A New Variance Estimator for Parameters of Semiparametric Generalized Additive Models

W. Dana FLANDERS, Mitch KLEIN, and Paige TOLBERT

Generalized additive models (GAMs) have become popular in the air pollution epidemiology literature. Two problems, recently surfaced, concern implementation of these semiparametric models. The first problem, easily corrected, was laxity of the default convergence criteria. The other, noted independently by Klein, Flanders, and Tolbert, and Ramsay, Burnett, and Krewski concerned variance estimates produced by commercially available software. In simulations, they were as much as 50% too small. We derive an expression for a variance estimator for the parametric component of generalized additive models that can include up to three smoothing splines, and show how the standard error (SE) estimated by this method differs from the corresponding SE estimated with error in a study of air pollution and emergency room admissions for cardiorespiratory disease. The derivation is based on asymptotic linearity. Using Monte Carlo experiments, we evaluated performance of the estimator in finite samples. The estimator performed well in Monte Carlo experiments, in the situations considered. However, more work is needed to address performance in additional situations. Using data from our study of air pollution and cardiovascular disease, the standard error estimated using the new method was about 10% to 20% larger than the biased, commercially available standard error estimate.

Key Words: Epidemiologic methods; Generalized additive models; Semiparametric models; Variance.

1. INTRODUCTION

Generalized additive models (GAMs), a relatively new approach to nonparametric or semiparametric smoothing and data analysis (Hastie and Tibshirani 1990), have become widely used, particularly in time series analyses of acute health effects of air pollution. In semiparametric models, the focus of this article, the mean of the dependent variable is modeled as a parametric, linear function of some predictors plus a sum of functions of other predictors, which in some applications may be confounders or nuisance factors. The form of the function used for these other predictors is quite general, hence the term semiparametric.

W. Dana Flanders is Professor, Rollins School of Public Health, Department of Epidemiology, Emory University, 1518 Clifton Road, Atlanta, GA 30327 (E-mail: flanders@sph.emory.edu). Mitch Klein is Assistant Professor, and Paige Tolbert is Associate Professor, Rollins School of Public Health, Department of Epidemiology and Department of Environmental and Occupational Health, Emory University, Rollins School of Public Health, 1518 Clifton Road, Atlanta, GA 30327.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 10, Number 2, Pages 246–257
DOI: 10.1198/108571105X47010

Schwartz proposed application of GAMS to time series studies assessing the association of air pollution with mortality or other outcome measures in 1994 (Schwartz 1994a), and initially presented GAM models as a sensitivity analysis augmenting a parametric approach (Schwartz 1994b). In the intervening years, GAMs have gained widespread popularity for use in these types of time series studies (e.g., Borja-Aburta et al. 1998; Michelozzi et al. 1998; Burnett et al. 1999; Conceicao et al. 2001; Moolgavkar 2000; Pope, Hill, and Villegas 1999; Samet et al. 2000).

GAMs can generally be fit using S-Plus or using `PROC GAM` in SAS (SAS 2001). As discussed in the following, the models can be fit using a backfitting algorithm. Hastie and Tibshirani (1990) discussed conditions that assure convergence of this approach. Two problems have recently surfaced, however, concerning implementation of these models. The first problem, easily corrected, was that the default convergence criteria were not adequately strict (Dominici, McDermott, Zeger, and Samet 2002; Katsouyanni et al. 2002). The other problem concerns the variance estimates produced by these programs for the parametric component of the semiparametic GAMs. The problem was noted independently by Klein, Flanders, and Tolbert (2002) and by Ramsay, Burnett, and Krewski (2003). They showed that the variance estimates could be as much as 50% lower than the simulated variance in some of the situations considered. This article addresses the second problem by deriving a relatively easily implementable variance estimator for these models.

One of the specific problems that motivated this work are the numerous published or ongoing studies of the associations between health outcomes, such as respiratory disease, and air pollution (e.g., Borja-Aburta et al. 1998; Michelozzi et al. 1998; Burnett et al. 1999; Conceicao et al. 2001; Moolgavkar 2000; Pope, Hill, and Villegas 1999; Samet et al. 2000; Tolbert et al. 2000). Some of the studies published by others have used generalized additive models to assess the association between air pollution and disease, but an appropriate variance estimator has been unavailable.

The purpose of this article is three-fold. First, we present an asymptotic variance estimator for the parametric component of GAM semiparametric models, providing an explicit formulation for up to three splines. Second, we empirically evaluate the performance of this estimator in finite samples using Monte Carlo simulations and base these simulations on actual data from an ongoing study of air pollution. Finally, we apply the estimator to an ongoing study of air pollution and emergency department visits. We illustrate that, in this study, the variance we estimate differs from the corresponding estimates produced by commercially available software.

