Generalized additive models

Trevor Hastie [*]
and
Robert Tibshirani [†]

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 modeling 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 an outcome variable. These methods are called "generalized additive models", and extend the traditional linear statistical model. They can be applied in any setting where a linear or generalized linear model is typically used. These settings include standard continuous response regression, categorical or ordered categorical response data, count data, survival data and time series.

One of the most commonly used statistical models in medical research is the logistic regression model for binary data. We use it here as a specific illustration of a generalized additive mode. Logistic regression (and many

[*] Department of Statistics and Division of Biostatistics, Stanford University, Stanford California 94305; trevor@stat.stanford.edu

[†] Department of Preventive Medicine and Biostatistics, and Department of Statistics, University of Toronto; tibs@playfair.stanford.edu; tibs@utstat.toronto.edu

other techniques) 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 a unspecified ("non-parametric") 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 first give some background on the methodology, and then discuss the details of the logistic regression model and its generalization. Some related developments are discussed in the last section.

2 Smoothing methods and generalized additive 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 generalized additive modeling.

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 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

$$\sum(y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx \tag{1}$$

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.

For any value of λ, the solution to (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

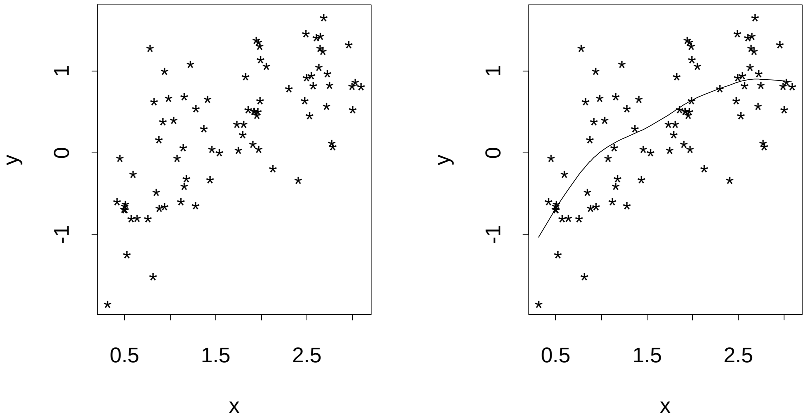

Figure 1: Left panel shows a fictitious scatterplot of an outcome measure y plotted against a prognostic factor x. In the right panel, a scatterplot smooth has been added to describe the trend of y on x.

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

$$\hat{y}_i \approx \sum_j f_j(x_{ij}) \tag{2}$$

A criterion like (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 is continues until the estimates $\hat{f}_j$ stabilize. These procedure is known as "backfitting" and the resulting fit is analogous to a multiple regression for linear models.

When generalized additive models are fit to binary response data (and in many other settings), the appropriate error 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. Details can be found in chapter 6 of Hastie & Tibshirani (1990).

3 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 $\sum f_j(x_{ij})$.

In the logistic regression 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:

$$\log \frac{p(y_i|x_{i1}, \ldots x_{ip})}{1 - p(y_i|x_{i1}, \ldots x_{ip})} = \beta_0 + x_{i1}\beta_1 + \ldots x_{ip}\beta_p \tag{3}$$

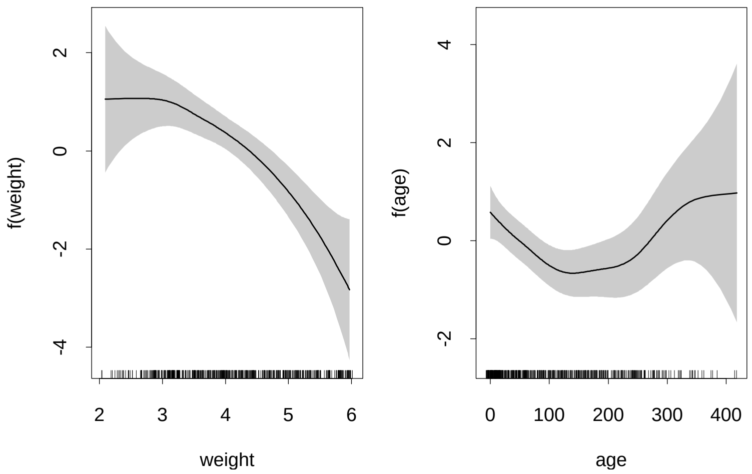

Figure 2: Estimated functions for `weight` and `age` for warm cardioplegia data. The shaded region represents twice the pointwise asymptotic standard errors of the estimated curve.

