Paper 203-29

Bayesian Model Averaging Using SAS® Software

Melissa Whitney and Long Ngo, Harvard School of Public Health, Boston, MA

ABSTRACT

The flexibility of the mixed-model framework in SAS is well-suited for a wide variety of statistical modeling applications. Smoothing methods that use penalized splines can be employed using generalized linear mixed-effects models. These models allow researchers to relax the strict linearity assumption required for traditional regression while simultaneously modeling other relationships of interest. For this paper, data on daily mortality, weather and seasonal variables, and average daily air pollution (PM_{10}) levels in Chicago were used to create log-linear models to predict total daily mortality. To allow for nonparametric adjustments for weather and seasonal variables, random-effects penalized splines were included in a generalized linear mixed-effects model using SAS. In order to estimate a PM_{10}-mortality dose-response curve, expected daily mortality counts were modeled as a function of PM_{10} using truncated linear fixed-effects splines with set numbers and locations of knots. All possible models in this pre-specified family of models were then averaged (weighted by posterior probabilities computed based on the model's BIC) to estimate a final dose-response curve. This technique results in a model-averaged dose-response relationship that incorporates model uncertainty with respect to both location and number of knots while allowing for non-parametric adjustment of continuous covariates such as weather and seasonality.

INTRODUCTION

In recent years, a wide variety of smoothing techniques have been proposed to allow for greater flexibility in modeling data by relaxing the linearity assumption required for standard parametric regression. Nonparametric regression relaxes the traditional regression assumption of linearity to examine non-linear relationships between variables of interest. Smoothing using mixed-model software is a powerful tool to model semi-parametric or non-parametric data. The mixed-model approach to smoothing offers several advantages over other smoothing techniques currently available in statistical packages [1].

The popularity of Hastie and Tbishirani's generalized additive modeling approach [5] has led to the development of the `gam()` function in SPLUS, and PROC GAM in SAS [6], which are also designed to fit non-parametric or semi-parametric additive models. However, PROC GAM requires a complex, weighted back-fitting algorithm; whereas, using the mixed model approach, one needs only to construct the design matrices for fixed and random effects, and maximum likelihood or quasi-maximum likelihood estimation method can then be used. The degree of smoothness (degree of freedom) of the nonparametric function depends on the smoothing parameters, which are a function of the variance components. User specification of degrees of freedom (or smoothing parameter) is possible in GAM, whereas in the mixed model framework, restricted maximum likelihood (REML) is used for estimating variance components, from which the degrees of freedom can be calculated. In addition, given the desired degrees of freedom, the appropriate variance components can also be derived and used in the mixed model. In the back-fitting algorithm used in PROC GAM, model fit criterion such as the Bayesian Information Criterion (BIC) is not available, so it is difficult to construct the model weight used in the weighted average of the fitted values or the estimated coefficients of the individual models. In addition, the parametric formulation of the smoothing functions using mixed models provides a more intuitive understanding and interpretation of the model.

Our analysis focused on two aims. First, the adjustment for the potential known confounders (relative humidity, temperature, and seasonality) of daily total mortality counts had to be taken into account. These confounders exhibited nonlinear relationships. Smoothing splines are ideal functions to capture the nonlinear effects of these variables. Secondly, the dose-response relationship between mortality counts and air pollution (PM_{10}) levels could also be nonlinear; however we were not certain of the nonlinear relationship between these two variables. For the three mentioned potential counfounders, we were not interested in the estimated coefficients of the smoothing functions, so we used three nonlinear smoothing functions to obtain the adjustment for the estimates of the exposure variable PM_{10}. For the exposure variable, PM_{10}, we wanted to obtain the estimated coefficients (the slopes) of the segmented linear functions that make up the dose-response relationship. We considered the relationship between mortality counts and PM_{10} to be linear, or to be segmented linear based on the intervals specified by predetermined knots of PM_{10}. All possible models created from different knot specifications generated estimated coefficients that were then weight-averaged to yield the overall dose response curve.

