GENERALIZED ADDITIVE MODELS FOR DATA WITH CONCURVITY: STATISTICAL ISSUES AND A NOVEL MODEL FITTING APPROACH

by

Shui He

B.S., Fudan University, 1993

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2004

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Shui He

It was defended on

October 29, 2004

and approved by

Sati Mazumdar, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Vincent C Arena, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Howard E. Rockette, Ph.D., Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Gong Tang, Ph.D., Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Nancy Sussman, Ph.D., Assistant Professor, Department of Environmental and Occupational Health, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Sati Mazumdar, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

GENERALIZED ADDITIVE MODELS FOR DATA WITH CONCURVITY: STATISTICAL ISSUES AND A NOVEL MODEL FITTING APPROACH

Shui He, PhD

University of Pittsburgh, 2004

The Generalized Additive model (GAM) has been used as a standard tool for epidemiologic analysis exploring the effect of air pollution on population health during the last decade as it allows nonparametric relationships between the independent predictors and response. One major concern to the use of the GAM is the presence of concurvity in the data. The standard statistical software, such as S-plus, can seriously overestimate the GAM model parameters and underestimate their variances in the presence of concurvity. We explore an alternate class of models, generalized linear models with natural cubic splines (GLM+NS), that may not be affected as much by concurvity. We make systematic comparisons between GLM+NS and GAMs with smoothing splines (GAM+S) in the presence of varying degrees of concurvity using simulated data. Our results suggest that GLM+NS perform better than GAM+S when medium-to-high concurvity exists in the data. Since GLM+NS result in loss in flexibility, we also investigate an alternative approach to fit a GAM. This approach, which is based on partial residuals, gives regression coefficients and variance estimates with less bias in the presence of concurvity, compared to the estimates obtained by the standard approach. It can accommodate asymmetric smoothers and is more robust with respect to the choice of smoothing parameters. Illustrative examples are provided. The public health significance of this study is that the proposed approach improves the estimate of adverse health effect of air pollution, which is important for public and governmental agencies to revise health-based regulatory standards for ambient air pollution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 OVERVIEW

1.1 BACKGROUND OF AIR POLLUTION STUDIES

The possibility of a significant relationship between short-term or long-term effects of air pollution and mortality is of concern to public and governmental agencies responsible for setting health-based regulatory standards for ambient air pollution. In the history, there were a series of air pollution "disasters" in the US and Europe, three most dramatic episodes of which happened in the Meuse Valley in Belgium in 1930, Donora, Pennsylvania in 1948, and London in 1952. During these episodes, there were evidence of increases in mortality and morbidity, coinciding with the increases in air pollution. Those disasters prompted investigation of the relationship between the air pollution and health. Numerous analyses have been performed to determine whether there is an increased relationship between air pollution level and mortality/morbidity. Many of the results from these studies have been considered by the United States Environmental Protection Agency (USEPA) when developing the air pollution regulations in the United Stated, National Ambient Air Quality Standard (NAAQS). Base on the design of studies, the air pollution studies can be divided into two groups: time series studies and prospective cohort studies. Cohorte studies, such as the Harvard six cities study [Dockery et al, 1993], the American Cancer Society study [Pope et al, 1995] and the Adventist Health Study [Abbey et al, 1995], followed a fixed group of people over a period of time, modeling health indicators as a function of air pollution measures after adjusting for some lifestyle confounding factors such as smoking and usual diet. A time series study is more like a population study and does not take the personal variation (life style factor) into account. It associates the daily indicators of health such as daily mortality and/or morbidity with the daily level of air pollutant after controlling for the confounding factor such as

long-term trend, seasonality and daily weather variations. Because of the availability of the data, there have been many more studies using the time series data.

1.2 STATE OF THE ART METHOD

In time series studies, the importance of removing the effects of long-term trends and seasonality, meteorological variability, and serial autocorrelation has been well recognized [Schwartz, 1994; Schwartz, 1999]. In the past decade, many epidemiological studies have shown an association between measurement of the ambient concentration of particulate matters less than 10 μm in aerodynamic diameter (PM_{10}) and non-accidental daily mortality [Schwartz and Marcus 1990; Pope et al., 1995; Schwartz, 1995]. Some of these studies have applied generalized additive models (GAMs) [Hastie and Tibshirani, 1990] in time series data of air pollution and mortality/morbidity because of the flexibility of these models. A GAM can be thought of as an extension of the family of generalized linear models. In a generalized linear model, a link function links the random component and the systematic component. Let Y to be the response variable and having exponential family density

$$\rho_Y(Y;\theta;\phi) = exp(\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y;\phi))$$

where θ is the natural parameter, and ϕ is the dispersion parameter. This is the random component. It is assumed that expectation of Y, denoted by μ, is related to the set of covariates $T_1, ..., T_p$ by

$$g(\mu) = \eta = \alpha + T_1\beta_1 + \cdots + T_p\beta_p.$$

The parameter η is the systematic component, called the linear predictor, and $g(\cdot)$is the link function. When an additive predictor (such as smoothing function) replaces the linear predictor in a generalized linear model, the model is called a generalized additive model. The mean $\mu = E(Y|T_1, ..., T_p)$ is linked to the predictor via the function

$$g(\mu) = \alpha + f_1(T_1) + \cdots + f_p(T_p).$$

In a generalized additive model, f_is are nonparametric smoothing functions that can be smoothing splines, kernels or local regression smoothing function (loess). These smoothing functions can be viewed as an extension of moving averages. Conceptually, for an individual smoothing function, the estimate of Y_i at point T_i is obtained by local average, i.e., averaging all of the Y_i, in a neighborhood about T_i. How to average the response values in each neighborhood and how to size the neighborhood become two important decisions and issues in the modeling endeavor. The way to average the response values in each neighborhood depends on which type of smoother one chooses. The closeness of neighborhood, reflected in terms of an adjustable smoothing parameter, e.g., span in loess, degrees of freedom in smoothing spline, determines the degree of smoothing [Hastie and Tibshirani, 1990]. For multiple smoothing functions, f_1, ..., f_p , these estimates can only be obtained by solving an iterative procedure. GAM uses the backfitting algorithm to sequentially smooth one smoother at a time, iterating until convergence occurs. The details of this fitting procedure can be seen in book: Generalized Additive Model [Hastie and Tibshirani, 1990]. GAMs are very flexible and there is no need for any assumption about the form for the dependence of Y on T. In air pollution study, the seasonality, long-term trend and weather variation are difficult to parameterize, as such GAMs became a standard analytic tool in time-series studies of air pollution and health [Schwartz, 1994; Dominici, et al, 2002a]. In order to quantify the health effect of air pollution, an air pollution measure is usually assumed to be linearly related to the indicators of health such as daily mortality and morbidity, but the relation to the calendar time and weather variables is not assumed to be parameterized. The model takes the form

$$E(Y|X,T) = g(\mu) = \eta = \alpha + X\beta + \sum f_i(T_i), \tag{1.1}$$

where X is a vector, representing air pollution measure and dummy variables for days of week. T_i represents calendar time and weather variables, etc. In model 1.1, we are most interested in estimating β and the smooth function, $f_i(T_i)$s, that can be regarded as nuisance parameters. There have been several approaches for estimating β. Backfitting [Hastie, et al., 1990] is a standard method and the GAM function of S-plus is widely used to perform the model fitting.

1.3 STATISTICAL ISSUES AND POSSIBLE SOLUTIONS

Recently, concerns were raised in the air pollution research community related to the use of GAMs in the assessment of pollution-health outcome associations by time series methodologies [Samet, et al 2003; Lumley, 2003]. GAMs are usually fitted by the S-plus software package. It has been demonstrated that the default convergence criteria of the gam function in S-plus do not assure convergence of the iterative estimation procedure and may result in biased estimates of regression coefficients and standard errors [Dominici et al., 2002b]. Studies conducted before May, 2002 using the gam function in S-plus with default convergence criteria may have overestimated the health effect of air pollution and underestimated its variance. For example, in National Morbidity, mortality, and Air Pollution Study (NMMAPS), the estimate of the average particulate pollution effect across the 90 largest U.S. cities changed from a 0.41% increase to a 0.27% increase in daily mortality per 10 $\mu g/m^3$ of PM_{10} when the more stringent convergence criteria were used [http://www.biostat.jhsph.edu/biostat/research/nmmapsfaq.htm]. NMMAPS has reported its reanalyzed results and concluded that there is strong evidence of an association between acute exposure to particulate air pollution (PM_{10}) and daily mortality occurring one day later, that the convergence issue has impact on the quantitative estimates, but the major conclusions do not change. The assurance of the convergence of the iterative procedure can be easily achieved by using more stringent convergence criteria when executing the software. However, the underestimation of the standard error and the presence of bias in the estimate of the regression coefficient due to concurvity, the nonparametric analogue of multicollinearity, still remained. It has been shown that this overestimation of parameters and underestimation of their variance might lead to significance tests with inflated type 1 error [Ramsay, et al., 2003a and 2003b], which may result in erroneously declaring a statistically significant effect when none exists. These researchers argue that some degree of concurvity between the transformed smooth functions and air pollutant levels is likely to be present in all epidemiological time series data, especially when time is used as an independent confounding variable. It has been suggested that other parametric approaches, such as generalized linear models with natural splines as smoothers (GLM+NS), may be used in place of nonpara-

metric GAMs [Ramsay et al., 2003a; Dominici, et al., 2002b; Samet, et al., 2003]. In the present dissertation, we make systematic comparisons between GLM+NS and GAM with smoothing splines as smoothers (GAM+S) in the presence of varying degrees of concurvity. The details of this comparison are given in chapter 2. Since GLM with natural splines is a parametric model, it has a few disadvantages. It has less flexibility in estimating smooth curves resulting in a slightly worse predictor error. Chapter 3 concentrates on the use of GAMs only allowing examination of the possibility of nonparametric association between factors. S-plus uses an ad hoc method to approximate the standard error of the parameter estimators, avoiding the expensive computation of the exact asymptotic standard error. It has been shown that this approximation underestimates the standard error [Ramsay, et al., 2003a]. Recently, the package called gam.exact, which is an extended S-plus function, has been developed to implement the expensive computation of the exact asymptotic standard error [Dominici, et al., 2004]. The use of this package allows a more robust assessment of uncertainty of air pollution effects. However, the bias due to concurvity, the nonparametric analogue of multicollinearity, still remains and needs to be corrected since the bias of the estimator of can asymptotically dominate the variance where T and X in model 1.1 are correlated [Rice, 1986] and some degree of correlation is likely to be present in majority of time series data in the air pollution studies. In chapter 3, we present an alternate way, using the partial residual regression approach, to fit the GAM. This method was first applied in the additive model with one smooth term, kernel [Speckman, 1988]. We extended it to the setting of time series analysis of air pollution and mortality data.

2.0 SYSTEMATIC COMPARISON BETWEEN GAMS AND GLMS

2.1 INTRODUCTION

As discussed in the previous chapter, if the concurvity is present in the data,which is usually the case in the air pollution data, the estimate of parameter is biased upwardly and the standard error of parameter is underestimated. This approximation underestimates the standard errors, leads to significance tests with inflated type 1 errors, and results in erroneously declaring a statistically significant effect when none exists [Ramsay et al., 2003]. Even when more stringent convergence criteria were used, it was shown that the parameter estimates are affected by concurvity, with larger bias when the size of true coefficient is small and concurvity is high, [Dominic et al., 2002b]. Given the fact that the effect of air pollution on health indicator is usually very small and concurvity is always present, the impact of concurvity on the parameter estimate could be very serious.

