BIOMETRICS 56, 139–144
March 2000

Negative Binomial Additive Models

Sally W. Thurston,[1,*] M. P. Wand,[1] and John K. Wiencke[2]

[1]Department of Biostatistics, Harvard School of Public Health,
655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
[2]Laboratory for Molecular Epidemiology, Department of Epidemiology and Biostatistics,
School of Medicine, University of California San Francisco, San Francisco, California 94143 U.S.A.
*email: sthursto@hsph.harvard.edu

SUMMARY. The generalized additive model is extended to handle negative binomial responses. The extension is complicated by the fact that the negative binomial distribution has two parameters and is not in the exponential family. The methodology is applied to data involving DNA adduct counts and smoking variables among ex-smokers with lung cancer. A more detailed investigation is made of the parametric relationship between the number of adducts and years since quitting while retaining a smooth relationship between adducts and the other covariates.

KEY WORDS: Backfitting algorithm; DNA adducts; Generalized additive models; Local polynomial regression; Semiparametric models; Smoothing.

1. Introduction

Cigarette smoke contains a complex mixture of carcinogens, some of which are capable of binding to DNA after metabolic activation within the cell. Once bound to DNA, these carcinogen-DNA complexes, called adducts, are thought to elicit mutational changes. Several recent studies have found a relationship between adduct levels and cancer (Schut and Shiverick, 1992; Wogan, 1992; Tang et al., 1995). It is thought that the mechanism by which cigarette smoking causes cancer is at least in part due to adduct formation in the p53 gene (Denissenko et al., 1996, 1997; Kure et al., 1996).

Since high adduct levels are associated with cancer, they could potentially be used as a good marker for cancer risk. Knowing who is at high risk could lead to early clinical diagnosis of cancer and thus to better survival rates. Unfortunately, measuring adduct levels is very expensive. The focus of this paper is to model the number of adducts from smoking variables, which can be measured cheaply and easily.

Considerable overdispersion has been found in adduct models. In a previous study using this dataset (Wiencke et al., 1999), a negative binomial model was used to estimate the number of adducts as a linear function of smoking variables. However, the relationship between the number of adducts and some of the covariates is nonlinear. Generalized additive modeling (GAM) (Hastie and Tibshirani, 1990) is a flexible and powerful regression technique since it allows for the effect of each covariate to be modeled as an unspecified smooth function rather than as a rigid parametric function. Presently, GAM models only allow for standard exponential family likelihoods, which is a limitation when the data are overdispersed.

Many methods for dealing with overdispersion within a generalized linear models (GLM) (McCullagh and Nelder, 1989) context have been used. For count data that might be modeled by a Poisson distribution in the absence of overdispersion, quasilikelihood (Wedderburn, 1974; McCullagh and Nelder, 1989, Section 9.3) extends the GLM framework by allowing the variance to be a constant times the mean. Alternatively, the mean of the Poisson distribution can be modeled to have a gamma distribution, which means the marginal distribution of the number of adducts is negative binomial and the variance increases with the mean (Lawless, 1987). Other methods include extended quasi-likelihood methods (Nelder and Pregibon, 1987) and double exponential families (Efron, 1986).

Methods for handling overdispersion within a GAM context have not been well developed. A quasi-likelihood approach based on introducing a multiplicative overdispersion parameter could be used (e.g., Fan, Heckman, and Wand, 1995), but this may have limitations in handling overdispersion in practice. Yee and Wild (1996) propose a general class of multiparameter generalized additive models based on smoothing splines but provide little detail on handling overdispersed count data using, e.g., the negative binomial distribution.

In this paper, we extend the GAM framework to allow for the mean number of adducts, given the covariates, to have a negative binomial distribution. This has a number of advantages over a quasi-likelihood approach in this application. While quasilikelihood is efficient if overdispersion is modest (Cox, 1983), in this case, the variance is more than 100 times larger than the mean, suggesting that the quasi-likelihood approach would not be optimal. Furthermore, a plot of the absolute value of the Pearson residuals versus the fitted values in a GAM model with Poisson likelihood suggests that the variance increases with the mean (Figure 1), which is not changed in quasilikelihood but is a natural part of a negative