AIR POLLUTION DATA

Numerous studies have demonstrated a consistent, positive association between particulate matter concentration and mortality. These findings motivate current research efforts to develop reliable methods to characterize the relationship between particulate matter (PM_{10}) and daily mortality counts in cities worldwide. The dataset described

in this section serves as an example of the utility of the mixed-model framework, paired with Bayesian Model Averaging techniques, to develop reliable methods to model dose-response relationships. The data for this study consist of daily citywide mortality counts, weather and seasonal variables, and average 24-hr. PM_{10} levels for Chicago over a 6-year span. Table 1 shows the variables used in the analysis of the data for the city of Chicago.

- 24-hr. average mean PM_{10} (variable name: `pmmean`) levels. This is particulate matter smaller than 10 microns in size, measured in $\mu g/m^3$. The mean is 35.6 and SD of 17.1. The median is 32.2.
- Daily relative humidity (`rhum`) has a mean of 71.1, SD of 12.4, and a median of 71.0.
- Daily average temperature (`temp`) has a mean of 49.8, SD of 19.2, and a median of 49.8.
- Day of the week indicator variables (`dow`) goes from 1 (Monday) to 7 (Sunday).
- Daily mortality counts from the National Center for Health Statistics (`totmort`) for Chicago
- Complete environmental time series data for approximately 2000 consecutive days (`daystud`)

Table 1. Summary of Air Pollution Dataset

ANALYSIS OF AIR POLLUTION DATA

In the following subsections, the process of creating a final, model averaged dose-response curve is outlined in detail.

EXPLORING NONLINEAR RELATIONSHIPS BETWEEN POTENTIAL CONFOUNDERS AND MORTALITY COUNTS

First, potential confounders were examined so that appropriate adjustment for the effect of PM_{10} could be made later in the analysis. For each of the three variables (relative humidity, temperature, seasonality), a smoothing method that uses linear basis functions with penalized splines was employed using generalized linear mixed-effects models with log links. The fixed effects were the intercept and the linear term of the single covariate, and the random effects were terms from the truncated linear function defined by the knots (see the section MODEL FORMULATION for the specification of the fixed and random design matrix X, and Z). The fitted values were obtained and plotted against each of the confounders.