The statistical issues regarding the underestimation of the standard error and the presence of bias in the estimate of the regression coefficient in the presence of concurvity, are not totally resolved. Moreover, due to readily availability of the data, there are more time series studies than prospective cohort studies. Results from these time series studies are being considered by the United States Environmental Protection Agency (USEPA) for developing the air pollution regulations in the United States, National Ambient Air Quality Standard (NAAQS). Therefore, it is important to adequately address the statistical issues related to concurvity either by methodological adjustments or by alternative methods.

Samet and his colleague (2003) discussed the obligation of the air pollution researchers for repeating analyses that have used the gam function and for considering further methodological issues via alternate strategies. An alternate strategy for estimating the variance of

the estimates through simulations suggested that the variance estimates produced by the parametric bootstrap, although less biased than those produced by S-plus, remain biased downwards [Ramsay et al., 2003]. A recently developed S-plus package, gam.exact, allows a more robust assessment of uncertainty of air pollution effects [Dominici et al., 2003]. However this package only works for symmetric smoothers and therefore is limited in its usefulness. A parametric smoother such as natural cubic splines appears to be a better approach. Hence,GAM can be effectively turned into a fully parametric model, generalized linear model with parametric natural cubic splines (GLM+NS), and iteratively reweighted least square (IRLS) algorithm can be used for inference. The backfitting, which is known to be slow to converge in the presence of concurvity and to contribute those bias and variance problems, is not required in GLM fitting as all the smooth terms are estimated in one step [Schwartz et al., 2003].

The inferential problems related to the use of GAMs in the presence of concurvity are discussed in several recent papers (Ramsay et al., 2003; Figueiras et al., 2003; Lumley and Sheppard, 2003). The problems are primarily related to unstable and correlated estimates and standard errors not reflecting the instabilities of the parameter estimates, due to the way in which the variances are estimated. These issues prompted the researchers to the use of alternate models, such as GLMs, for air pollution health effect studies with time series data. Moreover, when the association is weak, correct estimation of parameters and the associated standard errors become essential. Lumley and Sheppard (2003) noted that these problems with GAMs may be due to the uncertainty in the shape of the smooth seasonal function that are included to control for temporal confounding. The approximation used to compute the standard errors ignores the effect of correlation between the fitted smooth function and the temperature and pollution effects.

In the light of this, the use of GLMs with natural splines comes into play. We can note here that in GLMs, collinearity is taken into account in the calculation of standard errors (the greater the collinearity, the greater the standard error). For a fixed number of knots, natural spline models are parametric. This avoids reliance on the iteration and tuning parameters needed in the GAMs. The flexibility of GAMs is an attractive feature but the associated inferential issues also should not be ignored. A systematic comparison of the performance

of GAMs with GLMs over a range of degrees of concurvity is important to suggest a change in the strategy in the modeling of air pollution data.

However, to date, the impact of concurvity on the parameter estimates has not been fully investigated. For example, a comparison of the performance of GLM+NS with GAM+S, using mild and stringent convergence criteria, was carried out for a single level of concurvity (correlation=0.6) [Dominici, et al., 2002b]. Although this was a reasonable beginning, a more thorough evaluation is warranted the statistical concerns discussed earlier. This is the motivation behind the present work. This chapter presents a systematic comparison of GLM +NS and GAM +S for a range of degrees of concurvity, including low degrees, using simulated datasets for several 8-year (2882-day) time series mimicking a real life air pollution study. Bias and variances estimates, from simulated datasets, were used to measure the performance of these two methods.

GLM+NS provides a straightforward parametric modeling approach that can be used effectively with count or categorical data with different link functions if necessary. Flexibility and the ability to accommodate nonlinear functions are the main advantage for GAMs. Though GLMs may not be as flexible as GAM, it can be used for the final modeling task after an exploratory analysis performed by GAM.

2.2 SIMULATION STUDY

Our simulated time series was based on a real data analysis of air pollution PM_{10} on health effects in Pittsburgh, PA over the period 1987-1994. Data were downloaded from the NMMAPS website (http://ihapss.biostat.jhsph.edu/data/data.htm). The variables used in the model consisted of total number of deaths among people older than 75 years, daily average temperature, daily average PM_{10}, calendar time of the series and day of week. Following earlier work [Dominici et al., 2002b], we fitted the following GAM to the data:

$$Y_t = \text{Poisson}(\mu_t)$$
$$log(\mu_t) = \alpha + \beta PM_{10} + S(time, 7/year) + S(temp, 6) + \eta I_{dow} \quad (2.1)$$

where Y_t denotes the daily number of death among people older than 75 years and follows a Poisson distribution with mean μ_t, β denotes the log relative rate of Y_t associated with a 1 $\mu g/m^3$ increase in M_{10}, $S(time, 7/year)$ is the smoothing splines function of calendar time with 7 degrees of freedom per year, $S(temp, 6)$ is the smoothing splines function of average temperature with 6 degrees of freedom, I_{dow} is the indicator variables for days of the week and α, β and η are the parameters to be estimated. Figure 2-1 shows empirical seasonality and long-term trend effect on mortality and Figure 2-2 shows empirical temperature effect on mortality for this data.

Next, an 8-year (2882-day) series was constructed using the model given in Equation (2.2):

$$\begin{aligned} Y_t &= \text{Poisson}(\mu_t) \\ log(\mu_t) &= log(\mu_0) + \beta PM_{10} + 0.2 * Trend + Newtemp. \end{aligned} \tag{2.2}$$

We assume that the number of deaths per day followed a Poisson distribution. The range of β for the simulations is selected based on the observed health effects of PM_{10} reported in the literature. The NMMAP Pittsburgh data analysis described earlier showed that the increase in mortality is approximately 0.00053 (using mild convergence criteria) and 0.00035 (using stringent criteria) for 1 $\mu g/m^3$ increase in PM_{10}. We generated the time series data using β parameters of 0.00035, 0.00055 and 0.00075.The higher value 0.00075 is added to broaden the range of β values as we note that the increase in mortality is approximately 0.0007 in some other studies [Schwartz, 1994b]. The average daily mortality count, μ_0, over the period 1987-1994 was found to be 21.

The degree of concurvity in the simulated data was introduced in the following manner. We use an additive model $PM_{10} = S(Time, 7/year) + S(Temp, 6) + error$ and obtain the fitted values of the response variable, $FITTEDPM_{10}$. We define a new variable, $NEWPM_{10}$, by $NEWPM_{10} = FITTEDPM_{10} + N(0, \sigma^2)$ [Dominici et al, 2002], which is the fitted values plus noise. We chose different σ^2 so that the correlations between $NEWPM_{10}$ and $FITTEDPM_{10}$ were equal to 0.01, 0.15, 0.29, 0.40, 0.50, 0.60, 0.69,0.80, 0.86 and 0.96. We note that as σ^2 increases, the concurvity between PM_{10} and the smooth function

$S(Time, 7/year)$ and $S(Temp, 6)$ decreases. If σ^2 is chosen to be zero and we have exact concurvity in the data.

The *Trend*, an unobserved confounding variable, consisting of a seasonal and a long-term trend component is simulated using the following formula [Bateson et al., 1999; Figueiras et al., 2003]:

$$Trend = (1 + Time/2882)[1 + 0.6cos(2ptime/365.25)].$$

We multiplied the derived *Trend* value by 0.2 to rescale the trend effect. We note that the purpose of rescaling is to bring the empirical and simulated trend effects closer. This simulated trend effect is displayed in Figure 2-3. In Figure 2-1, we see that the empirical effect of seasonality and trend is in the range of -0.2 - 0.35. The simulated effect of seasonality and trend after being rescaled by 0.2 lies in the same range of -0.2 - 0.35 (Figure 2-3). We should note here that without the rescaling, the range would have been 1-1.75. The simulated time trend was made to be associated with daily count of outcomes to induce time varying confounding effect into simulated time series [Bateson et al., 1999]. This simulated effect of seasonality and trend shares the important features with the empirical effect of seasonality and trend (Figures 2-3 and 2-1). Both have the same period and reach the peak and trough at the same points in time, with amplitudes being different in every cycle. The only difference is that the real data show more irregularity. The curves are adjusted to mean zero.

The variable *Newtemp* represents the functional form of the temperature effect on mortality in Pittsburgh, PA. This variable is generated by fitting the GLM model given in Equation (2.3), where $NS(time, 7/yr)$ and $NS(temp, 6)$ represent natural cubic splines with appropriate degrees of freedom and other quantities as defined earlier.

$$\begin{aligned} Y_t &= \text{Poisson}(\mu_t) \\ log(\mu_t) &= \alpha + \beta PM_{10} + NS(time, 7/year) + NS(temp, 6) + \eta I_{dow} \end{aligned} \tag{2.3}$$

$NS(temp, 6)$ generates a basis matrix A of dimension 2882*6 that represents the family of piecewise-cubic spline. The model also reports 6 coefficients ($beta.Temp$) for each column of matrix A. Thus, we can reproduce the effect of temperature, $Newtemp$, by the following formula $Newtemp = A \times beta.Temp$ This simulated effect of temperature ($Newtemp$) is

displayed in Figure 2-4. It shares the same pattern with the empirical effect shown in Figure 2-2. The effect first increases then goes down, then goes up as the temperature rises, and the two curves turn at the same points. The curves are adjusted to mean zero.

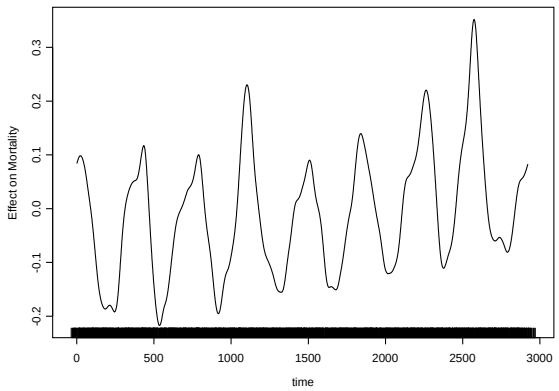

Figure 2.1: Empirical effect of seasonality and trend

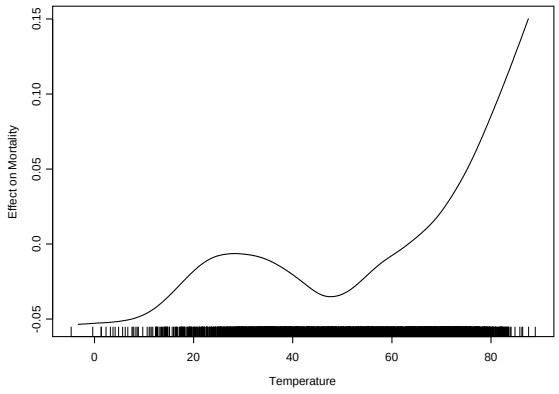

Figure 2.2: Empirical effect of temperature

For each level of concurvity, we generate 1000 mortality time series from Equation (2.2) by replacing PM_{10} with $NEWPM_{10}$. Each of the 1000 simulated datasets was fitted by the two models using S-plus software with the following specifications.