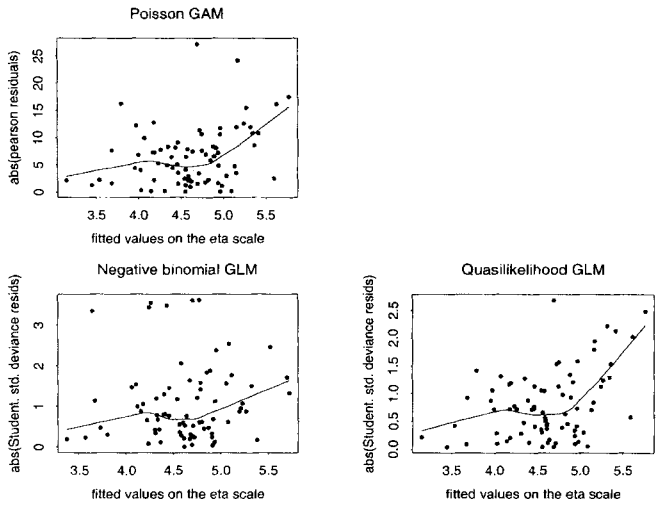

Figure 1. Top. Absolute value of Pearson residuals versus fitted values on the eta scale, with lowess smooth in a Poisson generalized additive model of number of lung adducts as a function of four smoking variables. **Bottom.** Absolute value of the Studentized standardized deviance residuals versus fitted values on the eta scale with lowess smooth, for the negative binomial model (bottom left) and for the quasilikelihood model (overdispersed Poisson: bottom right). The latter models are fully parametric, with a broken stick relationship (break at 9 years) for years since quitting, a quadratic relationships for years of smoking, and linear relationships for cigarettes per day and age of smoking initiation.

binomial model. Another method for examining the nature of overdispersion is given in Lambert and Roeder (1995).

In Section 2, we describe the nature of the data and the models we consider. Section 3 describes the algorithms for fitting these models. In Section 4, we briefly describe the method of weighted polynomial smoothing. We give details of implementation in Section 5. In Section 6, we discuss semiparametric, or partial linear, models and conclude with analysis of the adduct data in Section 7.

2. Data and Models

We intend to model the number of DNA adducts as a negative binomial variable so will start by describing this distribution. The data consist of the number of polyaromatic hydrocarbon adducts and smoking information (age of smoking initiation, years of smoking, cigarettes smoked per day, and years since quitting) for a set of ex-smoking lung cancer patients.

2.1 *Negative Binomial Likelihood*

Let $Y_1, \ldots, Y_n$ be a set of count data. The negative binomial model for these data is defined by

$$P(Y_i = y_i \mid \mu_i, k) = \binom{y_i + k - 1}{y_i} \left(\frac{\mu_i}{k + \mu_i}\right)^{y_i} \left(\frac{k}{k + \mu_i}\right)^{k},$$

where $\mu_i = \mathrm{E}(Y_i)$ and k is a shape parameter. Under this model, the variance of Y_i is $\mu_i + \mu_i^2/k$. Note that, for large k, the model approaches the Poisson model.

Writing this in exponential family format, the log likelihood is $\ell(\mu_i, k \mid Y_i) = Y_i \ln(\mu_i/\mu_i + k) - k \ln\{\mu_i/(\mu_i + k)\} + \ln \Gamma(Y_i + k) - \ln \Gamma(k) - \ln \Gamma(Y_i + 1) + k \ln k$, from which it can be seen that the canonical link is $\eta_i = \ln\{\mu_i/(\mu_i + k)\}$.

If k were known, this would be an exponential family. For a given k, the log likelihood for $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ is $\ell(\mu; k) = \Sigma_{i=1}^n Y_i \ln\{\mu_i/(\mu_i + k)\} - \Sigma_{i=1}^n k \ln(1 + \mu_i/k) + c(Y, k)$, where $c(Y, k)$ is a function of the Y_i's and k.

For a given μ, the log likelihood for k is

$$\ell(k; \mu_i) = n\{k \ln k - \ln \Gamma(k)\} + \sum_{i=1}^{n} \{\ln \Gamma(Y_i + k) - (Y_i + k) \ln(k + \mu_i)\} + d(Y, \mu) \quad (1)$$

for some function $d(Y, \mu)$. In the situation where the second parameter is replaced by an estimate, these are called the profile likelihoods of μ and k, respectively.