2. METHODS

In the semiparametric situations of interest here, the generalized additive model is given by:

$$E(Y_i \mid X_i, Z_{1i}, \ldots, Z_{Ji}) = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta X_i + f_1(Z_{1i}) + \cdots + f_J(Z_{Ji})), \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where Y_i is the number of events for the ith observation; g is a strictly monotone link function; $\eta_i = \alpha + \beta X_i + f_1(Z_{1i}) + \cdots + f_J(Z_{Ji})$; β is $(p \times 1)$ parameter of interest to be estimated; X_i is a $(1 \times p)$ vector of predictors; Z_{ji} is the value of the jth covariate for the ith observation; and $f_j(Z_{ji})$ is an arbitrary (smoothing) function with continuous second derivatives, for $j = 1$ to J. (Here, we limit consideration to $J \leq 3$, but results should extend in an analogous way for $J > 3$.) We assume here that the Y_i given X_i, and Z_{1i} are independent with a Poisson distribution whose mean is given by Equation (2.1). The Poisson distribution is typically used in applications in air pollution epidemiology. However, results hold with obvious modifications for other distributions in the exponential family (Hastie and Tibshirani 1990).

We derive an explicit expression for the variance estimator for a class of estimators of β in the model given by Equation (2.1), estimated by penalized likelihood (Hastie and Tibshirani 1990). That is, one maximizes

$$j(\beta, f) = l(\eta; Y) - \frac{1}{2} \sum_{j=1}^{3} \lambda_j \int \left\{ f_j''(z) \right\}^2 dz, \tag{2.2}$$

over η and over f_j in the class of functions with continuous second derivatives f_j''. Here, the λ_j are smoothing parameters which must be specified, or estimated from the data (Hastie and Tibshirani 1990). The functions that maximize Equation (2.2), are cubic smoothing splines (Hastie and Tibshirani 1990). An equivalent problem (Hastie and Tibshirani 1990) is to maximize

$$l(\eta; Y) - \frac{1}{2} \sum \lambda_j f_j^t K_j f_j, \tag{2.3}$$

where K_j are the $n \times n$ quadratic penalty matrices given, for example, by Buja, Hastie, and Tibshirani (1989) for $j = 1, 2, 3$; $\mathbf{f}_j$ are the $n \times 1$ vectors $f_j(Z_i), i = 1, 2, \ldots n$, $j = 1, 2, 3$; and the superscript "t" denotes a matrix transpose. Using $[\mathrm{A}]^-$ to denote the generalized inverse, the model can be fit using a local scoring procedure that incorporates the weighted smoothing matrices $S_j = (A + K_j)^- A$ in a backfitting algorithm as described by Hastie and Tibshirani (1990). We assume sufficient regularity and choice of smoothing parameters so that the local scoring procedure and the backfitting algorithm that it includes converge in probability as the sample size increases. In particular, $\hat{\beta}_i \to \beta_o$ for each i and $\hat{f}_{1,i} \to f_{1,0}$, $\hat{f}_{2,i} \to f_{2,0}$, and $\hat{f}_{3,i} \to f_{3,0}$, in probability where $\hat{\beta}_{i-1}, \hat{f}_{1,i}, \hat{f}_{2,i}, \hat{f}_{3,i}$ are the parameter estimates at step i; and β_o, $f_{1,0}$, $f_{2,0}$, and $f_{3,0}$ are the corresponding true values.

Our approach to estimating the large sample variance of $\hat{\beta}$ is straightforward: we find a large sample, approximate, linear expression for $\hat{\beta}$ in terms of $Y, E(Y)$, and known (or consistently estimable) functions.

By arguments presented by Hastie and Tibshirani (1990), the one-step updates for the Newton-Raphson step of the fitting algorithm at the ith step are given by:

$$X\hat{\beta}_i = X\left\{X^tA(I-S_2)X\right\}^{-1}X^tA\left(Z^c-\hat{f}_{1,i-1}-\hat{f}_{2,i-1}-\hat{f}_{3,i-1}\right), \quad (2.4)$$

$$\widehat{f}_{1,i} = S_1\left(Z^c-X\hat{\beta}-\hat{f}_{2,i-1}-\hat{f}_{3,i-1}\right), \quad (2.5)$$

$$\widehat{f}_{2,i} = S_2\left(Z^c-X\hat{\beta}-\hat{f}_{1,i-1}-\hat{f}_{3,i-1}\right), \quad (2.6)$$

and

$$\widehat{f}_{3,i} = S_3\left(Z^c-X\hat{\beta}-\hat{f}_{1,i-1}-\hat{f}_{2,i-1}\right), \quad (2.7)$$

where A is the $n \times n$ matrix $\partial^2 l/\partial\eta\partial\eta^t$; l is the Poisson log-likelihood; Z^c is the $n \times 1$ vector of linearized dependent variables $Z^c = \eta^c + A^{-1}u$; and where u is the $n \times 1$ vector $\partial l/\partial\eta^c$, all evaluated using the current estimates of β, f_1, f_2, f_3, and η.

