2011-01-01

Estimating The Effect Of Dust And Low Wind Events On Hospitalizations For Asthma While Adjusting For Hourly Levels Of Air Pollutants

Priyangi Kanchana Bulathsinhala
University of Texas at El Paso, priyangikb@yahoo.com

ESTIMATING THE EFFECT OF DUST AND LOW WIND EVENTS ON HOSPITALIZATIONS FOR ASTHMA WHILE ADJUSTING FOR HOURLY LEVELS OF AIR POLLUTANTS

PRIYANGI KANCHANA BULATHSINHALA

Department of Mathematical Sciences

APPROVED:

Joan Staniswalis, Chair, Ph.D.

Sara Grineski, Ph.D.

Amy Wagler, Ph.D.

Benjamin C. Flores, Ph.D.
Acting Dean of the Graduate School

to my

MOTHER and FATHER

with love

ESTIMATING THE EFFECT OF DUST AND LOW WIND EVENTS ON HOSPITALIZATIONS FOR ASTHMA WHILE ADJUSTING FOR HOURLY LEVELS OF AIR POLLUTANTS

by

PRIYANGI KANCHANA BULATHSINHALA

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2011

Acknowledgements

There are numerous people I wish to thank for their support in making this thesis possible.

First, I am indebted to my parents for their unconditional love, endless support and guidance.

Next, I wish to express my deepest gratitude to my advisor Dr. Joan Staniswalis for her invaluable insights and expertise. This thesis would not have been possible unless for her genuine support, encouragement, guidance, and patience. I am privileged to have work with such a great researcher and mentor.

I owe my deepest gratitude for Dr.Sara Grineski and Dr. Amy Wagler, my thesis committee members for their precious advice, comments and cooperation. I sincerely thank Prof. Terry Therneau for helping us modifying the ridge function in R-package to accommodate our goal. I am also heartily thankful to Maria Barraza-Rios, Lanna Tallmon, and Maria Salayandia, for their willing support extended to me throughout last two years in numerous ways.

I owe spacial thanks to the University of Texas at El Paso Center for Environmental Resource Management (CERM) and National Institutes of Health (NIH) score grant ISC3GM094073-01 for supporting my research work.

Last but not least, I thank my brother, sister and friends for their love, support and encouragement.

Abstract

El Paso, Texas is known as one of the dust hotspots in North America. We explore the effect of dust and low wind events on asthma admissions in El Paso, Texas between 2000 and 2005. Conditional logistic regression with a case-crossover design was used to estimate the probability of hospitalization after dust and low wind events while controlling for pollutants with hourly monitor measurements, and weather. The historical functional linear model is used to incorporate the hourly pollutant measures into the regression model with a continuous lag, as an alternative to a distributed lag model based on daily averages. The great advantage of the historical functional linear model was demonstrated, namely, the pollutant past exposure was included without having to choose a (short-term) lag. The nonparametric functional linear model in the conditional logistic regression framework is fit by first preprocessing the data, then applying the COXPH function in the R-package for survival analysis. Based on the theoretical relationship between P-splines and ridge regression we modified *COXPH* to allow for a penalty in the nonparametric functional linear model, which was a very time saving approach for us. We proposed that the ridge trace be used to guide the choice of the smoothing parameter in nonparametric functional linear model. The results obtained from the simulated examples suggest that the adapted ridge trace plot can be used to choose a suitable smoothing parameter for the P-spline estimator. We found that both the lag 0 storm and lag 2 low wind are significant at the 10% level of significance suggesting that the probability of asthma hospitalization on a given day is increased by occurrence of dust storms on the same day and low wind events on two days prior to the admissions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Background and Motivation

