience.

There have been some recent related work in this area. Kooperberg, Stone and Truong[13] describe a different method for flexible hazard modelling. Friedman[14] 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 and Tibshirani,[15] in which the parameter effects can change with other factors such as time. The model has the form.

$$h(t \mid x_{i1}, \ldots, x_{ip}) = h_0(t)\exp\sum_{j=1}^{p} \beta_j x_{ij} \tag{5.1}$$

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 and Wilks,[16] Chambers and Hastie,[17] 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).

Acknowledgements

We thank Lynn Sleeper for her work in preparing the manuscript Hastie *et al.*[3] from which the first example was taken, and a referee for helpful comments. The second author was supported by the Natural Sciences and Engineering Research Council of Canada.

References

1 Cox D. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B Series* 1972; **74**: 187–220.

2 Tibshirani R. A plain man's guide to the proportional hazards model. *Clinical and Investigative Medicine* 1982; **5**(1): 63–68.

3 Hastie T, Sleeper L, Tibshirani R. Flexible covariate effects in the cox model. *Breast Cancer Research and Treatment* – special issue, 1992.

4 Taylor SG, Knuiman MW, Sleeper LA *et al.* Six-year results of the eastern cooperative oncology group trial of observation versus cmfp versus cmfpt in postmenopausal patients with node-positive breast cancer. *Journal of Clinical Oncology* 1989; **7**: 879–89.

5 Kalbfleisch J, Prentice R. *The statistical analysis of failure time data.* New York: Wiley, 1980.

6 Hastie T, Tibshirani R. *Generalized additive models.* London and New York: Chapman and Hall, 1990.

7 Williams W, Rebeyka I, Tibshirani R *et al.* Warm induction cardioplegia in the infant: a technique to avoid rapid cooling myocardial contracture. *Journal of Thoracic and Cardiovascular Surgery* 1990; **100**: 896–90.

8 Hastie T, Botha J, Schnitzler C. Regression with an ordered categorical response. *Statistics in Medicine* 1989; **43**: 884–89.

9 Hastie T, Herman A. An analysis of gestational age, neonatal size and neonatal death using nonparametric logistic regression. *Journal of Clinical Epidemiology* 1990; **43**: 1179–90.

10 Green P, Silverman B. *Nonparametric regression and generalized linear models: a roughness penalty approach.* London and New York: Chapman and Hall, 1994.

11 Wahba G. *Spline models for observational data.* Philadelphia: DIAM, 1990.

12 Efron B, Tibshirani R. Statistical analysis in the computer age. *Science* 1991;

13 Kooperberg C, Stone C, Truong Y. Hazard regression. Technical report. Berkeley: Dept of Statistics, University of California, 1993.

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

15 Hastie T, Tibshirani R. Discriminant analysis by mixture estimation. *Journal of the Royal Statistical Society B Series* 1995 (in press).

16 Becker R, Chambers J, Wilks A. *The new S language.* Pacific Grove, CA: Wadsworth International Group, 1988.

17 Chambers J, Hastie T. *Statistical models in S.* Pacific Grove: Wadsworth/Brooks Cole, 1991.