Equations (2.4)–(2.7) are a system of four equations in four unknowns (vectors). A straightforward, though tedious derivation yields the following closed form estimate for the Newton-Raphson update of $\hat{\beta}_i$:

$$\hat{\beta}_i = \left[X^tA(I\text{–}V_1\text{–}V_2-V_3)X\right]^-\left(X^tA\,(I\text{–}V_1\text{–}V_2-V_3)\right)Z^c, \quad (2.8)$$

where

$$\begin{aligned}
V_1 &= \left[I-S_1S_3\text{–}S_1\,(I-S_3)\,(I-S_2S_3)^-\,S_2\,(I-S_3)\right]^- \\
&\quad S_1\left(I-S_{3^-}\,(I-S_3)\left[I-S_2S)3\right]^-\,S_2\,(I-S_3)\right) \\
V_2 &= \left[I-S_2S_1\text{–}S_2\,(I-S_1)\,(I-S_3S_1)^-\,S_3\,(I-S_1)\right]^- \\
&\quad S_2\left(I-S_1-(I-S_1)\,[I-S_3S_1]^-\,S_3\,(I-S_1)\right) \\
V_3 &= \left[I-S_3S_2-S_3\,(I-S_2)\,[I-S_1S_2]^-\right. \\
&\quad \left.S_1\,(I-S_2)\right]^-\,S_3\left(I-S_2-(I-S_2)\,[I-S_1S_2]^-\,S_1\,(I-S_2)\right),
\end{aligned}$$

and Z^c is expressed in terms of $\hat{\beta}_{i-1}$, $\hat{f}_{1,i}$, $\hat{f}_{2,i}$, $\hat{f}_{3,i}$, and η.

Equation (2.8) is an explicit form for multiple smoothing splines (in terms of the individual smoothers S_1, S_2, and S_3), of the result given by Hastie and Tibshirani (1990) of the form: $\hat{\beta} = [X^tA(I-S)X]^-(X^tA(I-S))Z^c$. By taking $S = (I-V_1-V_2-V_3)$, one sees that the one-step update for β in Equation (2.8) is consistent with Hastie and Tibshirani (1990) who gave explicit results for a single spline, but noted that the same form of the equation would hold for multiple splines. An explicit form for multiple smoothing splines (in terms of the individual smoothers) can be obtained by applying the recursive equation described in the Appendix.

Because the estimates converge in probability by assumption, we can imagine starting the process at the true value η^t. Then, with a large sample size, the one-step estimator $\hat{\beta}$ is

given by:

$$\hat{\beta} = \left[X^t A\left(I - V_1 - V_2 - V_3\right) X\right]^{-} \left(X^t A\left(I - V_1 - V_2 - V_3\right)\right) Z^t, \tag{2.9}$$

where $Z^t = \eta^t + A^{-1}u$—all parameters and expressions now evaluated at their true values. Thus, following the arguments of by McCullagh and Nelder (1989), the one-step estimator is a linear function of the observations, and subsequent updates should be negligible for a large sample size (asymptotically).

Thus, the asymptotic variance is given by:

$$\text{var}(\hat{\beta}) \approx W * \text{var}(Z^T) * W^t$$

where

$$W = W = \left[X^t A\left(I - V_1 - V_2 - V_3\right) X\right]^{-} \left(X^t A\left(I - V_1 - V_2 - V_3\right)\right). \tag{2.10}$$

Furthermore, we can estimate $\text{var}(Z^T)$ by A^{-1} and W, with all parameters evaluated at the estimated values.

3. SIMULATIONS

To evaluate the performance of this variance estimator (Equation (2.10)) in finite sample sizes, we performed Monte Carlo simulations in two different sets of situations—one based on real data, the other using hypothetical data. In the first situation, we used data from our ongoing study of emergency department (ED) visits for cardiorespiratory diseases and air pollution in Atlanta (Tolbert et al. 2000). The outcome variable was either daily ED visits for all types of cardiovascular disease (CVD) or for asthma from August 1, 1998, to July 31, 1999. We analyzed data for daily nitrogen dioxide (NO_2) and separately for daily particulate matter (PM_{10}), resulting in a total of four experimental conditions based on real data (Table 1). We controlled for other ED visits, temperature, dew point, day of the week, and time using smoothing splines in a GAM model:

$$E\left(Y_i \mid X_i, Z_{1i}, \ldots, Z_{Ji}\right)$$

$$= \exp\left(\alpha + \beta \mathbf{X}_i + f_1\left(Z_{1i}\right) + f_2\left(Z_{2i}\right) + f_3\left(Z_{3i}\right)\right), \quad \text{for} \quad i = 1, 2, \ldots, n, \tag{3.1}$$

where $\mathbf{X}_i^t =$ (air pollutant on day i, all ED visits on day i, I_{Tue}, I_{Wed}, I_{Thu}, I_{Fri}, I_{Sat}, I_{Sun}), $I_{\text{Tue}} - I_{\text{Sun}}$ are indicators for day i; $Z_{1i} =$ number of days since the start of the study, $Z_{2i} =$ the mean temperature for day i, and $Z_{3i} =$ the mean dew point on day i. To avoid ties and division by 0, we added a small random number (mean 0, variance .01) to each temperature and dew point. We then fit this GAM to the observed data, using 14 degrees of freedom for the time spline, corresponding approximately to monthly knots; and, 5 each for temperature and dew point splines. (We used 7 degrees of freedom for the CVD outcome, corresponding approximately to seasonal knots). We then saved these model predicted values, and used them in the simulations as the expected values, and generated independently for each day,

Table 1. Monte Carlo Simulation Results, Expected Values Calculated from Actual ED visits and Air Pollutants, in Atlanta

Outcome	Pollutant	True β_1[b]	SE ($\hat{\beta}$)[c]	$\widehat{SE}$ − old[d]	Coverage[d] 95% CI − old	$\widehat{SE}$ − new[e]	Coverage[e] 95% CI − new
CVD	PM_{10}	.011	.0124	.0111	90.7%	.0123	95.3%
CVD	NO_2	.006	.0336	.0313	93.4%	.0345	95.5%
Asthma	PM_{10}	.040	.0219	.0187	89.4%	.0226	95.7%
Asthma	NO_2	.014	.0378	.0314	90.2%	.0383	94.5%

[a] CVD (Asthma): Emergency room visits for cardiovascular disease (asthma), used to determine expected daily counts
[b] Value of β_1 used to determine expected value of Y, for each set of 1,000 Monte Carlo experiments.
[c] Standard error of $\hat{\beta}_1$ for each set of 1,000 experiments.
[d] Mean estimated standard error, and coverage of 95% CI produced by SAS PROC GAM (3.1); 95% CI calculated as the point estimate $+/-$ 1.96 times the estimated standard error.
[e] Mean estimated standard error, and coverage of 95% CI based on new variance estimator; 95% CI calculated as the point estimate $+/-$ 1.96 times the estimated standard error.

Poisson random variables. We analyzed this randomly generated series, by fitting the GAM model (Equation (3.1)), and saving the standard error generated by SAS and that calculated using Equation (2.10). We repeated this process 1,000 times for each outcome-air pollutant combination. We focus on β_1, the log rate ratio for the association of ED visits with the air pollutant.

The second set of situations used was similar to that just described, except that hypothetical, simulated data replaced the observed data used above to generate the daily expected counts of ED visits. Specifically, for $i = 1, 2, \ldots 200$ we generated (hypothetical) exposure and covariates: $\varepsilon_i, x_{2,i}, t_{2,i}, t_{3,i}$ as independent, standard Gaussian variables; $x_{1,i} = \varepsilon_i + x_{2,i}$ and $t_{1,i} = i$. We then defined the expected value of Y_i as:

$$E\left(Y_i \mid x_{1i}, x_{2i}, t_{1i}, t_{2i}, t_{2i}\right) = \exp\left(\beta_0 + \beta_1 x_1 - \beta_2 x_2 + \frac{1}{2}\cos(t_1/5) + \frac{1}{2}\sin(t_2/5) + (t_3^3 - t_3^2 + 3/4t_3)/200\right). \tag{3.2}$$

We chose $\beta_0 = 3, 4$, or 5 and $\beta_1 = .0$ or .04, and $\beta_2 = .1$, resulting in a total of six additional experimental conditions (Table 2). We randomly generated 200 *preliminary* values of Y_i as independent Poisson variables with mean given by Equation (3.2). So that the model would be correctly specified, we then fit a GAM with a smoothing splines using six degrees of freedom each for t_1, t_2, and t_3 to these *preliminary* observations and used the model-predicted values as the expected values to generate the 200 observations Y_i used for each Monte Carlo experiment. We adjusted the smoothing parameters $(\lambda_1, \lambda_2, \lambda_3)$ to correspond to the six degrees of freedom that we specified for use in `PROC GAM`. (That is, we retained the same x_1, x_2, t_1, t_2, and t_3 for each experiment, but used the predicted values from the model fit to the *preliminary* data as the expected values for all 1,000 Monte Carlo experiments in each set. This procedure should ensure that the expected values were in the space of possible fits. Because of this process, the "true value" of β_1 for each simulation