Located in the Chihuahuan Desert in the far western edge of Texas, El Paso is known as one of the dustiest cities in North America. "Blowing dust" is a typical weather event in El Paso. Novlan et al.(2007) classify blowing dust events in El Paso into two main categories: non-convective (dust storms) and convectively driven events (Haboob dust storms). The convective (Haboob) dust storms are caused by the downdraft of thunderstorms and has much smaller spatial and temporal effects compared to the (non-convective) dust storms[18]. These storms generally pass through a given area very quickly (less than an hour), perhaps covering just one part of the city. Other times, dust storms can cover the entire city of El Paso, Texas and Juarez, Mexico lasting for hours. In our study we ignore the convective/haboob storms and focus only on non-convective dust storms, which we refer to here after as *dust storm days*. According to the records from El Paso International Airport for the period of 1932 through 2005, on average 15 dust events per year have lasted at least for 2 hours [17]. Figure 1.1 shows the monthly frequency distribution of dust events in El Paso based on data from 1932 through 2006 [17]. Based on the monthly frequency distribution of dust events in El Paso, blowing dust events occur most often in March, April and May . Dust emitted from wind-erodible dry land surfaces have a significant impact on air pollution. Hosiokangas et al.(2004) report that both high wind and low wind conditions raise levels of pollutants and particulates in the air. Chronic exposure to increased particulate concentrations could cause major, even deadly respiratory health diseases. Thus, the dusty and windy climate around El Paso city has encouraged many researchers to study the possible causes of respiratory health problems in El Paso [7, 18, 25, 26].

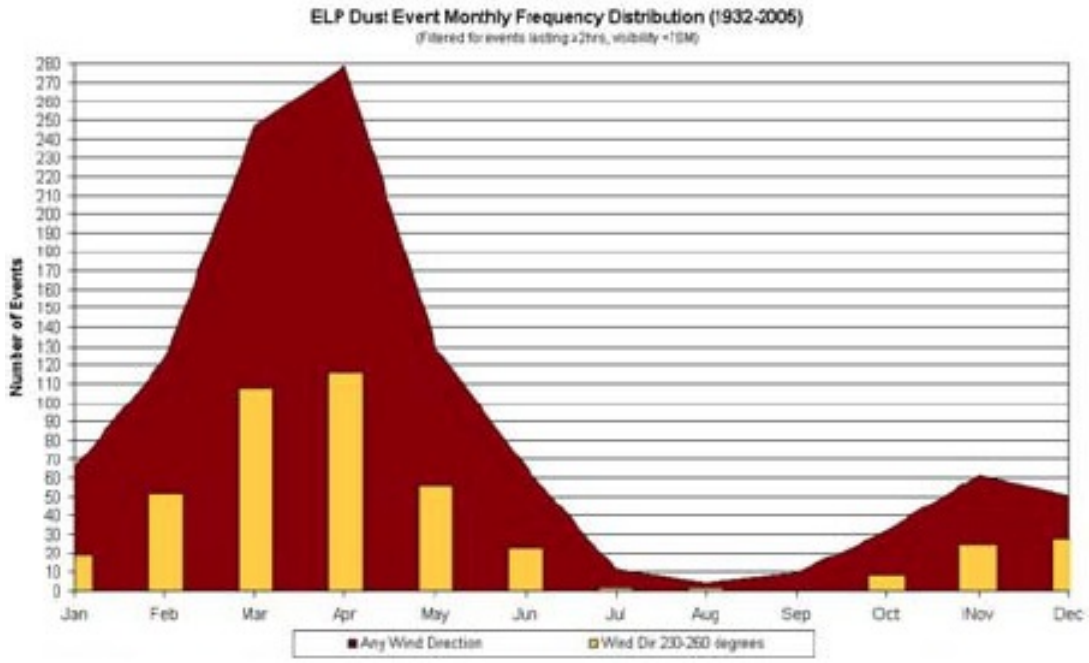


Figure 1.1: Monthly frequency distribution of dust events in El Paso, 1932-2006 (Taken from Novlan et al.(2007)[17]).

In these types of pollution studies, it is usual to consider patient information with the disease of interest together with a few or more pollutant variables such as $PM_{2.5}$ levels (airborne particulate matter with aerodynamic diameter smaller than 2.5 microns), NO_2 (Nitrogen Dioxide), SO_2 (Sulfur Dioxide) and O_3 (Ozone). In addition to these general features another important fact to be considered in such pollution studies is the effect of lagged weather and pollution conditions. That is, how the weather and pollution conditions of earlier dates affect the health outcomes on a given date. In particular, it is not reasonable to assume that the risk of hospital admissions on a give date only depends on weather conditions of that date. We need to consider the weather conditions on dates prior to the admission date as well. Selecting the best lag time to explain the relationship between hospital admissions and exposure to the pollutants is of major concern. Statistical approaches to select a suitable lag still requires further study.