2.2 *Incorporating Covariates*

Let $X_1, \ldots, X_n$ be a set of d-dimensional covariates. We use the notation $\mu(X_i)$ to indicate that the mean of Y_i may depend on covariates $X_{i1}, \ldots, X_{id}$. Allowing the mean response to depend on covariates X, in a linear model, we could model $\mu(X_i) = \alpha + \Sigma_{j=1}^d \beta_j X_{ij}$.

This model can be extended to a generalized linear model by allowing the link function η, which is a function of the mean μ_i, to be linear in the covariates, i.e., $\eta_i = g\{\mu(X_i)\} = \alpha + \Sigma_{j=1}^d \beta_j X_{ij}$.

The generalized additive models framework extends this by allowing nonlinearity in the relationship between η and the covariates, i.e., $\eta_i = g\{\mu(X_i)\} = \alpha + \Sigma_{j=1}^d f_j(X_{ij})$, where each f_j is a smooth function.

To fit the negative binomial model for Y given covariates X, several link functions η are possible. The canonical link has the disadvantage that η must be negative. This leads to problems when using iterative methods to fit a generalized additive model. To overcome this problem, we use the log link. A comparison of the Poisson model, the negative binomial model with the canonical link, and the negative binomial with the log link is given in Table 1.

We can write the model as $\mathrm{E}(Y_i \mid X_i) = \mu(X_i) = \exp\{\alpha + \Sigma_{j=1}^d f_j(X_{ij})\}$ and fit it iteratively, as described in more detail in subsequent sections. Here we briefly describe the idea. k can be estimated by maximizing the log likelihood for k given in (1), using the current estimates of μ. For any $\hat{k}$, fitting the model involves Fisher's scoring, which comes from a Taylor expansion, $Z_i = g(Y_i \mid X_i) \approx g(\mu_i) + (Y_i - \mu_i) g'(\mu_i)$, where $g'(\mu_i) = (\partial \eta_i / \partial \mu_i)$ and $\mathrm{var}(Z_i \mid X_i) = \mathrm{var}(Y_i \mid X_i) g'(\mu_i)^2 = V_i (\partial \eta_i / \partial \mu_i)^2$.

In the GLM framework, in which $\eta = g(\mu) = \alpha + \Sigma_{j=1}^d \beta_j X_j$, β can be estimated by regressing Z on X with weights $W = 1/\mathrm{var}(Z)$. We would then use $\hat{\beta}$ to update our estimate of μ and iterate between estimating β and μ (McCullagh and Nelder, 1989, p. 40). In the GAM framework, instead of regressing Z on X with weights W, the residuals, Z, are iteratively smoothed on each X_j with weights W and are centered in a process called backfitting.

3. Algorithms for Model Fitting

Hastie and Tibshirani (1990) explain how an exponential family GAM can be fit via a combination of two algorithms known as backfitting and local scoring. The beauty of this approach is that fitting can be done through a series of

Table 1
Comparison of models for count data

	Link	Inverse link	Weight
Poisson	$\eta = \ln(\mu)$	$\mu = e^{\eta}$	$\mu = e^{\eta}$
Negative binomial			
Log link	$\eta = \ln(\mu)$	$\mu = e^{\eta}$	$\frac{1}{\frac{1}{\mu}+\frac{1}{k}} = \mu\left(\frac{k}{\mu+k}\right) = \frac{ke^{\eta}}{e^{\eta}+k}$
Canonical link	$\eta = \ln\left(\frac{\mu}{\mu+k}\right)$	$\mu = \frac{k}{e^{-\eta}-1}$	$\mu + \frac{\mu^2}{k} = \mu\left(\frac{\mu+k}{k}\right) = \frac{ke^{\eta}}{(e^{\eta}-1)^2}$

single predictor (weighted) scatterplot smooths, and so the smoothing component can be handled through a single routine. In this paper, we use local polynomial fitting for this task.

The negative binomial model is only an exponential family when k is known. Therefore, the methodology of Hastie and Tibshirani (1990) can be used to fit the model for a given k. Conversely, if μ is known, estimation of k reduces to an ordinary maximum likelihood problem, the maximization of (1).