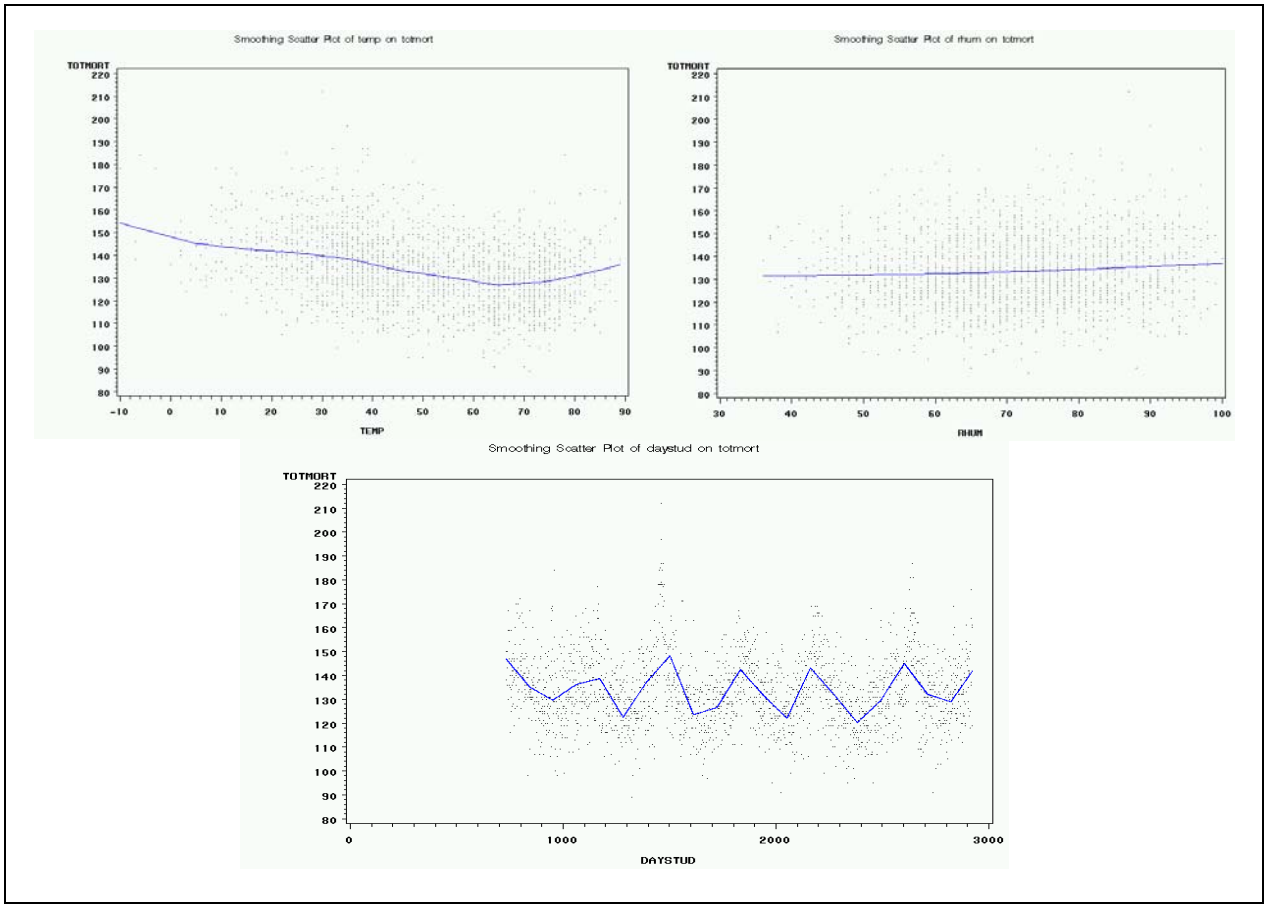

Figure 1. Univariate Plots of Smoothed Temperature, Relative Humidity, and Seasonality

Figure 1 shows the apparent nonlinear relationships for temperature and seasonality. The low mortality counts observed for temperature between 60 and 70 degrees appeared reasonable. Extreme temperatures tend to be associated with higher mortality counts. The seasonal cyclical trend observed for seasonality (`daystud`) was expected. Relative humidity appears to be linearly related to mortality counts, with a slight increase of mortality counts as humidity increases. Due to the availability of the large data set and the observed trends above, a decision was made to control these confounders nonparametrically, with 13 knots for relative humidity, 19 knots for temperature, and 35 knots for seasonality [2]. These plots were created using linear penalized splines with knot points at every 5 measurements of covariate, or a max of 35 knots total per covariate. Previous investigations have determined that the use of 35 knots per predictor leads to a stable model with a moderate level of smoothness.

Knots for smoothing splines for `temp`, `rhum`, and `daystud` were created using the `default_knots` macro below. This macro is called for each smoothing function to automatically configure the knot vector. A knot is set at every fifth value of the variable, and the number of knots is limited to 35. An option is also available for user-specified number of knots. The input parameter `librefknots` specifies the libname, `data` is the name of the dataset containing the confounders, `knotdata` is the name of the output dataset containing the knot vector, `varknots` is the variable name of the confounder, and the optional `numknots` is for user-specified number of knots.

```
%macro default_knots(librefknots=,data=,knotdata=,varknots=,numknots=);
proc sort data=&data (keep=&varknots) out=q1;
  by &varknots;
run;
data q2;
  set q1;
  by &varknots;
  if first.&varknots;
run;
data &librefknots..&knotdata;
  set q2 nobs=n;
  knotsp=int(n/5);
  if knotsp>=35 then kmx=35; else
  if knotsp<35 then kmx=knotsp;
  %if &numknots ne  %then %do;
    ktemp=&numknots;
    if 1 <= ktemp <= 35 then kmx=ktemp;
  %end;
  kintrvl=round(n/kmx);
  knotsok=mod(_n_,kintrvl);
  knots=&varknots;
  if knotsok=0 or _n_=n-1 then output;
keep knots;
run;
%mend;
```