Belleudi et al.(2010) is one of the best references which exemplifies this difficulty of choosing a lag. They have explored several lag times between exposures and health effects in single pollutant models (one pollution variable in the model at a time). In this study, they have evaluated the effect of PM_{10} (particulate matter with an aerodynamic diameter $\leq 10\mu m$), $PM_{2.5}$ and ultra fine particles (diameter $\leq 0.1\mu m$) on emergency hospital admissions for cardiac and respiratory diseases in the city of Rome; see Figure 1.2, a table extracted from Belleudi et al.(2010), they ran single pollutant models for each lag of the three pollutants. For example, they performed a conditional logistic regression analysis with a case-crossover design to study the association between $PM_{2.5}$ and hospitalizations for acute coronary syndrome. They explored the individual lag effects of $PM_{2.5}$ from day 0 to day 6, and then short term effects by taking average over lag 0 through 1 day and lag 0 through 2 days, and extended effects by averaging over lag 0-5 days. Even though this allows them to make comparisons with the past literature, it appears to be a time consuming approach and suffers from the statistical problems inherent in multiple comparisons.

As mentioned earlier, many studies have been done to investigate the respiratory health effects due to exposure to air pollutants [6, 7, 18, 25, 19, 26, 1], each taking different modeling approaches to overcome drawbacks of existing methods. One naive approach would be to model the relative risk of the disease including all the lags of pollutants of interest, in the regression model. The main difficulty with this approach is the existence of multicollinearity. The problem of multicollinearity arises between multiple pollution variables; whether they are different lags of the same pollutant or different pollutants such as SO_2 (Sulfur dioxide) and O_3 (Ozone). It is well known that the presence of multicollinearity, inflates the variance of the parameter estimators [11]. Also, it may complicate interpretations of the parameter estimates as well. For example, suppose we included lag 0 through 5 of a particular pollutant variable in our regression model and obtained the first two lags to be negative and the rest to be positive. In this kind of a situation it is quite difficult to explain

TABLE 3. Associations (Percentage Changes[a]) Between PM_{10}, $PM_{2.5}$ and Particle Number Concentration and Admissions for Acute Coronary Syndrome, Heart Failure, Lower Respiratory Tract Infections, and COPD in Patients Aged ≥35 Years in Rome During the Study Period