The generalized additive logistic model assumes instead that

$$\log \frac{p(y_i|x_{i1}, \ldots x_{ip})}{1 - p(y_i|x_{i1}, \ldots x_{ip})} = \beta_0 + f_1(x_{i1}) + \ldots f_p(x_{ip}) \tag{4}$$

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 study on the survival of children after cardiac surgery for heart defects, taken from Williams, Rebeyka, Tibshirani, Coles, Lightfoot, Freedom & Trusler (1990). The data were 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.

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 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.

An analysis of deviance can be carried out for inference from a generalized additive model, analogous to that done for generalized linear models. The only new twist is estimation of the degrees of freedom or effective number of parameters of the fitted model, which was discussed in the previous section. This analysis shows that the warming procedure is strongly beneficial to survival. 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

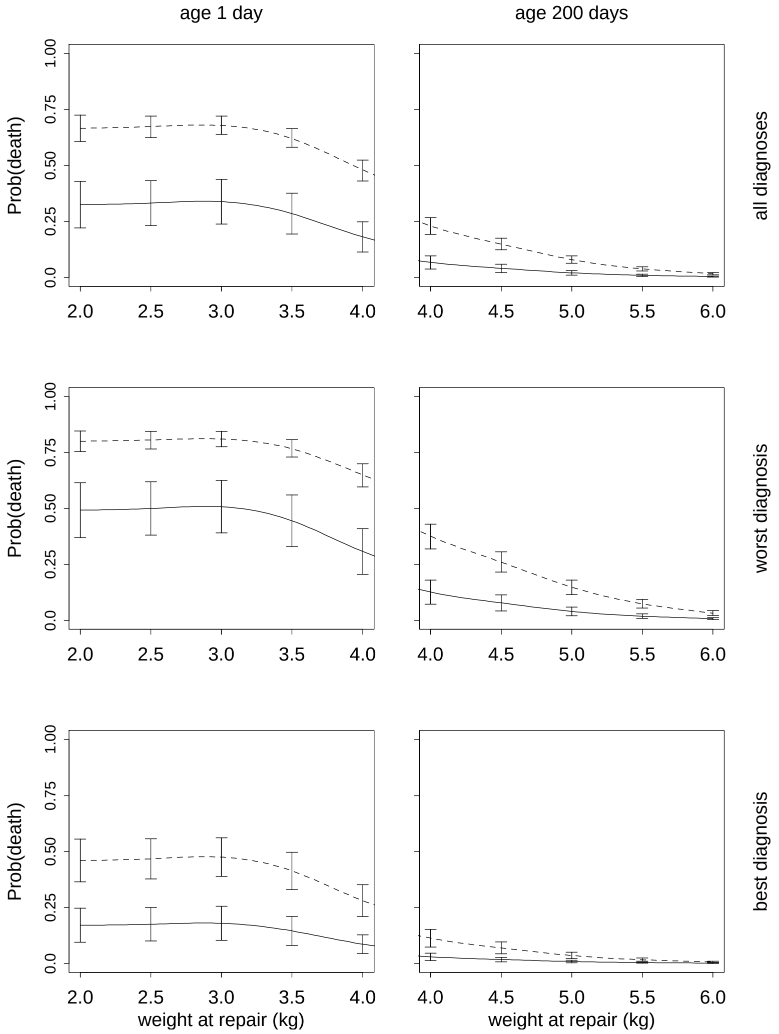

Figure 3: Estimated probabilities for warm cardioplegia data, conditioned on two ages (columns), and three diagnostic classes (rows). Broken line is standard treatment, solid line is warm cardioplegia. Bars indicate ± one standard error.

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. Hence there seems to be a significant overall improvement in survival after February 1988. For more details, see Williams et al. (1990).

4 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 prognostic factors and disease risk 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 covariate effects.

The most comprehensive source for generalized additive models is the text of that name by Hastie and Tibshirani (1990), from which the example was taken. Hastie, Sleeper & Tibshirani (1992) provide a detailed example of the use of generalized additive models in the proportional hazards setting. 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 nonmathematical audience.

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

$$h(t|x_{i1}, \dots, x_{ip}) = h_0(t) \exp \sum_{j=1}^{p} \beta_j(t) x_{ij} \tag{5}$$

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 modeling 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 in the S/Splus statistical environment (?, Chambers & Hastie 1991, `gam()`), 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, available from the authors.

5 Acknowledgments

Trevor Hastie was partially supported by grant DMS-9504495 from the National Science Foundation, and grant ROI-CA-72028-01 from the National Institutes of Health. Robert Tibshirani was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

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

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

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. (1993), Discriminant analysis by gaussian mixtures, In preparation.

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

Hastie, T., Sleeper, L. & Tibshirani, R. (1992), 'Flexible covariate effects in the cox model', *Breast Cancer Research and Treatment — special issue.*

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

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.