MODEL FORMULATION

Next, a generalized linear mixed-effects model (Poisson) was formulated for assessing the association between log of mortality count and PM_{10} controlling nonparametrically for temperature, relative humidity, and seasonality. Day of the week, a categorical variable to distinguish between weekdays, was also adjusted for in this analysis. The dose-response relationship between 24-hr. mortality count and PM_{10} was assumed to follow one of three possible forms: linear (no knots), segmented linear with one knot, or segmented linear with two knots. Using the above assumptions, one can then average across all these segmented linear models to obtain the final averaged model.

Let y_i be the mortality count for day i,

x_{1i} be the PM_{10} level,

x_{2i} be the mean-adjusted ($x_{2i} - \bar{x}_2$) relative humidity,

x_{3i} mean-adjusted temperature,

x_{4i} mean-adjusted day of study (seasonal), and

x_{5i} indicator variable for day of the week.

Also let $\kappa^{(1)}$ be the knot vector of length l_1 for PM_{10},

$\kappa^{(2)}$ and l_2 for relative humidity,

$\kappa^{(3)}$ and l_3 for temperature, and

$\kappa^{(4)}$ and l_4 for day of study.

Using the above notation, a Poisson generalized linear mixed-effects model can be formulated as:

$$\log E(y_i \mid U) = \beta_0 + g_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}) + \gamma_{(I=weekday)} x_{5i} \qquad (1)$$

where U is the set of specified random effects $\mu_m^{(j)} s$ in the model. The additive nonlinear functions for relative humidity, temperature, and seasonality are expressed as follows:

$$f_j(x_{ji}) = \beta_1 x_{ji} + \sum_{m=1}^{l_j} \mu_m^{(j)} (x_{ji} - \kappa_m^{(j)})_+ \quad \text{and} \quad j = 2,3,4 \qquad \mu_m^{(j)} \sim N(0, \sigma_j^2)$$

$$(x_{ji} - \kappa_m^{(j)})_+ = \begin{cases} 0 & \text{if } x_{ji} \le \kappa_m^{(j)} \\ x_{ji} - \kappa_m^{(j)} & \text{if } x_{ji} > \kappa_m^{(j)} \end{cases}$$

The design matrix for the fixed effects thus has the form:

$X = [1\ x_{1i}\ x_{2i}\ x_{3i}\ x_{4i}\ x_{5i}]_{1 \le i \le n}$, where n is the total number of days.

The design matrix for the random effects for relative humidity, temperature, and seasonality has the form:

$$Z = [(x_{2i} - \kappa_m^{(2)})_{+1 \le m \le l_2} \mid (x_{3i} - \kappa_m^{(3)})_{+1 \le m \le l_3} \mid (x_{4i} - \kappa_m^{(4)})_{+1 \le m \le l_4}]_{1 \le i \le n}$$

The knot vectors of PM_{10} is specified from the following set or no knots:

ψ = (15, 25, 35, 45, 55, 65, 75 $\mu g / m^3$).

There are a total of 29 models with no knots (1), 1 knot (7), and 2 knots (21). For the no knot model (1) becomes:

$$\log E(y_i \mid U) = \beta_0 + \beta_1 x_{1i} + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}) + \gamma_{(I=weekday)} x_{5i}.$$

For the 1-knot models, at knot $v_1 \in \psi$ we have the following model:

$$\log E(y_i \mid U) = \beta_0 + \beta_1 x_{1i} + \beta_{v_1} (x_{1i} - v_1)_+ + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}) + \gamma_{(I=weekday)} x_{5i}.$$

And for the 2-knot models, at specific knot v_1 and $v_2 \in \psi$:

$$\log E(y_i \mid U) = \beta_0 + \beta_1 x_{1i} + \beta_{v_1} (x_{1i} - v_1)_+ + \beta_{v_2} (x_{1i} - v_2)_+ + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}) + \gamma_{(I=weekday)} x_{5i}$$