Lag	Acute Coronary Syndrome % Change (95% CI)	Heart Failure % Change (95% CI)	Lower Respiratory Tract Infections % Change (95% CI)	COPD % Change (95% CI)
PM_{10}				
0	1.07 (−0.03 to 2.19)	1.77 (0.05 to 3.53)	−0.23 (−2.35 to 1.94)	0.40 (−1.41 to 2.25)
1	0.87 (−0.17 to 1.93)	0.89 (−0.74 to 2.56)	0.35 (−1.69 to 2.44)	−1.22 (−2.93 to 0.52)
2	1.04 (0.01 to 2.08)	0.39 (−1.21 to 2.01)	2.17 (0.17 to 4.21)	−0.66 (−2.34 to 1.04)
3	−0.35 (−1.71 to 1.03)	1.24 (−0.29 to 2.79)	1.21 (−0.69 to 3.15)	−1.64 (−3.21 to −0.04)
4	−1.27 (−2.61 to 0.08)	1.11 (−0.39 to 2.64)	0.62 (−1.25 to 2.53)	−0.86 (−2.44 to 0.75)
5	−1.66 (−2.99 to −0.31)	1.44 (−0.07 to 2.99)	0.30 (−1.56 to 2.20)	0.08 (−1.51 to 1.68)
6	−0.12 (−1.46 to 1.24)	1.09 (−0.43 to 2.62)	0.64 (−1.24 to 2.55)	0.17 (−1.41 to 1.78)
0–1	1.18 (−0.06 to 2.43)	1.70 (−0.24 to 3.67)	0.24 (−2.15 to 2.69)	−0.59 (−2.61 to 1.48)
0–2	1.46 (0.10 to 2.83)	1.55 (−0.57 to 3.72)	1.36 (−1.27 to 4.06)	−0.73 (−2.94 to 1.53)
0–5	−0.39 (−1.99 to 1.24)	2.77 (0.23 to 5.38)	1.56 (−1.55 to 4.76)	−1.96 (−4.54 to 0.69)
0–6	−0.55 (−2.24 to 1.17)	3.19 (0.50 to 5.95)	2.10 (−1.19 to 5.50)	−1.81 (−4.53 to 0.98)
$PM_{2.5}$				
0	2.29 (0.45 to 4.17)	2.38 (0.33 to 4.48)	−0.31 (−2.78 to 2.21)	1.88 (−0.27 to 4.09)
1	1.85 (0.11 to 3.61)	0.48 (−1.42 to 2.42)	1.38 (−0.99 to 3.80)	−2.50 (−4.44 to −0.52)
2	1.99 (0.32 to 3.68)	0.96 (−0.90 to 2.84)	2.82 (0.52 to 5.19)	1.76 (−0.18 to 3.73)
3	0.36 (−1.23 to 1.98)	1.49 (−0.29 to 3.30)	3.04 (0.83 to 5.30)	−0.04 (−1.88 to 1.84)
4	−0.66 (−2.21 to 0.91)	1.67 (−0.08 to 3.45)	1.52 (−0.61 to 3.68)	0.51 (−1.30 to 2.35)
5	−0.56 (−2.13 to 1.03)	1.24 (−0.51 to 3.01)	0.25 (−1.84 to 2.38)	0.02 (−1.76 to 1.83)
6	0.31 (−1.26 to 1.91)	0.31 (−1.40 to 2.06)	−0.18 (−2.28 to 1.97)	0.37 (−1.42 to 2.18)
0–1	2.67 (0.42 to 4.97)	2.17 (−0.30 to 4.70)	0.73 (−2.25 to 3.80)	−1.45 (−3.98 to 1.14)
0–2	2.94 (0.44 to 5.50)	2.58 (−0.20 to 5.44)	2.49 (−0.92 to 6.01)	0.02 (−2.83 to 2.96)
0–5	1.59 (−1.53 to 4.80)	2.58 (−0.93 to 6.22)	3.71 (−0.57 to 8.17)	−0.17 (−3.71 to 3.49)
0–6	1.47 (−1.90 to 4.96)	1.94 (−1.83 to 5.85)	3.62 (−0.96 to 8.42)	−0.08 (−3.87 to 3.87)
Particle number concentration				
0	−0.10 (−1.35 to 1.16)	1.80 (0.39 to 3.24)	−0.40 (−2.21 to 1.44)	1.59 (0.03 to 3.18)
1	0.12 (−1.06 to 1.31)	0.62 (−0.70 to 1.95)	−1.36 (−3.02 to 0.34)	0.18 (−1.26 to 1.63)
2	0.19 (−0.97 to 1.37)	1.65 (0.32 to 3.00)	0.19 (−1.48 to 1.90)	−1.01 (−2.41 to 0.41)
3	−0.48 (−1.63 to 0.69)	0.81 (−0.50 to 2.13)	0.29 (−1.37 to 1.98)	−0.70 (−2.09 to 0.71)
4	0.11 (−1.06 to 1.29)	1.02 (−0.30 to 2.35)	−0.01 (−1.66 to 1.67)	−0.83 (−2.23 to 0.59)
5	−0.01 (−1.18 to 1.17)	0.01 (−1.30 to 1.33)	−0.43 (−2.09 to 1.26)	0.13 (−1.28 to 1.57)
6	0.38 (−0.80 to 1.58)	−0.02 (−1.35 to 1.31)	0.77 (−0.95 to 2.51)	0.39 (−1.03 to 1.83)
0–1	−0.06 (−1.46 to 1.36)	1.65 (0.07 to 3.26)	−1.59 (−3.59 to 0.45)	0.95 (−0.80 to 2.73)
0–2	0.11 (−1.42 to 1.66)	2.00 (0.25 to 3.77)	−1.18 (−3.36 to 1.05)	−0.11 (−1.98 to 1.80)
0–5	0.07 (−1.87 to 2.04)	2.38 (0.17 to 4.65)	−0.99 (−3.73 to 1.83)	−0.70 (−3.04 to 1.70)
0–6	0.56 (−1.53 to 2.69)	2.68 (0.29 to 5.13)	−0.41 (−3.38 to 2.66)	−0.15 (−2.68 to 2.44)

[a]The percentage changes are estimated for 14 $\mu g/m^3$ PM_{10}, 10 $\mu g/m^3$ $PM_{2.5}$, and 9392 particles/cm^3 for particle number concentration.