1) $GAM(Y_t - NEWPM_{10} + S(time, 7/year) + S(temp, 6), family = Poisson)$ where S is

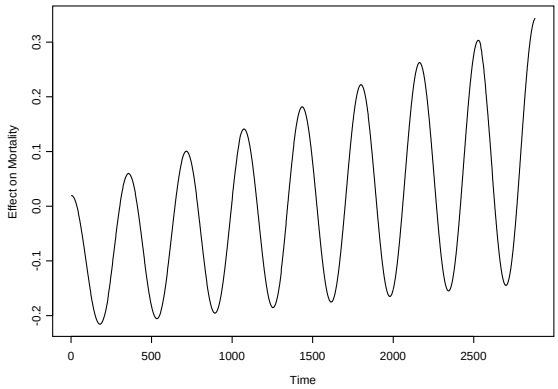

Figure 2.3: Simulated effect of seasonality and trend

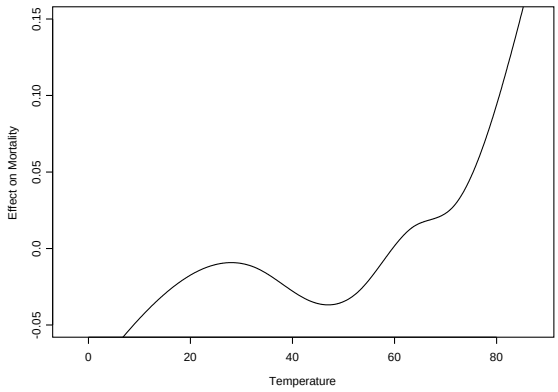

Figure 2.4: Simulated effect of temperature

the smoothing spline function. We used stringent convergence criteria with the parameters taking the values of epsilon:1e-15, bf.epsilon:1e-15, maxit :5000, bf.maxit:5000),

and

2) $GLM(Y_t - NEWPM_{10} + NS(time, 7/year) + NS(temp, 6), family = poisson)$ where NS is the natural cubic spline function using the default convergence criteria.

2.3 RESULTS

Table 2-1 summarizes the percent biases of the regression coefficient estimates β for PM_{10} in GAM+S and GLM+NS for different degree of concurvity in the data. Percent bias is a term well used in the air pollution community to reflect, on a percentage scale, how the biased estimate $\hat{\beta}$ increases or decreases relative to the true value of β [Dominici et al., 2002b] and is defined as $100(\hat{\beta}-\beta)/\beta$. The first column in the table is the correlation between $NEWPM_{10}$ and $FITTEDPM_{10}$, indicating the degree of concurvity. The second column represents the true regression coefficient β, used in the simulation. The third and fourth columns present the averages of β estimates and corresponding percent biases from 1000 simulated datasets using GAM+S. The fifth and sixth columns summarize the averages and percent biases of the β estimates using GLM+NS. Table 2-1 and the solid line in Figure 2-5 show that percent bias increases dramatically in GAM+S as concurvity increases and is maximum at 149%. In contrast, percent bias remains relatively constant and small (e.g. $\leq 3\%$) for GLM+NS over the range of concurvity. At lower concurvity (≤ 0.5), both GLM+NS and GAM+S provide comparable results.

Table 2-2 summarizes the percent discrepancy in variances for the regression coefficient β in GAM+S or GLM+NS in the presence of various degree of concurvity. We construct a measure to reflect the discrepancy between the two variances, the average of variances of β esimates from the simulated datasets ($VAR_{average}$) and the sample variance of β estimates from the simulated datasets (VAR_{sample}). This measure is defined as $100(VAR_{average} - VAR_{sample})/VAR_{sample}$. It represents a measure of bias in the variance estimate if we can say that with a large number of simulated datasets, VAR_{sample} represents

Table 2.1: Effect of concurvity on percent bias of the regression coefficient estimates

		GAM+S		**GLM+NS**	
Concurvity	β	$\hat{\beta}$	Percent Bias(%)	$\hat{\beta}$	Percent Bias(%)
0.96	0.00055	0.001370	149.1	0.000545	-0.9
0.86	0.00055	0.000803	46.0	0.000561	2.0
0.80	0.00055	0.000714	29.8	0.000554	0.7
0.69	0.00055	0.000619	12.6	0.000537	-2.4
0.60	0.00055	0.000600	9.1	0.000547	-0.6
0.50	0.00055	0.000581	5.6	0.000550	0.0
0.40	0.00055	0.000560	1.8	0.000542	-1.5
0.29	0.00055	0.000560	1.8	0.000551	0.2
0.15	0.00055	0.000551	0.1	0.000549	-0.2
0.01	0.00055	0.000550	0.0	0.000550	0.0

* Percent bias is defined as $100(\hat{\beta}$ - $\beta)/\ \beta$, where β is the true regression coefficient and $\hat{\beta}$ is the average of the thousand β estimates using GAM+S or GLM+NS.

Table 2.2: Effect of concurvity on percent discrepancy in variances

		GAM+S			**GLM+NS**	
Concurvity	Sample Variance	Average Variance	Percent discrepancy in Variances (%)	Sample Variance	Average Variance	Percent discrepancy in Variances (%)
0.96	7.01E-07	1.96E-07	-72.0	1.00E-06	9.98E-07	-0.2
0.86	2.20E-07	1.23E-07	-42.4	2.43E-07	2.62E-07	7.9
0.80	1.67E-07	9.99E-08	-40.3	1.79E-07	1.69E-07	-5.6
0.69	8.76E-08	6.40E-08	-26.9	9.06E-08	8.64E-08	-4.6
0.60	5.52E-08	4.49E-08	-18.6	5.66E-08	5.57E-08	-1.7
0.50	3.28E-08	2.92E-08	-10.7	3.31E-08	3.39E-08	2.2
0.40	1.96E-08	1.74E-08	-11.1	1.96E-08	1.90E-08	-2.8
0.29	1.00E-08	9.22E-09	-7.8	1.02E-08	9.60E-09	-5.9
0.15	2.70E-09	2.60E-09	-3.8	2.81E-09	2.67E-09	-4.9
0.01	6.24E-11	6.40E-11	2.6	6.40E-11	6.72E-11	5.1

* Percent discrepancy in variances is defined as $100(Var_{average}$ -Var_{sample} / Var_{sample}, where Var_{sample} is the sample variance of the one thousand β estimates. $Var_{average}$ is the average of the one thousand S-plus variances.

the "true" variance of the estimate. Similar measure was used by Ramsay (2003). Comparison of these two sets of variance estimates, $VAR_{average}$ and VAR_{sample}, can also be seen in statistical literature (Hu et al., 1998). In Table 2-2, the first column indicates the degree of concurvity. The second, third and fourth columns represent VAR_{sample}, $VAR_{average}$ and the percent discrepancy in variances using GAM+S. The fifth, sixth and seventh columns represent VAR_{sample}, $VAR_{average}$ and the percent discrepancy in variances using GLM+NS. Table 2-2 and the solid line Figure 2-6 show that for GAM+S the percent discrepancy in variances increase as concurvity increases; on the other hand, in GLM+NS, the percent discrepancy in variances remains constant and small even in the presence of high concurvity.

Table 2.3: Average Mean Square Errors of regression coefficient estimates

Concurvity	GAM+S	GLM+NS
0.96	1370	1000
0.86	284	243
0.80	194	179
0.69	92.2	90.4
0.60	57.5	56.5
0.50	33.5	33.2
0.40	19.5	19.7
0.29	10.2	10.2
0.15	2.73	2.75
0.01	0.06	0.06

* Average Mean Square Errors (10^{-9}) is obtained for simulation when β=0.00055

Table 2-3 presents the average of the 1000 mean square errors for regression coefficient estimates. The first column indicates the degree of concurvity. The second column represents the AMSE using GAM+S; the third column represents the AMSE using GLM+NS. From this table, we see that in general the estimates of GLM+NS have smaller AMSE than those of GAM+S. This suggests that GLM+NS have better performance than GAM+S, even under the circumstance that the variance by using GAM is not underestimated.

We also investigated whether the percent bias and the percent discrepancy in variances depend on the size of the true regression coefficient β by repeating the simulations using βs equal to 0.00035 and 0.00075. Figure 2-5 and 2-6, discussed earlier, summarize the results. In Figure 2-5, we see that the percent bias increases as the size of the true regression coefficient decreases. In Figure 2-6, we see that the percent discrepancy in variances did not depend on the "true" regression coefficient in both models