We propose fitting both k and μ by iterating between these two processes. Such an approach might be called alternating profile likelihood and will reduce to backfitting in the least squares case.

The structure of the full algorithm is as follows.

Iterate the alternating profile likelihood algorithm. Each iteration requires implementation of the local scoring algorithm.

Iterate the local scoring algorithm. Each iteration requires implementation of the backfitting algorithm for a weighted additive model.

Iterate the backfitting algorithm for a weighted additive model. Each iteration requires a weighted local polynomial smooth.

We now describe the three outer algorithms. Weighted local polynomial smoothing is described in Section 4. For ease of reading, some of the details such as convergence criteria are postponed until Section 5.

The Alternating Profile Likelihood Algorithm

Step 0: Initialize $\hat{k}$.
Step 1: Obtain $\hat{\eta}_i = \hat{\alpha} + \Sigma_{j=1}^{d} \hat{f}_j(X_{ji}; p_j h_j)$, a fitted generalized additive model, using the negative binomial likelihood with $k = \hat{k}$ using the local scoring algorithm and set $\hat{\mu}_i = e^{\hat{\eta}_i}$.
Step 2: Set $\hat{k} = \text{argmax}_k\, \ell(k, \hat{\mu})$.
Step 3: Repeat steps 1 and 2 until convergence.

The Local Scoring Algorithm

Step 0: Initialize $\hat{\alpha} = \ln(\bar{Y})$, $\hat{f}_1 = \cdots = \hat{f}_d = 0$.
Step 1: Set $\hat{\eta} = \hat{\alpha} + \Sigma_{j=1}^{d} \hat{f}_j$, $w = ke^{\hat{\eta}}/(e^{\hat{\eta}}+k)$, $Z = \hat{\eta} + (Y - e^{\hat{\eta}})/e^{\hat{\eta}}$, and $\hat{\alpha} = \bar{Z}$ and fit the w-weighted additive model with dependent variable Z and independent variables $X_1, \ldots, X_d$.
Step 2: Repeat step 1 until convergence.

Let Y and X_j, $j = 1, \ldots, d$, each be $n \times 1$ vectors. Let S_j^w denote the smoother matrix for a w-weighted local polynomial regression of Y on X_j with bandwidth h_j and degree p_j. The details of S_j^w are described in Section 4. The following algorithm describes how to fit the w-weighted additive model, $\mathrm{E}(Y \mid X_1, \ldots, X_d) = \alpha + \Sigma_{j=1}^{d} f_j$.

The Backfitting Algorithm for Weighted Additive Models

Step 0: Let $\hat{\alpha}$ and $\hat{f}_j$, $j = 1, \ldots, d$, be some initial values.
Step 1: Cycle $j = 1, \ldots, d$: $U = Z - \hat{\alpha}\mathbf{1} - \Sigma_{k \neq j} \hat{f}_k$, $\hat{f}_j = S_j^w U - \overline{S_j^w U}\mathbf{1}$, where $\overline{S_j^w U}$ is the mean of the smooth and $\mathbf{1}$ is the vector of ones.
Step 2: Repeat step 1 until convergence.

4. Weighted Local Polynomial Scatterplot Smoothing

We now give a brief description of the method by which a scatterplot may be smoothed using local polynomial fitting. Detailed descriptions of this approach to smoothing can be found in Wand and Jones (1995) and Fan and Gijbels (1996). However, this section does describe the extension to weighted scatterplot smoothing, which is not normally discussed in literature on local polynomial fitting.

Let Y, X, and w each be $n \times 1$ vectors, where the entries of w are positive. The w-weighted local polynomial smooth of Y on X of degree p and with bandwidth h is $S_j^w Y$, where S_j^w is an $n \times n$ matrix with (i, i') entry given by $(S_j^w)_{ii'} = e_1^{\mathrm{T}}(X_{X_i}^{\mathrm{T}} W_{X_i} X_{X_i})^{-1} X_{X_i}^{\mathrm{T}} W_{X_i} e_{i'}$, where

$$X_x = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \cdots & (X_n - x)^p \end{bmatrix},$$

$$W_x = \operatorname*{diag}_{1 \leq i \leq n} K\left(\frac{X_i - x}{h}\right) w_i,$$

and e_i is a column vector with one in the ith position and zeroes elsewhere. Here K is a positive symmetric function known as a kernel function.