Figure 1.2: Results of single pollutant models (Taken from Belleudi et al.(2010)[1])

the relationship between the pollutant and the health outcome of the study. In the past, the most common approach to dealing with multicollinearity was to drop non-significant variables using variable selection methods. But using variable selection for dropping variables might still introduce specification bias [2]. Specification bias is a consequence of variable deletion from a regression model, this is explained next. In detail, suppose we have a response variable Y and q predictor variables (X_1, X_2 ,...., X_q). Then the multiple

linear regression model can be written as:

$$y_i = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij} + \epsilon_i, \quad i = 1, 2, ...n$$

where y_i is the response of the i^{th} subject, x_{ij} is the value of the j^{th} predictor variable corresponding to the i^{th} subject, β_j are regression parameters and ϵ_i are the random errors. Instead of a regression model with the full set of predictor variables, consider the reduced regression model with a subset of variables ($p < q$).

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

Consider the following vectors and matrices:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \; X = \left[\begin{array}{cccc|ccc} 1 & x_{11} & \cdots & x_{1p} & x_{1(p+1)} & \cdots & x_{1q} \\ 1 & x_{21} & \cdots & x_{2p} & x_{2(p+1)} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} & x_{n(p+1)} & \cdots & x_{nq} \end{array}\right]_{n\times q}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ -- \\ \beta_{p+1} \\ \vdots \\ \beta_q \end{bmatrix}_{q\times 1} = \begin{bmatrix} \boldsymbol{\beta}_p \\ \boldsymbol{\beta}_r \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}.$$

The matrix $X_{n\times(q+1)}$ is partitioned into two submatrices X_p and X_r of dimensions $(n \times (p+1))$ and $(n \times r)$, where $r = q - p$. The β vector is also partitioned into $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_r$ components. Then the regression model containing the full set of variables (q variables) can be written as;

$$Y = X\beta + \epsilon = X_p\beta_p + X_r\beta_r + \epsilon \text{ , where } \epsilon \sim N(0, \sigma^2 I). \tag{1.1}$$

The reduced model with only p variables is;

$$Y = X_p\beta_p + \epsilon. \tag{1.2}$$

Let $\hat{\beta}^*$ and $\hat{\beta}_p$ be the least squares estimates of the regression parameters of the full model (1.1) and the reduced model (1.2),

$$\hat{\beta}^* = (X^T X)^{-1} X^T Y \text{ and } \hat{\beta}_p = (X_p^T X_p)^{-1} X_p^T Y.$$

Recall that $\hat{\beta}^*$ is an unbiased estimator of β. It can be shown that $\hat{\beta}_p$ is a biased estimator for β_p where the expected value of $\hat{\beta}_p$ is given by[2]

$$E(\hat{\beta}_p) = \beta_p + A\beta_r, \text{ where } A = (X_p^T X_p)^{-1} X_p^T X_r.$$

Nevertheless, databased variable selection methods to choose the best lag for pollutants have been used in quite a number of studies: Staniswalis et al. (2005), Perez et al. (2008), and Grineski et al.(2011). Staniswalis et al. (2005) explain how to employ a backward variable selection procedure to select the best lag for the weather variables in modeling the association between daily mortality and temporal dynamics of PM_{10} particle concentration occurring within a 24 hour period. Perez et al. (2008) use the highest odds ratio to select the best lag-time of PM values ($PM_{2.5}$ and PM_{10}) on daily mortality, in studying whether

the outbreaks of Saharan dust exacerbate the effects of man-made pollution. Grineski et al.(2011) also employ the highest odds ratio approach in studying the effects of dust and low wind events on hospital admissions for asthma and acute bronchitis in El Paso, Texas.