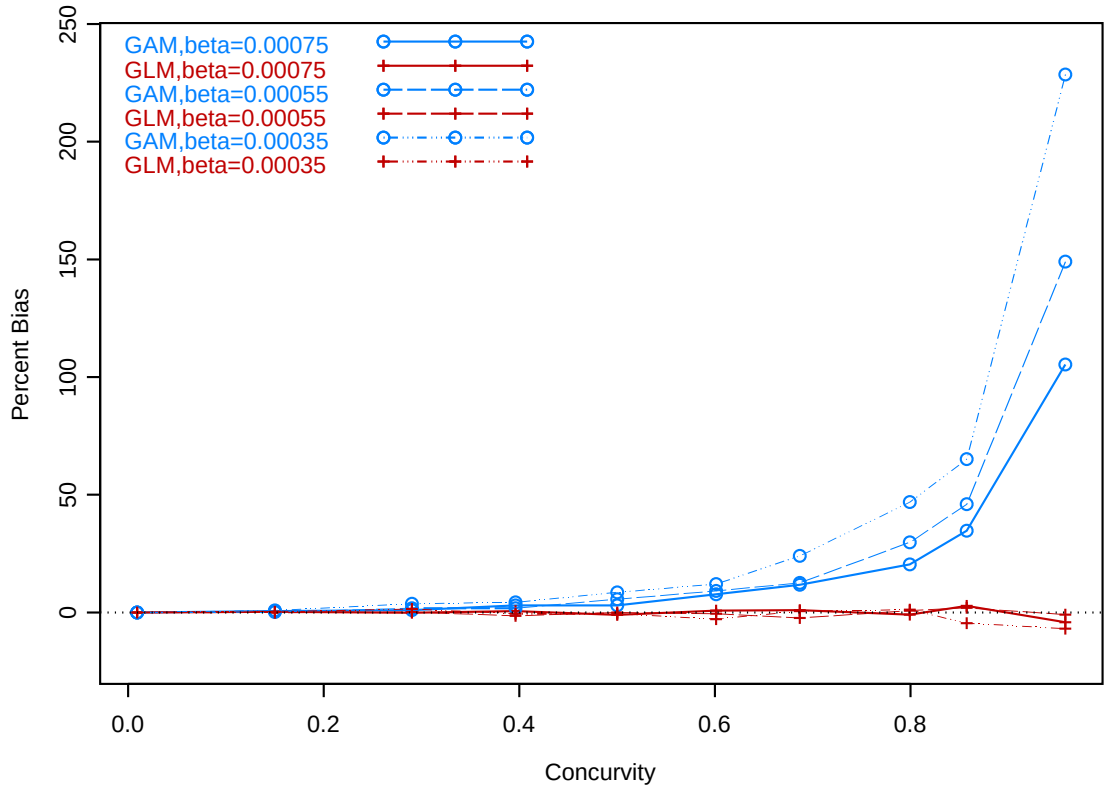

Figure 2.5: Effect of concurvity on percent bias

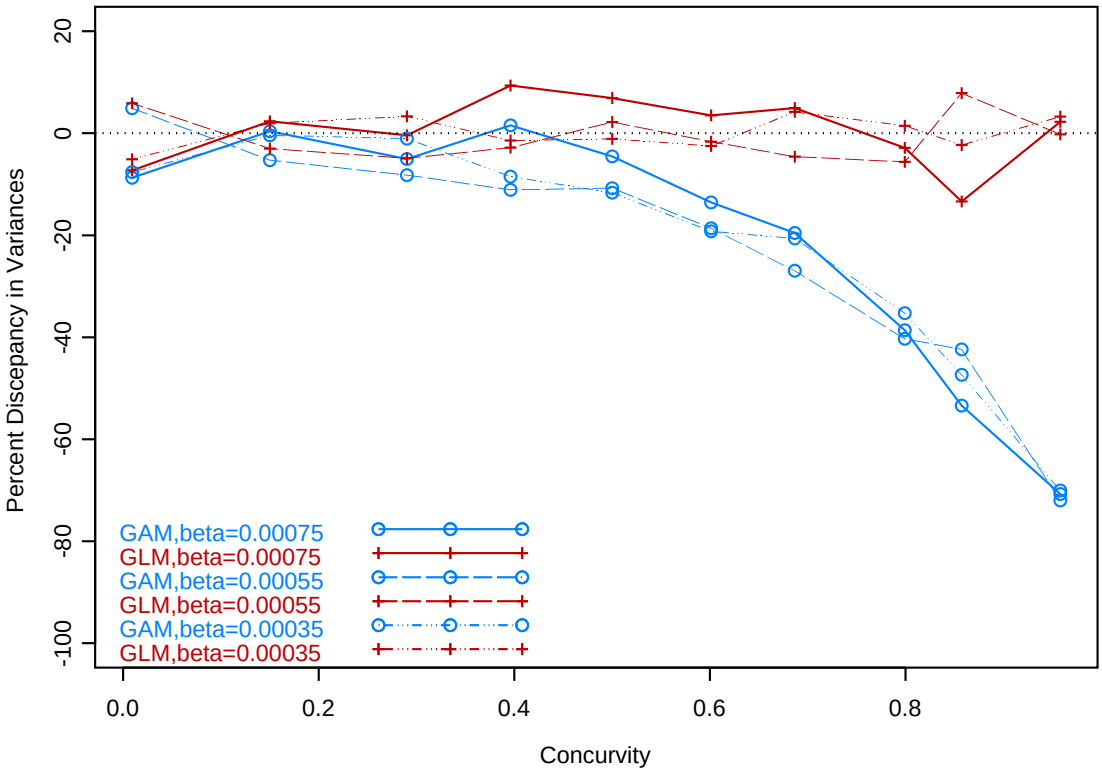

Figure 2.6: Effect of concurvity on variance estimate

2.4 DISCUSSION

In this chapter, we made systematic comparisons between GLM+NS and GAM+S in the context of time series analysis of air pollution and mortality using simulated data. The simulated mortality data were constructed from a linear model using pollution measures, a trend/seasonality component and a nonlinear temperature component. We simulated the temperature component with the aid of natural splines. In the Pittsburgh data analysis, that served as the framework for the simulation study, the empirical temperature effect on mortality using GAM+S was comparable with the simulated effect using GLM+NS (Figures 2-2 and 2-4). Hence, we believe that the results from our simulation study do not extremely favor towards the use of GLM+NS. The relationship between trend/seasonality and mortality is considered to be very important in assessing the mortality and PM_{10} relationship. Sometimes, it is believed that this relationship is more important than the PM_{10} and temperature relationship, as seasonality accounts for the temperature effect to a large extent. The nonlinear function necessary to generate the trend/seasonality component was similar for both the methods. Based on these observations, we are confident about the correctness of our simulation study.

Our results suggest GLM+NS perform better than GAM+S when medium-to-high concurvity is present in the data. At low concurvity, GLM+NS and GAM+S are comparable. Since some degree of concurvity is likely to be present in air pollution time series data, the use of GAM+S may give erroneous results. The results also show that there is more benefit from using GLM+NS than using GAM+S in the presence of smaller effects, which is typically the case in U.S. It can be argued that using GLM+NS will result in some loss in flexibility, as it needs a parametric model. We suggest a strategy with a two-stage approach. First, GAM would be used for the exploratory analysis because of the flexibility of GAM will make the exploratory analysis easy and less time consuming. Then the fully parametric models (GLM+NS) would be fitted for the final parameter estimates.

In a specific study, nonparametric models (GAM with smoothing splines or loess) would be used to explore the data, then, after the appropriate variables and smoothing parameters are identified, a fully parametric model (GLM+NS) with the same predictors and the same

degree of smoothness can be used. For example, the NMMAP Pittsburgh data had been analyzed using the GAM and 56 degrees of freedom were chosen for the smoothing spline of time and 6 degrees of freedom were chosen for temperature [Dominici et al., 2002b]. In our work, we re-fitted the GAM+S with the same smoothing parameters. This step can be thought as the first stage of the analysis. The β was found to be 0.00033 with a corresponding variance of 3.74E-08. Next, we used the same degrees of freedom to fit the GLM model. The β from GLM was found to be 0.00030 with a variance of 4.54E-08. This was a 10% difference from the GAM estimate of β with an increase in variance. The concurvity in the Pittsburgh data is approximately 0.6 and based on our simulation study we would expect a 9.1% bias in the GAM estimate of β and -0.5% in GLM β estimate. Furthermore, we would expect an underestimation of 18% in the variance from GAM and 1.7% from GLM. Thus, our results for the Pittsburgh data are consistent with our findings from the simulations. Figure 2-7 presents the two fitted models imposed on the observed data along with the plots of the residuals. The residual plots did not show any seasonality or long term trend and the values are scattered evenly around '0' value. Figure 2-8 shows similar patterns in residuals when plotted against temperature. Formal tests of goodness of fit were made with the null model and the tests were found to be highly significant.

In conclusion, we recommend more GLM modeling with natural splines in time series analysis of air pollution studies with substantial concurvity in the data.

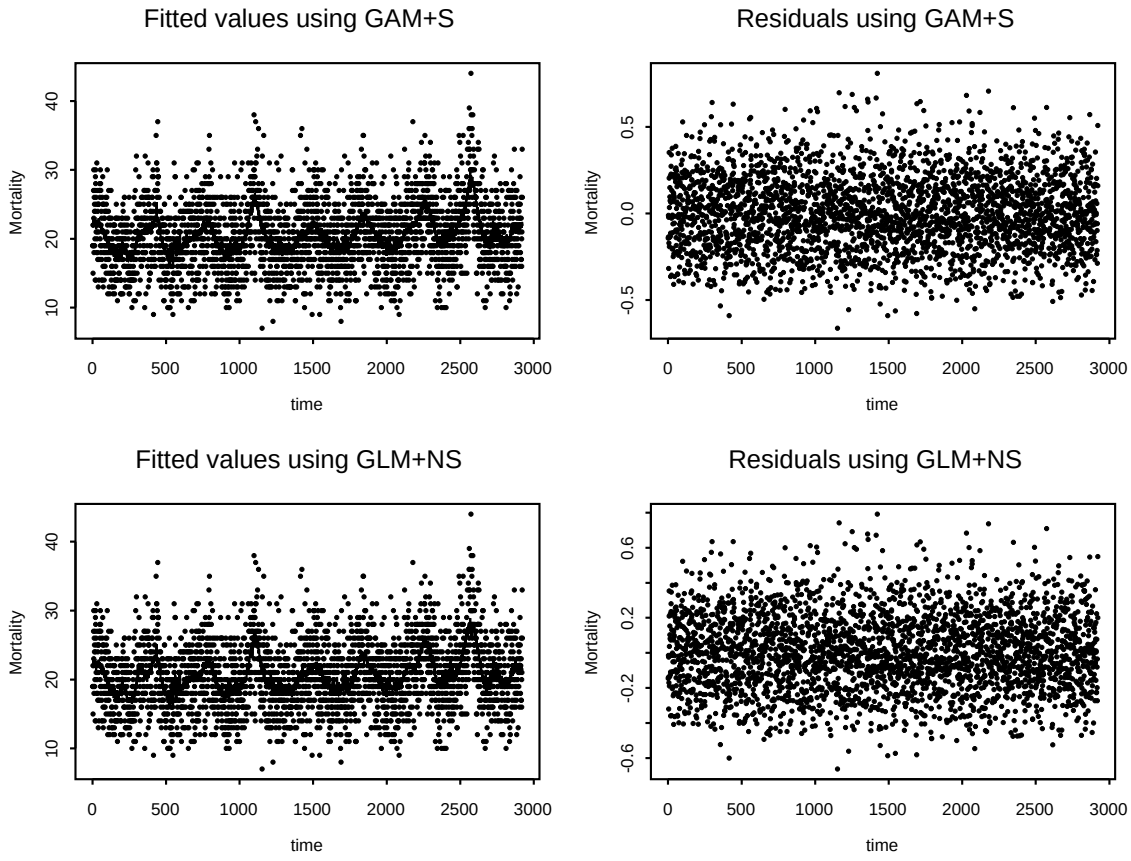

Figure 2.7: Fitted values and residuals for daily mortality against time

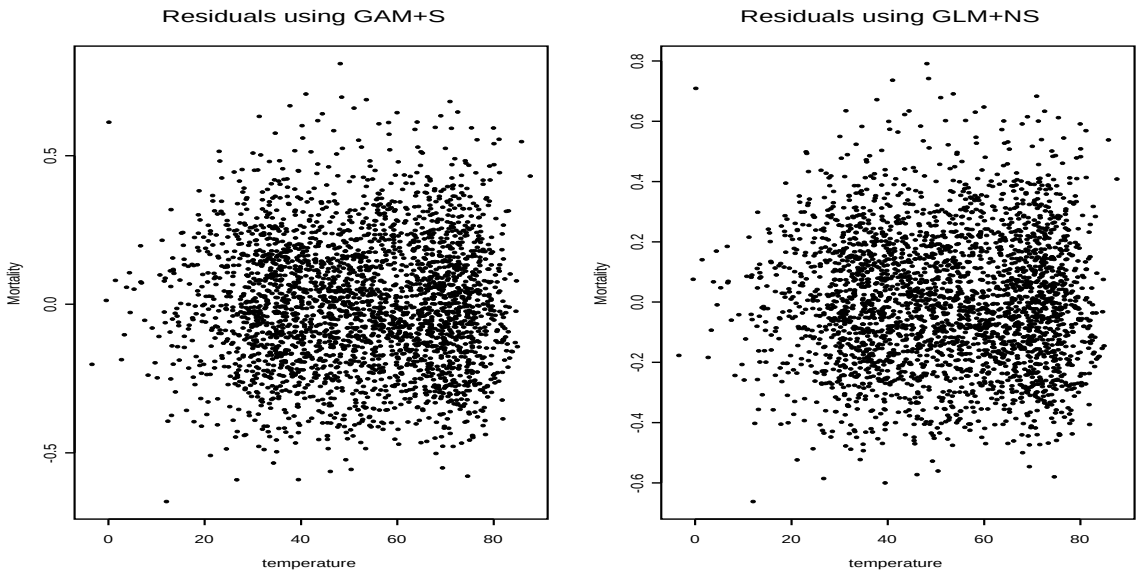

Figure 2.8: Residuals for daily mortality against temperature

3.0 PARTIAL REGRESSION APPROACH

3.1 INTRODUCTION

In chapter 2, we show that the GLM with NS perform better than GAM+S and because of the difficulty to select the number and locations of knots, especially with missing data and we recommended a two-stage strategy. But a two-stage strategy is time and labor intensive, and moreover the use of GLM in the second stage may result in some loss in flexibility. Hence, we are interested in fitting the GAM by an alternate approach.