Table 2. Monte Carlo Simulation Results, Expected Values Calculated Hypothetical Data

Experiment	Mean of Expected Y	True β_1 [a]	SE ($\hat{\beta}$) [b]	$\widehat{SE}$ − old [c]	Coverage [c] 95% CI − old	$\widehat{SE}$ − new [d]	Coverage [d] 95% CI − new
1	21.9	.0010	.0158	.0123	87.8%	.0152	93.3%
2	51.2	−.0001	.0096	.0073	85.7%	.0096	95.0%
3	157	.0050	.0057	.0042	84.6%	.0056	94.7%
4	22.3	.049	.0161	.0125	87.6%	.0163	95.3%
5	59.9	.064	.0098	.0076	89.1%	.0099	95.0%
6	165	.065	.0058	.0046	88.6%	.0060	95.3%

[a] Value of β_1 used for each set of 1,000 Monte Carlo experiments.
[b] Standard error of $\hat{\beta}_1$ for each set of 1,000 experiments.
[c] Mean estimated standard error, and coverage of 95% CI produced by SAS PROC GAM (3.1); 95% CI calculated as the point estimate +/− 1.96 times the estimated standard error.
[d] Mean estimated standard error, and coverage of 95% CI based on new variance estimator; 95% CI calculated as the point estimate +/− 1.96 times the estimated standard error.

set was equal to the initial estimate, and differed slightly from 0 or .04 depending on the baseline estimates.) Using `PROC GAM` in SAS (2001), we estimated β_1 1,000 times for each randomly generated series. We calculated the standard error of the estimated β_1's, and compared this standard error with the estimated standard errors produced by the SAS program, and with the new standard error estimate in Equation (2.10).

In the third set of experiments (Table 3), we considered estimation with a smaller number of time points (either 100 or 50), and chose either 6, 4, or 2 degrees of freedom for the smoothing splines. Otherwise, this last set of experiments is like the second set, specifically experiment 6. In two experiments (12 and 14), we used two alternative types of error structures, the normal and the binomial distributions.

4. RESULTS

Results of the first set of Monte Carlo experiments—based on the actual observations of ED visits and air pollutants in Atlanta (situation 1)—are shown in Table 1. These results illustrate several points. First, they illustrate a tendency for under-estimation of the standard errors by the SAS procedure `PROC GAM` (3.1). For example, the "old" standard estimate for β_1 (which relates emergency room visits to PM_{10}) averaged about .0111, compared to .0123 ,the sample standard error of estimated β_1's. On the other hand, the new standard error estimator was about the same, on average, as the sample standard error. A similar pattern held for the other pollutant, and when asthma emergency room visits were used to determine the expected values. The second point, related to the first, is that coverage of the 95% confidence limits based on the "old" variance estimates was consistently less than the nominal value of 95%. The coverage in these experiments was as low as 89% in one instance. In contrast, the coverage of confidence intervals based on the new variance estimator was close to the nominal level.

Results of the second set of Monte Carlo experiments—based on hypothetical data (situation 2)—are shown in Table 2. These results further support these same patterns.

Table 3. Monte Carlo Simulation Results, Expected Values Calculated Hypothetical Data

Experiment / Distribution	*N, df*	*True* β_1 [a]	*SE* $(\hat{\beta})$ [b]	$\widehat{SE}$ *– old* [c]	*Coverage* [c] *95% CI – old*	$\widehat{SE}$ *– new* [d]	*Coverage* [d] *95% CI – new*
8[†]/ Poisson	100, 6	.031	.0100	.0076	89.4%	.0100	95.1%
9 /Poisson	100, 4	.044	.0083	.0072	82.4%	.0085	93.4%
10/Poisson	50, 6	.031	.0134	.0072	71.6%	.0133	94.4%
11/Poisson	50, 4	.056	.0121	.0111	92.2%	.0119	94.8%
12/Poisson	50, 2	.076	.0119	.0112	93.3%	.0119	94.9%
13/Normal	100, 2	−.038	.0920	.0904	94.7%	.0913	94.9%
14/Binomial	100, 4	.037	.0116	.0088	79.9%	.0113	94.4%

[†] Each experiment is like Experiment 6 in Table 2, except for: the error distribution, the number of time points (*N*), and the degrees of freedom (df).

[a] Value of β_1 used for each set of 1,000 Monte Carlo experiments.

[b] Standard error of $\hat{\beta}_1$ for each set of 1,000 experiments.