The following code shows the use of the GLIMMIX macro for fitting the Poisson generalized linear model with the smoothing splines. This is a model with linear PM_{10} only (no knots model).

```
%include 'c:\glimmix.sas';
    %glimmix(data=paper.an1,
       procopt=order=data,
       stmts=%str(
          model totmort=rhum temp pmmean daystud dow/solution;
          class dow; *day of the week indicator;
          random Z1_1 Z1_2 Z1_3 Z1_4 Z1_5 Z1_6 Z1_7 Z1_8
```

```
            Z1_9 Z1_10 Z1_11 Z1_12 Z1_13
            / solution type=TOEP(1); *rhum;
      random Z2_1 Z2_2 Z2_3 Z2_4 Z2_5 Z2_6 Z2_7 Z2_8
            Z2_9 Z2_10 Z2_11 Z2_12 Z2_13 Z2_14 Z2_15
            Z2_16 Z2_17 Z2_18 Z2_19
            / solution type=TOEP(1); *temp;
      random Z3_1 Z3_2 Z3_3 Z3_4 Z3_5 Z3_6 Z3_7 Z3_8
            Z3_9 Z3_10 Z3_11 Z3_12 Z3_13 Z3_14 Z3_15
            Z3_16 Z3_17 Z3_18 Z3_19 Z3_20 Z3_21 Z3_22
            Z3_23 Z3_24 Z3_25 Z3_26 Z3_27 Z3_28 Z3_29
            Z3_30 Z3_31 Z3_32 Z3_33 Z3_34 Z3_35
            / solution type=TOEP(1); *daystud;

   parms
     (3.6985025e-6)
     (1.1996286e-6)
     (1.3058401e-5)
     (1)/hold=1,2,3,4;),
  error=poisson,link=log)
```

The model statement includes fixed effects terms for relative humidity, temperature, PM_{10}, seasonality, and day of the week. The random statements include the nonparametric smoothing splines for relative humidity (`rhum`), temperature (`temp`), and seasonality (`daystud`). The parms statement contains the starting values for variance components for `rhum`, `temp`, and `daystud`. In order to hold these constant for all models, the `hold` statement is used. These `parms` values correspond to pre-specified degrees of freedom for each smoothing term. Without the `hold` statement, the variance components are estimated by default via REML. To run the one-knot model at 55 $\mu g/m^3$, include the variable $(\text{pmmean}-15)_+$ in the `model` statement above. Note the multiple specification of the `RANDOM` statement and the type of variance-covariance structure specified for the random effects.

BAYESIAN MODEL AVERAGING TO OBTAIN FINAL MODEL

A total of 29 candidate models were fit using the above SAS macro GLIMMIX. For each model, the estimated coefficients of the linear term (in all models), the knot-related terms of PM_{10}, and the model Bayesian Information Criterion (BIC) were obtained [3,4]. The estimated coefficients of the PM_{10} linear term and each of the seven knots for the overall model (the Bayesian model averaging model) were calculated as

$$E(\alpha \mid y) = \sum_{k=1}^{29} P(M_k \mid y) E(\beta \mid y, M_k)$$

where

$P(M_k \mid y)$ is the posterior probability of the model M_k given the data vector y, and

$E(\beta \mid y, M_k)$ is the model-specific estimate of β.

The posterior probability $P(M_k \mid y)$ in the above formula for each model was estimated as

$$P(M_k \mid y) = \frac{P(M_k) e^{-0.5(BIC(M_k) - \overline{BIC(M_k)})}}{\sum_{k=1}^{29} P(M_k) e^{-0.5 BIC(M_k) - \overline{BIC(M_k)}}}$$

The above equation is a modified version of an empirical approximation to a fully Bayesian form of model averaging [3], thus bridging classical Frequentist and Bayesian estimation methods. The BIC is a function of the log likelihood of the model that extracts a penalty according to the number of terms in the model. Thus the above function of the BIC can be used to form a weighted average over all models, in which weights depend on the degree to which data support each model. In addition to heavily weighting the best fitting models, the penalty extracted for dimensionality of the model ensures that parsimonious models are favored as well.