We explored an alternate way, using the partial residual regression approach, to fit the GAM. This method was first applied in the additive model with one smooth term, kernel [Speckman, 1988]. We extended it to the setting of time series analysis of air pollution and mortality data. We developed the package *gam.partial.residual*, an extension of gam for its ready use.

3.2 METHODS

3.2.1 The Standard Backfitting Algorithm to Fit a GAM

Consider a very simple additive model

$$Y = f(T) + \epsilon, \tag{3.1}$$

where Y is the vector of observed values of $(Y_1, ...Y_n)$, the error term ϵ is assumed a vector of independent identically distributed random variables with mean zero and variance σ^2 and

$f(T)$ represents a smooth nonparametric function that can be a smoothing spline, kernel or local regression smoothing (loess). S is defined as a smoother matrix (Hastie and Tibshirani, 1990), which transforms $Y = (Y_1, ...Y_n)'$ to fitted value $\hat{Y}$, where $\hat{Y} = f(\hat{T}) = SY$.

For the semiparametric additive model

$$Y = X\beta + f(T) + \varepsilon \tag{3.2}$$

we can consider two smoothers S_1 and S_2. $S_1 = X(X'X)^{-1}X'$ produces a least square fit $X\hat{\beta}$ for the parametric part, and S_2, another smoother for the nonparametric part, produces an estimate $\widehat{f}(T)$. The backfitting steps (Hastie and Tibshirani, 1990) can be given as follows:

$$f_0 = S_1(Y - f_1) = X(X'X)^{-1}X'(Y - f_1) = X\beta,$$
$$f_1 = S_2(Y - X\beta).$$

$\hat{\beta}$ and $\hat{f}_1$ can be solved by following the above iterative procedure and an explicit expression for the value of $\hat{\beta}$ can also be derived.

$$\hat{\beta} = \{X'(I - S_2)X\}^{-1}X'(I - S_2)Y = HY,$$
$$\hat{f}_1 = S_2(Y - X\hat{\beta}) \tag{3.3}$$

The covariance matrix of is estimated by,

$$\widehat{Var}(\hat{\beta}) = H'Var(Y)H \tag{3.4}$$

where $H = \{X'(I - S_2)X\}^{-1}X'(I - S_2)$

If a model has two and more nonparametric terms such as

$$Y = X\beta + \sum f_i(T_i) + \varepsilon, \tag{3.5}$$

the same expressions for $\hat{\beta}$ and covariance matrix of $\hat{\beta}$ as in equations (3.3) and (3.4) could be obtained. In the expression, S_2 produces an estimate $\sum \widehat{f_i(T_i)}$. We have to put all the additive smooth terms $f_i(T_i)$ together, and S_2 represents the operator for computing the additive fit of the nonparametric part. As such, S_2 represents the smoothing matrix in the

last iteration of backfitting procedure on these terms (Durban et al., 1999). We can note here that with a nonidentity link function, model (3.5) becomes a generalized additive model

$$E(Y|X,T) = g(\mu) = \eta = X\beta + \sum f_i(T_i) \tag{3.6}$$

This model can be fitted by using local scoring procedure (Hastie and Tibshirani, 1990), which iteratively fits weighted additive models by backfitting. This iterative procedure is quite similar to iteratively reweighted least square algorithm in generalized linear model (McCullagh and Nelder, 1989). A GAM differs from a generalized linear model (GLM) in that an additive predictor replaces the linear predictor. An explicit expression for $\hat{\beta}$ can be derived as

$$\hat{\beta} = \{X'W(I - S_2)X\}^{-1}X'W(I - S_2)Z,$$

where l is the likelihood function, Z is the working response from the final iteration of the iteratively reweighted least square (IRLS) algorithm $Z = \eta + (Y - \mu)\frac{d\eta}{d\mu}$, W is diagonal in the final IRLS weights, $W = -\frac{d^2l}{d\eta\eta'}$ and S_2, as previous defined, is the smoothing matrix producing the estimate $\sum \widehat{f_i(T_i)}$. The variance estimate for $\hat{\beta}$ is

$$\widehat{Var}(\hat{\beta}) = \{X'W(I - S_2)X\}^{-1}X'W(I - S_2) = H'W^{-1}H.$$

In S-plus package, when the backfitting algorithm is used to fit the standard GAMs, the standard errors are approximated by an ad hoc method. The calculation of the standard errors depends on S_2 ($n \times n$ matrix) described in the earlier section. The calculation of S_2 is computationally expensive when n is large. Moreover, in the current version of gam function in S-plus, the ad hoc method approximates the standard errors by assuming that the smoother, $f_i(T_i)$, is linear. In air pollution studies, as the time effect has a cyclic effect, this assumption of linearity is often inadequate. A shortcut has been developed to calculate the exact asymptotic standard error without the expensive computations (Durban et al.; 1999). Also, a package, *gam.exact*, is available (Dominici et al., 2003). A few issues

regarding gam.exact can be noted here. If S_2 is a symmetric smooth matrix and as W is always symmetric, we have

$$WS_2 = S_2'W,$$
$$H = \{X'W(I - S_2)X\}^{-1}X'W(I - S_2)$$
$$= \{X'(WX - WS_2X)\}^{-1}(WX - WS_2X)'.$$

The knowledge of S_2X is sufficient to calculate $\widehat{Var}(\hat{\beta})$. To calculate S_2X, we can fit a model and extract W from S-plus output. We note that S_2X is the corresponding fitted values of the assumed model. Only p additive models (where p is the rank of X) instead of n need to be fitted to calculate the required standard errors.

The *gam.exact* works well for symmetric smoothers, such as smoothing splines, but does not work well for smoother that are not symmetric such as "loess" (Dominici et al., 2003).

3.2.2 Partial Regression Approach to Fit a GAM

This approach was described by Speckman (1988) with kernel smoothing. For the semiparametric model given in equation(3.3),

$$Y = X\beta + f(T) + \varepsilon.$$

The partial residuals could be defined as the variable Y and X after 'adjustment' for the dependence on T. The partial residuals of Y could be defined as the residuals of the partial additive model, $E(Y) = f(T)$, and take the form of $\tilde{Y} = (I - S_2) = Y - Y_{partial\_fitted}$. Similarly, the partial residuals of X could be defined as the residuals of the partial additive model $E(X) = f(T)$, and take the form of $\tilde{X} = (I - S_2) = X - X_{partial\_fitted}$. A simple linear regression of $\tilde{Y}$ on $\tilde{X}$ provides new estimates:

$$\hat{\beta}_r = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$$
$$= \{X'(I - S_2)^2X\}^{-1}X'(I - S_2)^2Y, \tag{3.7}$$

$\hat{\beta}_r$ is a new estimator. Compared to $\hat{\beta}$ from equation (3.3), the bias of $\hat{\beta}_r$ is of lower order and asymptotically (at least) negligible, which suggests that $\hat{\beta}_r$ is more appropriate if main interest is inference on β (Speckman, 1988).

Even though Speckman only focused on the kernel smoothing and a simple additive model, this method can be extended to model (3.5). In this model, g can be non-identity link, $f_i(T_i)$s are any smoothing function(s) for two and more nonparametric terms. The new estimate of $\hat{\beta}_r$ can be obtained and takes the form:

$$\hat{\beta}_r = \{X'W(I - S_2)^2X\}^{-1}X'W(I - S_2)^2Z,$$

where l is the likelihood function, Z is the working response from the final iteration of the iteratively reweighted least square (IRLS) algorithm $Z = \eta + (Y - \mu)\frac{d\eta}{d\mu}$, W is diagonal in the final IRLS weights, $W = -\frac{d^2l}{d\eta\eta'}$ and S_2 is the smoothing matrix producing the estimate $\sum \widehat{f_i(T_i)}$. The details were described later. From the close form solution for $\hat{\beta}_r$, the variance estimate can be calculated as

$$\widehat{Var}(\hat{\beta}_r) = H_1'W^{-1}H_1 = \{X'W(I - S_2)^2X\}^{-1}X'W(I - S_2)^2.$$

When the smooth matrix S_2 is not symmetric, the partial regression approach does not pose any problem.

We should also note that the fitting of the model (3.5),

$$E(Y|X, T) = g(\mu) = \eta = X\beta + \sum f_i(T_i)$$

is actually implemented by fitting the model

$$E(Z|X, T) = X\beta + \sum f_i(T_i)$$

with weight W, where $Z = \eta + (Y - \mu)\frac{d\eta}{d\mu}$ and $W = -\frac{d^2l}{d\eta\eta'}$. As Z is the new dependent variable, we can obtain partial residuals of Z by fitting the partial additive model, $E(Z|T) = \sum f_i(T_i)$ with weight$W$, and obtained the partial residuals Z_{res}. Similarly, we can obtain partial residuals of X by fitting the partial additive model, $E(X|T) = \sum f_i(T_i)$ with weight W, and obtained the partial residuals X_{res}. Then, a linear regression of Z_{res} on X_{res} with the weight W provides $\hat{\beta}_r$.

The above model fitting and calculation are implemented in an S-plus function *gam.partial.residual*. It is applicable to the entire class of link functions for GAM.

3.3 SIMULATION STUDY

Our simulated time series was based on a real data analysis of air pollution PM_{10} on health effects in Pittsburgh, PA over the period 1987-1994. We refer to Section 2.3 for the details.

For each level of concurvity, we generate 500 mortality time series from the equation in previous section by replacing PM_{10} by $NEWPM_{10}$. Each of the 500 simulated datasets was fitted by the two models, standard approach (gam.exact) and partial regression approach (gam.partial.residual), using S-plus software with the following specifications.

$$GAM(Y_t - S(time) + S(Temp, df = 6) + NewPM_{10}, family = Poisson)$$

where S is the smoothing spline function. We assign different degrees of freedom for time, ranging from 4 df/yrs to 8 df/yrs.

Table 3-1 summarizes the comparison of these two approaches under different concurvity and degrees of freedom when the true β=0.00055. The first column indicates the degree of concurvity. The second column indicates the degrees of freedom for time. The third column gives the approaches used to obtain the β estimates. The fourth column is $\hat{\beta}$, the average of 500 β estimates ; the fifth and the sixth columns represent the empirical standard error of the five hundreds β estimates and the average of the five hundreds standard errors. We can see that, as the concurvity increases, the bias increases both in the standard approach (using *gam.exact*) and the partial regression approach (using *gam.partil.residual*). However the magnitudes of the bias in the partial regression approach are always smaller than the magnitudes in the standard approach. Results from df=7/year, df=5/year and β=0.00055 are reported here, similar results are found with other degrees of freedom and other values of β. We note that the average standard errors are very close to the empirical standard errors in both approaches and that empirical standard errors in the partial regression approach are slightly larger than those in the standard approach. However, the inflation in standard error is negligible, compared to the bias reduction.