In practice, significant gains in speed are realized by binning the data onto a fine grid and working with the corresponding counts. A full explanation of this approach is given in Fan and Marron (1994).

5. Details of Implementation

We now fill in the implementational details that were omitted in Sections 3 and 4.

5.1 *Convergence Criteria*

Following Hastie and Tibshirani (1990, p. 141), convergence of the local scoring and backfitting algorithms was assessed through evaluation of the criterion $\Delta = \Sigma_{j=1}^{d} \|f_j^{\text{new}} - f_j^{\text{old}}\| / \Sigma_{j=1}^{d} \|f_j^{\text{old}}\|$, where $\|f\| = (\Sigma_{i=1}^{n} f^2(x_i))^{1/2}$, the length of the vector f. We used a default stopping rule for the iterations of Δ falling below 0.01.

Buja, Hastie, and Tibshirani (1989) and Opsomer and Ruppert (1997) provide conditions under which the backfitting algorithm converges. According to their theory, the potential hindrance to convergence is multicollinearity and nonlinear forms of relatedness (concurvity) among the predictors. When analyzing the adduct data with all four covariates, we did experience problems with convergence of the backfitting algorithm. Although Δ fell below 0.1, it failed to get below 0.01. A graphical check showed that the functions oscillated between two practically identical sets of curves. So while strict mathematical convergence was not obtained (according to the above convergence criterion), we are satisfied that convergence in a practical sense was obtained. Buja et al. (1989) and Hastie and Tibshirani (1990, pp. 124–125) proposed a modified backfitting algorithm that is more efficient when data are correlated, although this has not been explored.

5.2 *Binned Local Polynomial Smoothing*

All scatterplot smooths were obtained using a binned implementation of the local linear kernel smoother over a set of 401 equally spaced grid points. The kernel is the normal kernel truncated to the interval $[-4, 4]$. To avoid problems with sparse design, we used a locally adaptive bandwidth chosen so that the $100\alpha\%$ of the scatterplot data were contained in the kernel window, where $0 < \alpha < 1$. We refer to α as the span.

5.3 *Choice of the Span*

While the adoption of a span-based locally adaptive bandwidth reduces the bandwidth selection problem to a single parameter, we are still faced with the problem of its choice. Although there has been a considerable amount of research into automatic bandwidth selection during the past 15 years, most of it has been confined to simple single predictor models with little work done on the practicalities of bandwidth selection for additive models. Recently, Opsomer and Ruppert (1998) developed an automatic algorithm for choosing the bandwidths in additive models but did not treat the generalized case. Because of the immaturity of automatic bandwidth selection for generalized additive models, we chose the span so that all unweighted smooths have 4 d.f., defined to be the trace of the smoother matrix, and made judicious adjustments to these smoothing parameters based on the graphical output.

5.4 *Standard Error Bars*

Pointwise standard error bars at $k = \hat{k}$ were obtained using the methodology advocated by Hastie and Tibshirani (1990, Section 6.8.2) based on linearization arguments. The details are given there, but essentially it involves an estimation of the asymptotic covariance matrix based on the covariance of the working response variable (denoted by Z in the local scoring algorithm) at convergence. In the parametric model in which the negative binomial mean is linear in the covariates, β and k are asymptotically independent. This is approximately true in the negative binomial additive model, which implies that estimation of k does not affect the standard error bars for the mean.

6. Semiparametric Models

In a semiparametric or partial linear generalized additive model, we could model $\eta = g(\mu) = \beta_0 + \Sigma_{j=1}^{d_1} \beta_j X_j + \Sigma_{j=d_1+1}^{d} f_j(X_j)$. We report an estimate of $\beta = \beta_0, \ldots, \beta_{d_1}$ motivated by the backfitting algorithm (Hastie and Tibshirani, 1990, Section 6.7), which, in the unweighted case, is the estimator given by Green et al. (1985). In our notation,

$$\hat{\beta} = \left\{ X_p^{\mathrm{T}} W \left(I - S_{(p)}^{W} \right) X_p \right\}^{-1} X_p^{\mathrm{T}} W \left(I - S_{(p)}^{W} \right) Z, \quad (2)$$