For this particular example, the prior probability $P(M_k)$ was assumed to be from the uniform distribution,

$$P(M_k) = \frac{1}{29}.$$

The partial prediction from PM_{10} was calculated for each model as well as for the final model. The final model's partial prediction due to PM_{10} was calculated as

$$\hat{\alpha}_0 + \hat{\alpha}_m x_{1i} + \sum_{p=1}^{7} \hat{\alpha}_p (x_{1i} - (5+10p)_+)$$

where $\hat{\alpha}_0$ is the averaged intercept, $\hat{\alpha}_m$ is the averaged slope, and $\hat{\alpha}_p$ are the averaged knot coefficients.

Table 2 shows the estimated coefficients for the final model. The estimated coefficients are the weighted estimates of the 29 models' coefficients with the weight being the model's posterior probability. The total posterior probability is the sum of all probabilities from all models in which the coefficient exists. Intercept and PM_{10} estimates have probabilities equal to 1 because these two terms exist in all 29 models. The coefficient at knot 55 is the largest value among the knot coefficients and has the largest summed posterior probability. This indicates that the data support the existence of a knot at a PM_{10} value of 55 μg/m^3 more than any other location among the pre-specified set of knots. Coefficients at knots 15, 25, and 35 are the least supported by the data, and thus contribute less to the final model-averaged fit.

Coefficient	Coefficient Value	Posterior Probability
Intercept	5.090354937	1.00000
PM_{10}	0.000384463	1.00000
Knot at 15	0.000037107	0.02946
Knot at 25	-0.000000772	0.02294
Knot at 35	0.000005290	0.02851
Knot at 45	-0.000005207	0.04503
Knot at 55	-0.000174602	0.14406
Knot at 65	-0.000012574	0.04445
Knot at 75	-0.000038110	0.03564

Table 2. BMA Coefficients and Total Posterior Probabilities of Coefficients

Figure 2 below was created by fitting each of the 29 models to the data, extracting out coefficient estimates for the intercept term, PM_{10} term, and all PM_{10}-related knots, and then plotting these against log daily mortality counts. The resulting plot s are referred to as partial prediction plots, as they plot a prediction of the partial effects of PM10 on daily mortality, with coefficients obtained from models that also included other potential confounders. Finally, the Bayesian model averaged curve of all 29 partial predictions was superimposed (in red) over the 29 contributing models.

The red curve is the partial prediction of the BMA final model, as well as all 29 model's partial prediction curve in blue. Note that the y-axis represents the contribution of PM_{10} to logarithm of mortality counts. The final dose-response model appears fairly linear for these data, with total mortality increasing with PM_{10} levels. The slope appears to decrease slightly when PM_{10} exceeds 55 μg/m^3.