Table 3-2 gives more details about the comparison between the standard approach and the partial regression approach for different degree of concurvity in the data. Similar results are found with other degrees of freedom and other values of β. We use percent bias to reflect, on a

Table 3.1: Comparison of the approaches under different concurvity and degrees of freedom

Concurvity	Df/Yr	Approaches	$\hat{\beta}$	Sample SE	Average SE
0.8	7	standard	0.000685	0.000381	0.000385
		partial	0.000546	0.000396	0.000396
	5	standard	0.000981	0.000364	0.000370
		partial	0.000619	0.000385	0.000391
0.6	7	standard	0.000600	0.000231	0.000226
		partial	0.000555	0.000234	0.000227
	5	standard	0.000705	0.000228	0.000222
		partial	0.000578	0.000232	0.000226
0.4	7	standard	0.000572	0.000137	0.000133
		partial	0.000556	0.000138	0.000133
	5	standard	0.000609	0.000137	0.000132
		partial	0.000565	0.000137	0.000132
0.0	7	standard	0.000550	0.000008	0.000008
		partial	0.000550	0.000008	0.000008
	5	standard	0.000550	0.000008	0.000008
		partial	0.000550	0.000008	0.000008

*β=0.00055

*Standard: standard approach, using *gam.exact*

*Partial: partial regression approach, using *gam.partil.residual*

Table 3.2: Point estimates under different degree of concurvity

	GAM+S (Standard)		**GAM+S (Partial)**	
Concurvity	$\hat{\beta}$	Percent Bias*(%)	$\hat{\beta}$	Percent Bias (%)
0.96	0.001303	136.91	0.000595	8.1
0.85	0.000790	43.69	0.000579	5.31
0.80	0.000685	24.51	0.000546	-0.65
0.68	0.000632	14.98	0.000561	2.07
0.60	0.000600	9.13	0.000555	0.82
0.50	0.000581	5.55	0.000553	0.53
0.40	0.000572	3.91	0.000556	1.15
0.30	0.000560	1.75	0.000552	0.35
0.21	0.000554	0.73	0.000551	0.09
0.16	0.000550	-0.04	0.000548	-0.42
0.03	0.000550	-0.04	0.000550	-0.04

*β=0.00055

*Degrees of freedom: 7/year for time, 6 for temperature

*Percent bias: defined as 100($\hat{\beta}$ - β)/ β, where β is the true regression coefficient and $\hat{\beta}$ is the average of the five hundred regression coefficient estimates.

percentage scale, how the bias increases or decreases relative to the true value of β (Dominici et al., 2002). The percent bias is defined as 100($hat\beta$ - β)/ β. The first column indicates the degree of concurvity. The second and third columns present the averages of regression coefficient estimates and corresponding percent biases from 500 simulated datasets using the standard approach. The fourth and fifth columns summarize the averages and percent biases of the regression coefficient estimates using the partial regression approach. Table 3-2 shows that percent bias increases dramatically in the standard approach as concurvity increases and is maximum at 137%. In contrast, percent bias remains relatively constant and small for partial regression approach over the range of concurvity.

Table 3.3: Standard error estimates under different degree of concurvity

	Standard Approach			**Partial Regression Approach**		
Concurvity	Sample SE	Average SE	Percent* discrepancy	Sample SE	Average SE	Percent discrepancy
0.96	0.000800	0.000812	1.58	0.000947	0.000963	1.70
0.85	0.000471	0.000472	0.13	0.000496	0.000493	-0.46
0.80	0.000381	0.000385	1.08	0.000396	0.000396	-0.03
0.68	0.000292	0.000280	-3.94	0.000297	0.000284	-4.67
0.60	0.000231	0.000226	-2.17	0.000234	0.000227	-3.03
0.50	0.000177	0.000177	0	0.000178	0.000177	-0.67
0.40	0.000137	0.000133	-3.20	0.000138	0.000133	-3.84
0.30	0.000096	0.000095	-1.20	0.000097	0.000095	-2.27
0.21	0.000067	0.000067	-0.61	0.000068	0.000066	-1.72
0.16	0.000049	0.000050	2.42	0.000049	0.000050	1.55
0.03	0.000008	0.000008	-2.03	0.000008	0.000008	-2.82

*β=0.00055

*Degrees of freedom: 7/year for time, 6 for temperature

*Percent discrepancy in standard error, is defined as $100(SE_{average} - SE_{sample})/SE_{sample}$.

Table 3-3 summarizes the estimates of standard errors for the regression coefficient estimates obtained by the two approaches in the presence of various degree of concurvity. The first column indicates the degree of concurvity. The second and third columns represent the sample standard error of the five hundreds regression coefficient estimates and the average of the five hundreds standard errors obtained by the standard approach (using *gam.exact*). The fourth represents the percent discrepancy in standard errors to reflect the discrepancy between the two standard errors, the average standard error ($SE_{average}$) and the sample standard error (SE_{sample}), on the premise that the sample standard error reflects the true standard error of the regression coefficient. This measure is defined as $100(SE_{average}$ -$SE_{sample})$ / SE_{sample}. The fifth and sixth columns represent the sample standard error and the average of the standard errors obtained by the partial regression approach (using *gam.partial.residual*). The seventh column represents the percent discrepancy in standard errors using partial regression approach. From the table, we can see that the estimates of the standard error, reflected by average standard errors, are close to their corresponding "true" standard errors, reflected by sample standard errors, which suggests that no underestimation exists. We also note that the sample standard errors in the partial regression approach are a little larger than the sample standard errors in the standard approach. However, the inflation in standard errors is negligible, compared to the magnitude of the bias reduction.

Figure 3-1 presents the 95% confidence interval coverage for a range of regression coefficients, concurvity levels and degrees of freedom per year for time for the two approaches. The horizontal lines represent the 95% CI coverage. We can see that the partial regression approach has a better confidence interval coverage than the standard approach in most of the situation and is never worse. Figure 3-2 presents the 95% confidence interval coverage against temperature for the two approaches. The horizontal lines represent the 95% CI coverage. We can see that the 95% confidence interval coverages are stable over the range of degrees of freedom in both approaches.