where X_p is the parametric part of the design matrix $(\mathbf{1}, X_1, \ldots X_{d_1})$, Z is the adjusted dependent variable, and W^{-1} is the variance of Z, where W is the diagonal weight matrix. $S_{(p)}^{W}$ is the weighted generalized smoothing matrix resulting from a smooth on $X_{(p)}$, where $X_{(p)}$ refers to all the X's except X_p. Thus, $S_{(p)}^{W} Z$ is the vector of fitted values resulting from smoothing Z on $X_{(p)}$ with weights W, and $S_{(p)}^{W} X_p$ is the matrix of fitted values from the smooth of each $(X_i \in X_p)$ on $X_{(p)}$. Note that $S_{(p)}^{W}$ is the same matrix in both cases (Opsomer and Ruppert, 1999). Since $S_{(p)}^{W}$ could not be calculated explicitly, we computed $S_{(p)}^{W} X_p$ and $S_{(p)}^{W} Z$ by backfitting. In the model $g(\mu) = X_p \beta_p + X_{(p)} \beta_{(p)}$, $\beta_p = (X_p^{\mathrm{T}} W (I - H_{(p)}^{W} W) X_p)^{-1} X_p^{\mathrm{T}} W (I - H_{(p)}^{W} W) Z$, where $H_{(p)}^{W}$ is the weighted hat matrix ($H_{(p)}^{W} = X_{(p)} (X_{(p)}^{\mathrm{T}} W X_{(p)})^{-1} X_{(p)}^{\mathrm{T}}$) with $\mathrm{var}(\hat{\beta}_p) = (X_p^{\mathrm{T}} W (I - H_{(p)}^{W} W) X_p)^{-1}$. This estimator can be derived from added variable plots (Weisberg, 1985, Section 2.4), specifically from a weighted regression of the residuals from the weighted regression of Z on $X_{(p)}$, on the matrix of residuals from the weighted regression of each $(X_i \in X_p)$ on $X_{(p)}$. The GAM analog to this, given in (2), is approximately equivalent to doing a weighted regression of the residuals from a weighted smooth of Z on $X_{(p)}$ on the matrix of residuals from the weighted smooth of each $(X_i \in X_p)$ on $X_{(p)}$. An alternative estimator to β that involves squaring the smoother matrix is asymptotically superior to (2) (Green et al., 1985), but little difference was found between the two estimators in an application (Hobert et al., 1997). We could not calculate the alternative since we could not calculate $S_{(p)}^{W}$ explicitly.

The variance based on the estimator of β in (2) can be approximated by $\widehat{\mathrm{var}}(\hat{\beta}) = \{X_p^{\mathrm{T}} W (I - S_{(p)}^{W}) X_p\}^{-1}$. This variance is exact only if $S_{(p)}^{W}$ can be written as the product of a smoothing matrix S and W, where S is symmetric and idempotent (all of which are true in the semiparametric GLM case when $S = H_{(p)}^{W}$). The semiparametric models were fit in a separate S-PLUS program.

7. Analysis of Adduct Data

The adduct dataset consists of data from ex-smokers, all of whom had lung cancer. We modeled the number of polyaromatic hydrocarbon adducts from four variables: age of smoking initiation (`age.start`), years of smoking (`cig.`