[c] Mean estimated standard error, and coverage of 95% CI produced by SAS PROC GAM (3.1); 95% CI calculated as the point estimate $+/-$ 1.96 times the estimated standard error.

[d] Mean estimated standard error, and coverage of 95% CI based on new variance estimator; 95% CI calculated as the point estimate $+/-$ 1.96 times the estimated standard error.

Specifically, the old standard error estimates are consistently lower than the sample standard error of the estimated β's, and the coverage of the associated confidence intervals was consistently lower than the nominal 95% level.

In the third set of Monte Carlos experiments, the new estimator of the standard error yielded results close to the simulated standard deviation of the estimated β, when we reduced the number of time points from 100 as in earlier experiments to either 100 or 500 in these. In a few experiments with the lower number of time points and *if* we used more than 2–4 degrees of freedom for each spline, the average $\hat{\beta}$ differed slightly, but significantly from the true β (data not shown).

5. EXAMPLE: APPLICATION OF NEW METHOD

We applied the estimator to data from an ongoing study of air pollution and ED visits for cardiorespiratory diseases in Atlanta (Tolbert et al. 2000), the data used here from August 1, 1998, to July 31, 1999. Our rationale for using GAMs reflects their inherent appeal due in part to the semiparametric nature of the time dependency and consequent relaxation of assumptions, and the frequent use of these models in the air pollution literature. We calculated the standard error for parameter estimates from a semiparametic generalized additive model, with the Poisson distribution and log link, executed using `PROC GAM` in SAS, and compared this estimate to the standard error estimated with our new method. We chose the degrees of freedom for the splines to be similar to those we used in parametric Poisson regression; Use of generalized cross-validation, the default approach in SAS (SAS Institute 2001), suggested slightly fewer degrees of freedom, but led to the same results. We evaluated the association between nitrogen dioxide (NO_2) and ER visits for all CVD, using the three-day moving average of NO_2, in part due to a priori interest, and controlling for

time, for mean temperature, and for dew point using cubic splines with 7, 7, and 5 degrees of freedom, respectively. We also controlled for day of the week using indicator variables and the number of emergency room visits for noncardiovascular disease. To simplify calculations by avoiding ties, we added a small random number to the temperature and dew point (which did not change the estimate of the parameter or its standard error). We found little evidence of autocorrelation of residuals (Durbin-Watson $= 2.155$, $p = .40$ by simulation). In the model for CVD visits, the parameter estimate for NO_2 was .020 (rate ratio $= 1.020$) and the standard error estimated in `PROC GAM` was .018. The standard error estimated using the new estimator was .020, about 10% larger than that obtained using SAS. This difference is important for evaluating the stability of results, for interval estimation, and would affect any meta-analysis that used this result.

6. DISCUSSION

Our results provide evidence that the variance estimator in Equation (2.10) works well with finite samples—at least for the situations considered. More work, however, needs to be done to verify that its perfomance remains good under other conditions. Our results also further support and are consistent with the work of Klein et al. (2002) and of Ramsey et al. (2003) who showed that the variance estimation procedures used in commercially available programs could be inadequate. These tendencies of commercial software to underestimate variances have been attributed to concurvity in the data (Ramsay et al. 2003), and to inadequate linear approximations used for the smooth functions (Dominici, McDermott, and Hastie 2003). Recognition of these and other problems has motivated reanalyses of at least 20 studies of air pollution and health effects with the overall conclusion that use of GAMs, implemented with the faulty variance estimator, was associated with smaller standard errors than use of generalized linear models (Health Effects Institute 2003).

We have presented and evaluated a variance estimator for up to three splines, extending the previous work of Hastie and Tibshirani (1990) who presented explicit results for a single spline, and of Flanders, Klein, and Tolbert (2003). Implementation of this approach for more splines is straightforward; for example, perhaps using the recursive equations given, but as the number of splines increases, of course, computations become more and more onerous. Our result also applies directly to the case of only one or two splines by simply taking S_1 and/or S_2 equal to 0.

Our arguments depend heavily on the assumption of consistency of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, and convergence of the backfitting algorithm as argued by Hastie and Tibshirani (1990). We have not investigated performance of the variance estimator when that assumption might fail. Conditions other than those noted by Hastie and Tibshirani (1990) may also lead to consistency. In particular, we might also expect consistency if the number of (say, time) points remains fixed, but the expected mean increases for each point, other parameters remain constant, and the model is correctly specified with judicious choice of degrees of freedom for the splines. More work on consistency, not the focus of this article, remains. For example, the application to time series should probably be based on futher specification of