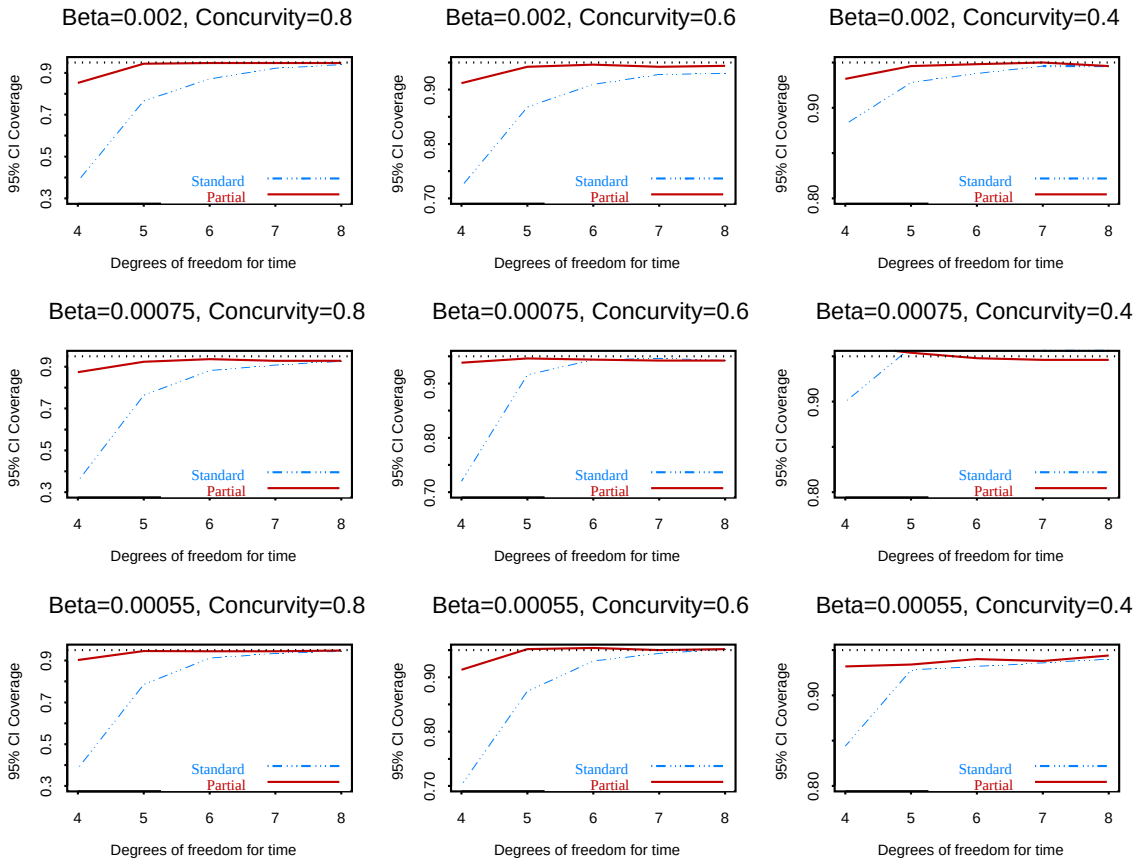

Figure 3.1: Confidence interval coverage against time

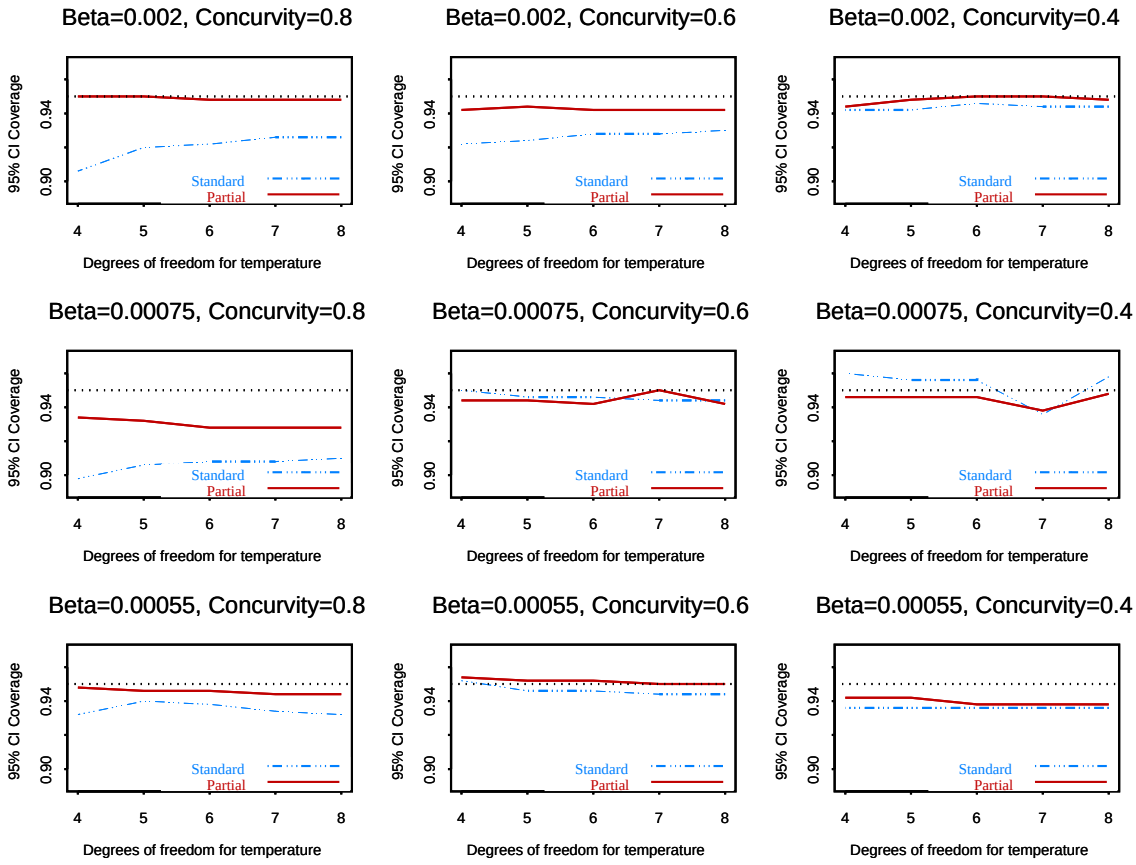

Figure 3.2: Confidence interval coverage against temperature

3.4 DISCUSSION

In GAM fitting by S-plus, the default statement underestimates the standard error of the regression coefficients because of the use of the ad hoc approximation. The gam.exact package provides better estimate of standard error, but the package works well only when the smoother matrices are symmetric. One of the widely used smoothers, loess, which is not symmetric, is unable to benefit from this package. And the package only improves the estimation of the standard errors but bias due to the concurvity remains. The partial regression approach in this chapter is aimed at this issue. This approach was first used by Speckman [Speckman, 1988] and we have extended it to the setting of air pollution studies with time series data. Our results show that this approach performs better than standard GAM fitting when concurvity is present and also can handle asymmetric smoothers.

Our sensitivity analyses with varying degrees of freedom for the time component of the model also demonstrate that this approach performs better than the standard approach.

4.0 ILLUSTRATIVE EXAMPLES

4.1 DATA

The illustrative examples used the NMMAPS database which is comprised of daily time series of air pollution levels, weather variables, and mortality counts for the largest 90 cities in the US from 1987 to 2000. A full description of the NMMAPS database is detailed by Samet (2000) and data can be downloaded from the web site http://www.ihapss.jhsph.edu. To analyze effectively the association, we chose the data with daily PM_{10} measurements and few missing values. Most U.S. cities only measured PM_{10} for 1 of 6 days, but a number of locations had daily measures available. Four cities with roughly daily PM_{10} were selected These cities were Pittsburgh, Chicago, Detroit, and Minneapolis. Table 4-1 describes the time periods for which data were available. Among all the four cites, Pittsburgh had PM_{10} data only from 1/1/1987 to 12/31/1998 and the analysis was based on the 12-year series data covering this period. Detroit had PM_{10} data from 1/1/1987 to 4/30/2000. It is more convenient to assign the degrees of freedom for the smoothing purpose to time with complete year data and we decided to discard the period of 1/1/2000-4/30/2000 and retained the 13-year time-series data. For the other two cities, we had complete data from 1/1/1987-12/31/2000 and the analyses covered this period.

4.2 METHODS

Both the standard approach and the partial regression approach were used to analyze the data for these four cities. We modeled the association between the non-accidental mortality

for the people 75 years of age and older and the same day PM_{10} level. The modeling of the association between the mortality and the lag of air pollution could have been done. However, the purpose of this analysis is to illustrate the use of the standard approach and partial regression approach with some real examples.

We used a smoothing spline with 7 degrees of freedom per year as the smoothing function for the calendar time. The total degrees of freedom changed for the four cities, since we had

Table 4.1: Time period for available data

Variable	Pittsburgh	Chicago	Minneapolis	Detroit
Deaths	1/1/87-12/31/00	1/1/87-12/31/00	1/1/87-12/31/00	1/1/87-12/31/00
Daily PM_{10}	1/1/87-12/31/98	1/1/87-12/31/00	1/1/87-12/31/00	1/1/87-04/30/00
Dptp	1/1/87-12/31/00	1/1/87-12/31/00	1/1/87-12/31/00	1/1/87-12/31/00
Tmpd	1/1/87-12/31/00	1/1/87-12/31/00	1/1/87-12/31/00	1/1/87-12/31/00
Time frame	1/1/87-12/31/98	1/1/87-12/31/00	1/1/87-12/31/00	1/1/87-12/31/99
# of Years	12	14	14	13

* Death: Daily death of person 75 years of age and older, all cause excluding accident; PM_{10}: 24 hourly mean PM_{10} ($\mu g/m^3$); Dptp: Dew point temperature (^0F); Tmpd: Average of Tmax and Tmin.

Table 4.2: Descriptive statistics for the selected variables

City	Deaths (ages 75+)	PM_{10} ($\mu g/m^3$)	Temperature (^0F)	Dew point temperature (^0F)	Days of PM_{10} Measured
Pittsburgh	21	35.5	52.0	41.1	4358
Chicago	58	37.1	50.2	40.4	4863
Minneapolis	18	28.2	48.0	35.2	4449
Detroit	22	41.0	50.0	39.8	4324

* For the time period given in Table 4-1.

different number of years of data. The average of 24-hour maximum and 24-hour minimal temperature was used to represent the daily temperature measure. We were unable to use the 24-hour average temperature as this measure was not always available during the whole time period. We used a smoothing spline with 6 degrees of freedom for the temperature. The dew point temperature was included in the model and a smoothing spline with 3 degrees of freedom was assigned for this variable. We also included the day of week variables and PM_{10} measure. The inclusion of the variables and assignment of degrees of freedom described above were in accordance with the NMMAP study [Dominici et al., 2002]. In the NMMAP study, a variable, agacat, was included to indicate the age group of the people and an interaction of agacat and dew point temperature was also included. Since we were modeling the mortality for the people 75 years of age and older, it was not necessary for us to include these variables. Table 2 lists some descriptive statistics from the data used in the present study.

We also performed sensitive analyses with respect to the smoothing parameter, degrees of freedom, choosing variables one at a time. First, we fixed the degrees of freedom for temperature at 6 and the degrees of freedom for dew point temperature at 3, and then assigned the following degrees of freedom: 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5 and 9 df/year for time. Second, we fixed the degrees of freedom for time at 7/year and the degrees of freedom for dew point temperature at 3, and then assigned the following degrees of freedom: 4, 5, 6, 7, and 8 for temperature. Finally, we fixed the degrees of freedom for time at 7/year and the degrees of freedom for temperature at 6, and then assigned the following degrees of freedom: 1, 2, 3, 4, 5 and 6.

4.3 RESULTS

Table 4-3 summarizes the result of the analyses of the four chosen cities. The first column gives the name of the city. The second column indicates the concurvity in the β. The third and fourth columns represent the estimate of β using the standard approach and partial regression approach. The fifth column is the default standard given by S-plus, the sixth column is the exact standard error given by *gam.exact* and the seventh column is the

standard error given by *gam.partial.residual*. Form the table, we find that the concurvity does ubiquitously exist in air pollution data, which is around 0.55 in all the four cities. Thus it is very important to adjust for the influence of concurvity. We note that the estimates using the partial regression approach are always lower than the estimates using the standard approach. From our simulation study and the other study [Ramsay et al., 2003b], the estimate of β using the standard approach was shown to have upward bias. The estimate of β by using the partial regression can also have upward bias. However, the magnitude o the bias in the partial regression is smaller than that in the standard approach over the whole range of concurvity level, seen in Section 3. We note that there are some discrepancies between the estimates of β from these two approaches in all the four cities and those discrepancies are not negligible. The default standard errors are smaller than the standard errors given by the *gam.exact* and *gam.partial.residual*, which is just as we expected. We also find that the standard errors given by *gam.exact* are very close to the standard errors given by *gam.partial.residual*, which suggest that, compared the magnitude of bias reduction, the inflation in standard error is negligible.

Figure 4-1, 4-2 and 4-3 show the results from our sensitivity analyses. In figure 4-1, we change the degrees of freedom for time with fixed degrees of freedom for temperature and dew point temperature. The plots show that the standard approach always has a larger estimate of β than the partial regression approach, which are consistent with our simulation

Table 4.3: Results of the real data analyses

		β		**Standard error**		
City	Concurvity	Standard	Partial	Default	Standard	Partial
Pittsburgh	0.56	0.0004595	0.0004253	0.0001613	0.0001744	0.0001737
Chicago	0.57	0.0004729	0.0004153	0.0001066	0.0001185	0.0001298
Minneapolis	0.51	0.0003769	0.0002897	0.0002684	0.0002947	0.0002983
Detroit	0.58	0.0004812	0.0003922	0.0001669	0.0001907	0.0001973

*Degrees of freedom: 7/year for time, 6 for temperature and 3 for dew point temperature.