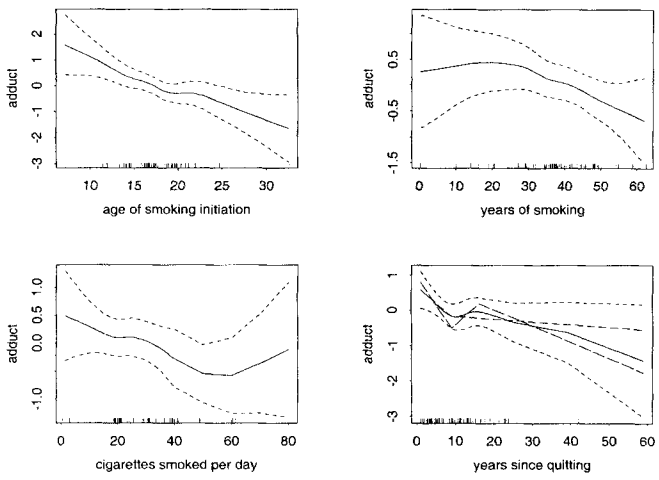

Figure 2. Fitted values in the negative binomial model of adduct counts using a smooth fit for all variables. Solid lines show the effect of age of smoking initiation, years of smoking, cigarettes smoked per day, and years since quitting, respectively. Dashed lines show plus and minus two standard errors. For the plot versus years since quitting, the short dashed line shows the fitted values for a broken stick model with one break at 9 years and the long dashed line shows the fitted values for breaks at 9 and 16 years for a semiparametric model with smooth relationships for the remaining variables.

`time`), number of cigarettes smoked per day (`cig.per.day`), and years since quitting (`yrs.quit`). The number of adducts ranged from 0 to 781. We excluded one person who started smoking at the age of 60, leaving 77 observations with `age.start` between 7 and 33.

Overdispersion is a common phenomenon (McCullagh and Nelder, 1989, p. 124). One estimate of overdispersion, used in quasilikelihood, is $\hat{\sigma}^2 = X^2/(n-p)$ (McCullagh and Nelder, 1989, p. 328), where X^2 is the sum of the Pearson residuals, $X^2 = \Sigma\,(y_i - \hat{\mu}_i)^2/V(\hat{\mu}_i)$, and $(n-p)$ is the residual degrees of freedom for the model. Under a Poisson model, $\hat{\sigma}^2$ would be one, whereas in the adducts dataset $\hat{\sigma}^2 = 98.7$. In the negative binomial model, the variance for an observation with mean μ_i is $\mu_i + \mu_i^2/k$. Under a Poisson model, k would be ∞, whereas in the adducts dataset $\hat{k} = 0.89$. The considerable overdispersion in this dataset suggests that variables not included in the model influence the number of adducts.

Plots from the negative binomial model with smooth functions for all four covariates (Figure 2) indicate that people who start smoking at a young age tend to have more adducts, after controlling for the other variables. This suggests that, in young smokers for whom the lung is still growing, adducts are either formed more rapidly than in mature lungs or adducts formed at that time are less able to be removed later than adducts formed in mature lungs (Wiencke et al., 1999). For years since quitting, it appears that the number of adducts declines linearly on a log scale for 8–10 years, then either rises slightly before declining slightly or remains relatively constant. These interpretations should be taken with caution since we are using cross-sectional data to make longitudinal interpretations.

We were interested in exploring more fully a semiparametric model in which the relationship between the number of adducts and `yrs.quit` is modeled parametrically and the relationships between the number of adducts and the other covariates are modeled nonparametrically. We fit two piecewise linear regression (broken stick) models for `yrs.quit` and are interested in comparing the relationship between the number of adducts and `yrs.quit` using a smooth fit for `yrs.quit` to fits from broken stick models with one and with two breaks, respectively, while allowing for smooth fits for the other covariates. Although the fits from both broken stick models are within the standard errors from the fully smoothed model, only the model with one break appears to reasonably approximate the smooth fit (Figure 2).

For a model with one break, we used a break at 9 years, which was the location of the breakpoint based on visual inspection of the smooth plot. Defining $(\texttt{yrs.quit} - 9)_+$ to be `yrs.quit` if `yrs.quit` > 9 and 0 otherwise, we parameterized this model as $g(\mu) = \beta_0 + \beta_1 \texttt{yrs.quit} + \beta_2(\texttt{yrs.quit} - 9)_+ + \cdots$, where $\cdots$ indicates smooth functions of the other variables. Under this model, the estimated slope for the first 9 years (β_1) was -0.1017 with SE $= 0.0580$. The difference in slopes between the two time periods (β_2) was 0.1006 with SE $= 0.0667$. This gives weak evidence of a changepoint at 9 years (p-value ≈ 0.14). Under this model, the number of adducts declines by 9.67% per year for 9 years, then remains almost constant after that. Under this model, 40.04% of the original number of adducts would remain after 9 years.