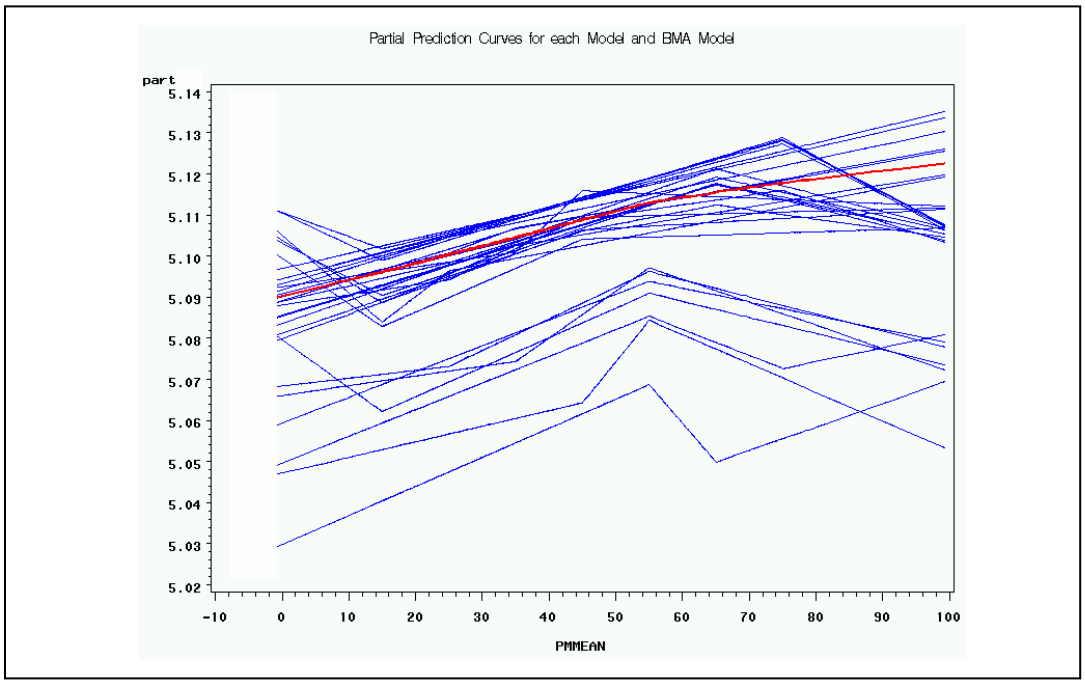

Figure 2. BMA Dose Response Curve (in red) Superimposed over Entire Family of 29 Considered Models (in blue)

CONCLUSION

The mixed-model approach to smoothing presented in this paper provides a flexible framework for employing non-parametric regression techniques in statistical modeling. Smoothing with mixed models allows researchers to relax the strict linearity assumption required for traditional regression techniques, while simultaneously modeling other trends of interest, such as dose-response relationships. The air pollution example utilizes this technique to build a model-averaged dose-response relationship that incorporates model uncertainty with respect to both location and number of knots while allowing for non-parametric adjustment for weather and seasonal covariates. The easily adaptable techniques described in this paper allowed for the development of a fully parametric characterization of the effects of PM_{10} on daily mortality while accounting for the nonparametric effects of other covariates. Future work on this project will aim to address convergence issues with using cubic smoothing splines for long-term seasonal covariates, model-averaging over more extensive families of models, constructing confidence intervals for average model presented above, employing these methods to explore heterogeneity across other cities and regions of the U.S., and developing Bayesian model averaging software for widespread use.

REFERENCES

[1] Ngo L. and Wand M.P. (2004), "Smoothing with Mixed Model Software," *Journal of Statistical Software*, Volume 9, Number 1, 1-56.

[2] Dominici F., Samet M.J. and Zeger L.S. (2000), "Combining Evidence on Air Pollution and Daily Mortality from the Twenty Largest U.S. Cities: A Hierarchical Modeling Strategy," *Journal of the Royal Statistical Society*, Ser. A, 163, 263-302.

[3] Clyde M, Model Averaging (2003), Subjective and Objective Bayesian Statistics (Editor: James Press), Wiley, 320-333.

[4] Daniels, M., Dominici, F., Samet, J.M. and Zeger, S.L. (2000), "Estimating PM10-Mortality Dose-Response Curves and Threshold Levels: An Analysis of Daily Time-Series for the 20 Largest U.S. Cities," *American Journal of Epidemiology*, 152, 397-412.

[5] Hastie T.J. and Tibshirani R.J. (1990), Generalized Additive Models. Chapman and Hall, New York, N.Y., 136-171.

[6] Xiang D., "Fitting Generalized Additive Models with the GAM Procedure," SAS SUGI 26 Proceedings, Statistical and Data Analysis Section, SAS Institute Inc., Cary, NC.

ACKNOWLEDGMENTS

We thank Dr. Louise Ryan, Dr. Joel Schwartz, and Dr. Brent Coull for the data, ideas and advice on this project.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Melissa Whitney
Harvard School of Public Health
Department of Biostatistics, HSPH2
655 Huntington Ave.
Boston, MA 02115
Email: mwhitney@hsph.harvard.edu