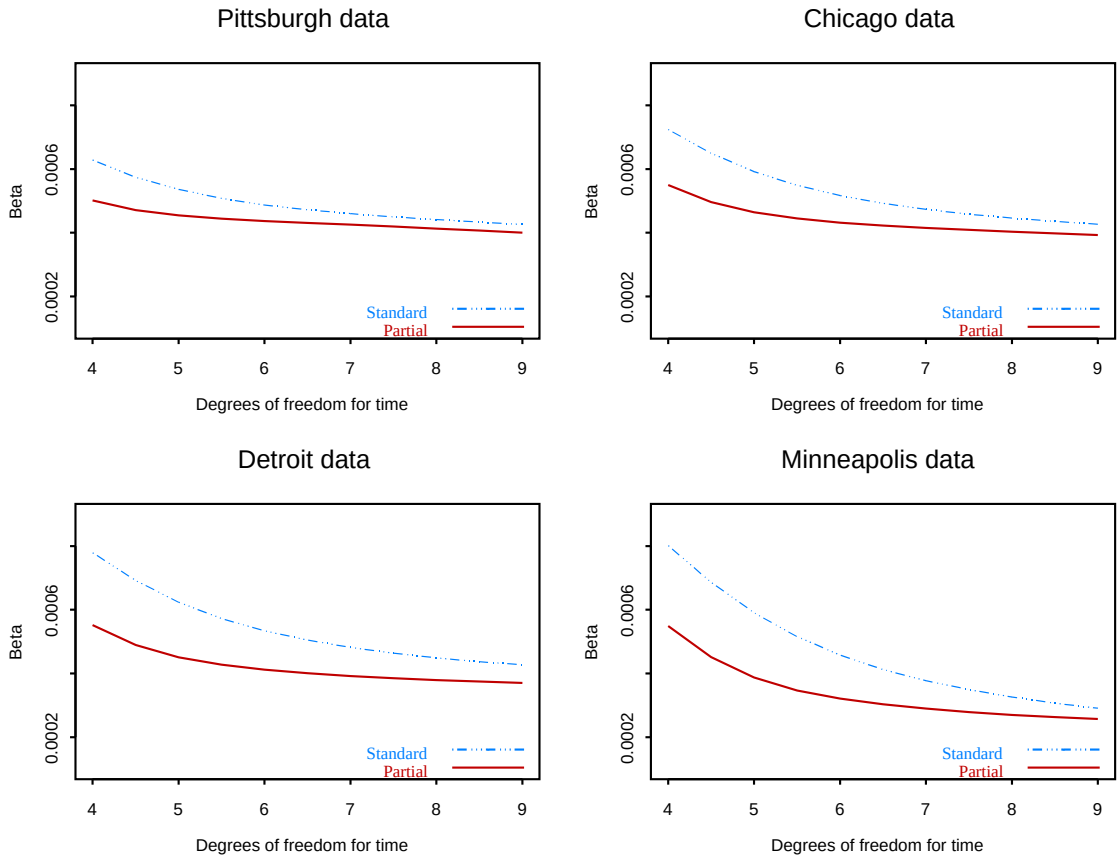

Figure 4.1: Sensitive analyses with difference degrees of freedom for time

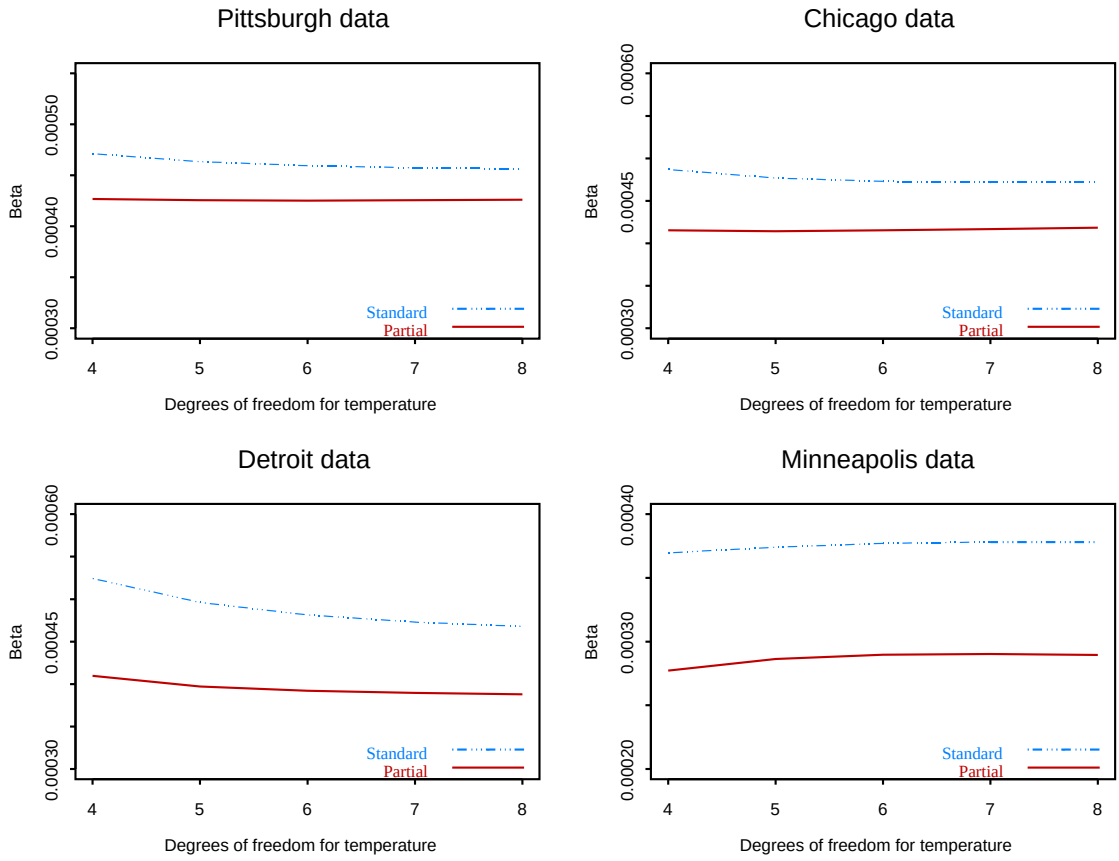

Figure 4.2: Sensitive analyses with difference degrees of freedom for temperature

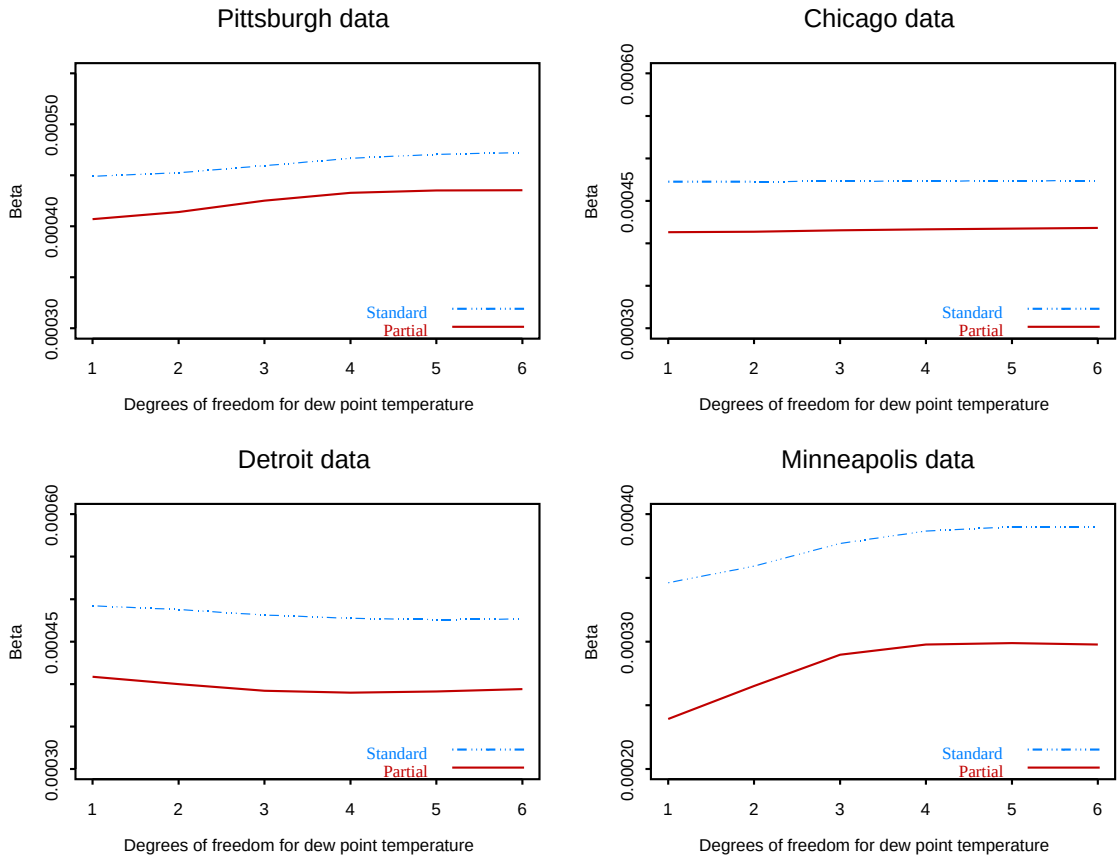

Figure 4.3: Sensitive analyses with difference degrees of freedom for dew point temperature

results, in which the standard approach always has a larger upward bias than the partial regression approach. Also the estimates of β change quicker in the standard approach than the partial regression approach and the partial regression approach curves are more flat that the standard approach curve. This confirms that the results from the partial regression approach are less sensitive to choice of the degree of smoothing, which is a critical concern in air pollution study. In figure 4-2, we changed the degrees of freedom for temperature with fixed degrees of freedom for time and dew point temperature. In figure 4-3, we change the degrees of dew point freedom for temperature with fixed degrees of freedom for time and temperature. We find that there is no big difference between the standard approach and the partial regression approach. The estimates from both approaches are relatively insensitive to how much to smooth the temperature and dew point freedom, which is just as we expected - the time effect play a much more important role than the temperature effect.

5.0 CONCLUSION

The result from the simulation study and real data analyses show that the partial regression approach performs better than the standard approach (using *gam.exac*) when concurvity is present in the data. Even without concurvity, the partial regression approach performs as well as the standard approach. Given the fact that some degree of concurvity is alway present in the air poluution data, we recommend more use of the partial regression approach.

BIBLIOGRAPHY

[1] Abbey DE, Nishino N, McDonnell WF, Burchette RJ, Knutsen SF, Beeson WL and Yang JX (1999) Long-term inhalable particles and other air pollutants related to mortality innonsmokers. *Americal Journal of Respiratory and Critical Care Medicine* 159: 373-382.

[2] Bateson TF, Schwartz J (1999) Control for seasonal variation and time trend in case-crossover studies of acute effects of environmental exposures. *Epidemiology* 10: 539-44.

[3] Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, Ferry BG and Speizer FE (1993) An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* 329:1753-1759.

[4] Dominici F, Daniels M, Zeger SL (2002a) Air pollution and mortality: estimating regional and national dose-response relationships. *Journal of the American Statistical Association* 97:100-11.

[5] Dominici F, McDermott A, Zeger SL, Samet JM (2002b) On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology* 156:193-203.

[6] Dominici F, McDermott A, Hastie T (2004) Improved Semi-Parametric Time Series Models of Air Pollution and Mortality. Revision *http://www.biostat.jhsph.edu/f̃dominic/jasa .R2.pdf*

[7] Durban M, Hackett C, and Currie I (1999) Approximate Standard Errors in Semiparametric Models. *Biometrics* 55:699-703.

[8] Figueiras A, Roca-Pardias J and Cadarso-Surez C (2003) Avoiding the effect of concurvity in Generalized Additive Models in Time-series studies of Air Pollution. *The ISI International Conference on Environmental Statistics and Health* (http://isi-eh.usc.es/trabajos/110_70_fullpaper.pdf)

[9] Hastie TJ, Tibshirani RJ (1990) *Generalized additive models*. New York, Chapman and Hall, Inc.

[10] Lumley T, Sheppard L (2003) Time Series Analyses of Air Pollution and Health: Straining at Gnats and Swallowing Camels? *Epidemiology* 14:113-14.

[11] McCullagh, P. and Nelder JA (1989) *Generalized Linear Models (Second Edition)*. New York: Chapman & Hall.

[12] Pope CA III, Thun MJ, Namboordiri MM, Dockery DW, Evans JS, Speizer FE, Heath Jr. CW (1995) Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory and Critical Care Medicine* 151:669 - 674.

[13] Ramsay TO, Burnett RT, Krewski D (2003a). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology* 14:18-23.

[14] Ramsay TO, Burnett RT, Krewski D (2003b). Exploring Bias in a Generalized Additive Model for Spatial Air Pollution Data. *Environmental Health Perspectives Environmental Health Perspectives* 111:1283-1288.

[15] Rice J (1986) Convergence rates for partially splines models. *Statistics & probability letters* 4:203-208.

[16] Samet JM, Dominici F, McDermott A, Zeger SL (2003) New problems for an old design: time series analyses of air pollution and health. *Epidemiology* 14:11-2.

[17] Schwartz J and Marcus A. (1990) Mortality and Air pollution in London: A time serires analysis. *American Journal of Epidemiology* 131:185-194.

[18] Schwartz J (1994) The use of generalized additive models in epidemiology. *Proceedings of the17th International Biometric Conference* 55-80.

[19] Schwartz J (1995) Air Pollution and Daily Mortality in Birmingham, Alabama. *American Journal of Epidemiology* 137:1136-1147.

[20] Schwartz J (1999) The distributed lag between air pollution and daily death. *Epidemiology* 50:413-436.