To evaluate how well the negative binomial GAM model fit the data, we plotted the absolute value of the Studentized standardized deviance residuals versus the fitted values on the η scale (McCullagh and Nelder, 1989, p. 398) from a negative binomial adduct model with a broken stick relationship for `yrs.quit` (break at 9 years), a quadratic relationship for `cig.time`, and linear relationships for `cig.per.day` and `age.start`. For comparison, we plotted the same quantities from the quasi-likelihood, overdispersed Poisson model of the same form (Figure 1). The plot from the quasi-likelihood model shows that the variance of the residuals clearly increases with the fitted values. The plot from the negative binomial model shows much less heteroscedasticity but shows a clustering of points with large negative residuals corresponding to people with adduct counts of zero. We conclude that the negative binomial model is an improvement over quasilikelihood for this data.

Acknowledgement

We are grateful to Professor Louise Ryan for discussions that initiated this project.

Résumé

Le modèle additif généralisé est étendu pour prendre en compte les réponses binomiales négatives. L'extension est compliquée par le fait que la distribution binomiale négative a deux paramètres et n'est pas dans la famille exponentielle. La méthodologie est appliquée à des données concernant la fréquence des "adduits" sur l'ADN et les variables liées au tabagisme chez des ex-fumeurs avec un cancer du poumon. Une investigation plus détaillée est faite sur la relation paramétrique entre le nombre d' "adduits" et les années depuis l'arrêt, alors que l'on conserve une relation lissée entre les "adduits" et les autres covariables.

References

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics* **17,** 453–510.

Cox, D. R. (1983). Some remarks on overdispersion. *Biometrika* **70,** 269–274.

Denissenko, M. F., Pao, A., Tang, M.-S., and Pfeifer, G. P. (1996). Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. *Science* **274,** 430–432.

Denissenko, M. F., Chen, J. X., Tang, T.-S., and Pfeifer, G. P. (1997). Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *The Proceedings of the National Academy of Sciences, USA* **94,** 3893–3898.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* **81,** 709–721.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* London: Chapman and Hall.

Fan, J. and Marron, J. S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics* **3,** 35–56.

Fan, J., Heckman, H. E., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90,** 141–150.

Green, P., Jennison, C., and Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society, Series B* **47,** 299–315.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* New York: Chapman and Hall.

Hobert, J. P., Altman, N. S., and Schofield, C. L. (1997). Analyses of fish species richness with spatial covariate. *Journal of the American Statistical Association* **92,** 846–854.

Kure, E. H., Ryberg, D., Hewer, A., Phillips, D. H., Skaug, V., Baera, R., and Haugen, A. (1996). p53 mutations in lung tumours: Relationship to gender and lung DNA adduct levels. *Carcinogenesis* **17,** 2201–2205.

Lambert, D. and Roeder, K. (1995). Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association* **90,** 1225–1236.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* **15,** 209–225.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models,* 2nd edition. London: Chapman and Hall.

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74,** 221–232.

Opsomer, J.-D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics* **25,** 186–211.

Opsomer, J. D. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association* **93,** 605–619.

Opsomer, J.-D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics,* in press.

Schut, H. A. J. and Shiverick, K. T. (1992). DNA adducts in humans as dosimeters of exposure to environmental, occupational, or dietary genotoxins. *FASEB J* **6,** 2942–2951.

Tang, D., Santella, R. M., Blackwood, A. M., Young, T.-L., Mayer, J., Jaretzki, A., Grantham, S., Tsai, W.-Y., and Perera, F. P. (1995). A molecular epidemiological case–control study of lung cancer. *Cancer Epidemiology, Biomarkers, and Prevention* **4,** 341–346.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* London: Chapman and Hall.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61,** 439–447.

Weisberg, S. (1985). *Applied Linear Regression.* New York: Wiley.

Wiencke, J. K., Thurston, S. W., Kelsey, K. T., Varkonyi, A., Wain, J. C., Mark, E. J., and Christiani, D. C. (1999). Early age at smoking initiation and tobacco carcinogen DNA damage in the lung. *Journal of the National Cancer Institute* **91,** 614–619.

Wogan, G. N. (1992). Molecular epidemiology in cancer risk assessment and prevention: Recent progress and avenues for future research. *Environmental Health Perspectives* **98,** 167–178.

Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society, Series B* **58,** 481–494.

Received August 1998. *Revised June* 1999.
Accepted June 1999.