assumptions because, as the number of time points increases, the complexity and in general the number of parameters in the underlying model could potentially increase proportionately. Yet additional work could allow for potential serial autocorrelation, although our example did not suggest important residual autocorrelation. We also note that after development of the expression for variance (Flanders, Klein, and Tolbert 2002, Flanders et al. 2003; equation 10), we became aware that another estimator was in the process of being developed to correct the errors in the commercially available software (Dominici, McDermott, and Hastie 2002). Comparison now shows our formulation to be equivalent to theirs (Dominici, McDermott, and Hastie 2003): one simply substitutes a combined smoothing matrix, S, in place of $(I–V_1–V_2 - V_3)$ in the expression for W in Equation (2.10). Our independent derivation, results of our empiric evaluations and our example show that the commercially available estimates can be too small, and that the alternative estimator has good finite sample properties, at least in the situations considered. In particular, our work provides empiric evidence, complimenting work of Dominici et al. (2003), that use of Equation (2.10) for analysis of real data can lead to confidence limits with appropriate coverage properties, again at least for the situations considered.

Some investigators in air pollution epidemiology have so far avoided use of GAMs as the primary method of analysis because of concerns about the variance estimator (Klein et al. 2002), choosing instead to use Poisson regression models with splines for time with many knots and chosen, in part, based on a priori considerations. The Monte Carlo experiments suggest—in agreement with simulations of Klein et al. (2002) and of Ramsey et al (2003)—that the standard error provided by commercially available software can have substantial error. The estimator evaluated here performed nicely with finite samples in simulations completed so far. Importantly, as illustrated in our simulations, the correction has a substantial effect on the estimated standard errors and confidence interval when applied to data based on an ongoing study in Atlanta (Tolbert et al. 2000), and on hypothetical data. This underestimation is further illustrated in the example, using data from our ongoing study of air pollution in Atlanta. The simulations suggest that, at least for situations like those considered here, the assumptions may be adequately achieved and that the new estimator can perform well in such real situations. As noted by Lumley and Sheppard (2003), major challenges air polllution epidemiology remain, particularly including model selection in the face of measurement error and confounding.

APPENDIX

Equations for $\hat{\beta}$ and var$(\hat{\beta})$ in terms of the individual smoothing matrices like Equations (2.8) and (2.10), but that apply with any number of splines can be obtained recursively as follows. Start with expressions for $\hat{\beta}$ (such as Equation (2.8)) and var$(\hat{\beta})$ (i.e., Equation (2.10)) that apply with $J - 1$ splines, in terms of smoothing matrices $S_1, S_2, \ldots, S_{J-1}$, matrix A (the $n \times n$ matrix $\partial^2 l/\partial\,\partial\eta\partial\eta^t$, as in Equation (2.8)). Expressions for $\hat{\beta}$ and var$(\hat{\beta})$ that apply with one additional spline (say, S_J) is obtained by substituting: $(I - S_iS_J)^- S_i(I - S_J)$ in place of S_i for $i = 1, 2, \ldots, J - 1$ and $A(I - S_J)$ for A throughout.

This recursive approach is justified by writing the system of $J + 1$ linear Equations in unknowns $\hat{\beta}$ and $\hat{f}_1 \ldots, \hat{f}_J$, comparable to Equations (2.4)–(2.7); eliminating $\hat{f}_J$; and rearranging to obtain a reduced system of J equations in J unknowns. The new system of equations has the same form as the original, provided we identify $(I - S_i S_J)^- S_i(I - S_J)$ with S_i for $i = 1, 2, \ldots, J - 1$ and $A(I - S_J)$ with A throughout.

ACKNOWLEDGMENTS

This work was supported by grants from the U.S. Environmental Protection Agency (R82921301-0) and from the National Institute of Environmental Health Sciences (R01ES11294).

[Received April 2004. Revised November 2004.]

REFERENCES

Borja-Aburta, V.H., Castillejos, M., Gold, D.R., Bierzwinski, S. and Loomis, D. (1998), "Mortality and Ambient Fine Particles in Southwest Mexico City, 1993–1995," *Environmental Health Perspectives*, 106, 849–855.

Buja, A., Hastie, T., Tibshirani, R., (1998), "Linear Smoothers and Additive Models," *The Annals of Statistics*, 17, 453–510.

Burnett, R.T., Smith-Doiron, M., Stieb, D., Cakmak, S., and Brook, J., (1999), "Effects of Particulate and Gaseous Air Pollution on Cardiorespiratory Hospitalizations," *Archives of Environmental Health*, 54, 130–139.

Conceicao, G.M.S., Miraglia, S.G.E.K, Kishi, H.S., Saldiva, P.N.H., and Singer, J.M. (2001), "Air Pollution and Child Mortality: A Time-Series Study in Sao Paolo, Brazil," *Environmental Health Perspectives*, 109, 347–350.

Dominici, F., McDermott, A., Zeger, S.L., and Samet, J.M. (2002), "On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health," *American Journal of Epidemiology*, 156, 193–203.

Dominici, F., McDermott, A., and Hastie, T. (2002), "Semiparametric Regression in Time Series Analyses of Air Pollution and Mortality: Generalized Additive and Generalized Linear Models," Presentation on Variance of GAM Estimators, Environmental Protection Agency Workshop on GAM-Related Statistical Issues in PM Epidemiology, November 4–6, 2002, Durham, NC.

——— (2003), "Improved Semi-Parametric Time Series Models of Air Pollution and Mortality" [on-line], http://www.biostat.jhsph.edu/~fdominic/jasa.R2.pdf.

Flanders, W.D., Klein, M., and Tolbert, P. (2002), "A New Variance Estimator for Parameters of Semi-parametric Generalized Additive Models. A Report to the U.S. Environmental Protection Agency," Based on a Presentation at the Environmental Protection Agency Workshop on GAM-Related Statistical Issues in PM Epidemiology, November 4–6, 2002, Durham, NC.

Flanders, W.D., Klein, M., and Tolbert, P. (2003), "A New Variance Estimator for Parameters of Semi-parametric Generalized Additive Models," Technical Report, Rollins School of Public Health, Emory University, Department of Biostatistics, Atlanta, GA.

Hastie, T.J., and Tibshirani, R.J. (1990), *Generalized Additive Models*, Monographs on Statistics and Applied Probability 43, New York: Chapman & Hall.

Health Effects Institute (2003), "Revised Analyses of Time Series Studies of Air Pollution and Health," Special Report, Boston, MA: Health Effects InstituteBoston, MA.

Katsouyanni, K., Touloumi, G., Samoli, E., Gryparis, A., Monopolis, Y., LeTertre, A., Boumghar, A., Rossi, G., Zmirou, D., Ballester, F., Anderson, H.R., Wojtyniak, B., Paldy, A., Braunstein, R., Pekkanen, J., Schindler,

C., and Schwartz, J. (2002), "Different Convergence Parameters Applied to the S-Plus GAM Function," *Epidemiology*, 13, 742–743.

Klein, M., Flanders, W. D., and Tolbert, P. E. (2002), "Variances may be Underestimated Using Available Software for Generalized Additive Models," *American Journal of Epidemiology*, 155, s106.

Lumley, T., and Sheppard, L., (2003), "Time Series Analyses of Air Pollution and Health: Straining at Gnats and Swallowing Camels," *Epidemiology*, 14, 13–14.

McCullagh, P., and Nelder, J.A. (1989), *Generalized Additive Models*, New York: Chapman and Hall, pp. 327–329.

Michelozzi, P., Forastiere, F., Fusco, D., Perucci, C.A., Ostro, B., Ancona, C., and Palotti, G. (1998), "Air Pollution and Daily Mortality in Rome, Italy," *Occupational and Environmental Medicine*, 44, 605–610.

Moolgavkar, S. (2000), "Air Pollution and Hospital Admissions for Diseases of the Circulatory System in Three U.S. Metropolitan Areas," *Journal of Air Waste Management Association*, 50, 1199–1206.

Pope, C.A., Hill, R.W., and Villegas, G.M. (1999), "Particulate Air Pollution and Daily Mortality on Utah's Wasatch Front," *Environmental Health Perspectives*, 107, 567–573.

Ramsay, T., Burnett, R., and Krewski, D. (2003), "The Effect of Concurvity in Generalized Additive Models Linking Mortality to Ambient Air Pollution," *Epidemiology*, 14, 18–23.

Samet, J.M., Dominici, F., Curriero, F., Coursac, I., and Zeger, S.L. (2000), "Fine Particulate Air Pollution and Mortality in 20 U.S. Cities: 1987–1994," *New England Journal of Medicine*, 343, 1742–1757.

SAS Institute (2001), *The SAS system for Windows*, Release 8.02, TS Level 02M0, Cary, NC: SAS Institute.

Schwartz, J. (1994a), "The Use of Generalized Additive Models in Epidemiology," XVIIth International Biometric Conference, Hamilton, Ontario, Canada, August 8-12, 1994. Proceedings, Volume 1: Invited papers.

——— (1994b), "Air Pollution and Hospital Admissions for the Elderly in Birmingham, Alabama," *American Journal of Epidemiology*, 139, 589–598.

Tolbert, P.E., Klein, M., Metzger, K.B., Peel, J., Flanders, W.D., Todd, K., Mulholland, J.A., Ryan, P.B., and Frumkin, H. (2000), "Interim Results of the Study of Particulates and Health in Atlanta (SOPHIA)," *Journal of Exposure Analysis and Environmental Epidemiology*, 20, 446–460.