There are some modeling approaches that avoid variable selection. One such approach that has been employed in the past is ridge regression. Ridge regression controls the high variances of the estimated coefficients at the expense of incurring some bias [14]. Recall that, an unbiased estimator of the vector of regression coefficients is $\hat{\beta} = [X'X]^{-1}X'Y$. When covariates in the regression model are correlated, the matrix $X'X$ tends to be singular. Ridge regression was developed to overcome the singularity that prevents inverting $X'X$ by adding a small amount to its diagonal. With this adjustment, $\hat{\beta}$ in ridge regression is given by $\hat{\beta} = [X'X + kI]^{-1}X'Y$ which is a biased estimator for β. Ridge trace is a popular tool use to find the biasing constant k in the ridge estimator (see Chapter 4 for more details about ridge regression and ridge trace). Other than that it also has been used for variable selection as well. The method of variable elimination based on ridge trace was originally suggested by Hoerl and Kennard(1970)[8]. Their procedure is[15]:

- Examine the stable coefficients and eliminate the factors with the least predictive power. In other words, the variables with stable coefficients that are small in absolute value are considered for elimination.
- Examine the unstable coefficients and eliminate those factors that cannot hold their predicting power.
- Delete one or more of the remaining unstable coefficients.

McDonald and Schwing (1973) [15] discuss the instability of ordinary least squares estimates of linear regression coefficients in the presence of multicollinearity in pollution studies and illustrate the problems by regressing mortality rates with various socioeconomic, weather and pollution variables. HC (hydrocarbons), NO_x (oxides of nitrogen) and SO_2 (sulfur

dioxide) was used as pollution variables. Total age adjusted mortality rate was used as the response. They found a high sample correlation between HC and NO_x. They discussed the modeling of total mortality data utilizing ridge regression and eliminating explanatory variables with the C_p criteria and the ridge trace plot. By comparing parameter estimates from these three methods: ridge regression, variable selection using C_p criteria and ridge trace, McDonald and Schwing (1973) have attested that the best strategy is to use ridge regression with a "good' value of k, and retain all important variables in the analysis.

Another systematic approach to specifying the lag association between air pollutants and health outcomes is discussed in Schwartz (2000)[24]. More specifically, the method is known as the distributed lag model[21] and it relates the response to the lagged valued predictor variables in a parsimonious fashion. To illustrate the effectiveness of the approach he utilized daily deaths of elderly people (65 years of age and older) with daily PM_{10} measurements in 10 U.S cities (New Haven, Birmingham, Pittsburgh, Canton, Detroit, Chicago, Minneapolis, Colorado Springs, Spokane and Seattle). For each city, a generalized additive Poisson regression model was fit. A smooth function of time was included to capture the seasonal and other long term trends. Other than that, the temperature of the same day and the previous day, relative humidity, barometric pressure on the same day and day of the week were included as smooth functions while PM_{10} was treated as having a linear association with the response. The linear effect of PM_{10} on the daily deaths was incorporated into the generalized additive model through distributed lags. Thus, the proposed model is a semi-parametric regression model which is

$$Log(E(Y)) = covariates + \alpha_0 + S_1(X_1) + ... + S_p(X_p) + \beta_0 Z_0 + ... + \beta_q Z_q \qquad (1.3)$$

where Z_0 is the exposure to PM_{10} on the concurrent day, Z_1 is the exposure on the previous day and etc. The coefficients of the distributed lags of PM_{10} was restricted to be polynomials of of degree d,

$$\beta_j = \sum_{k=0}^{d} \theta_k j^k, \qquad \text{or} \qquad \beta_j = P(j), \text{ where} \qquad P(x) = \sum_{k=0}^{d} \theta_k x^k.$$

By substituting the constraints on β_j's in (1.3) and defining new variables $W_d = Z_1 + 2^d Z_2 + ... + q^d Z_q$, the model was further simplified

$$Log(E(Y)) = covariates + \alpha_0 + S_1(X_1) + ... + S_p(X_p) + \theta_0 W_0 + ... + \theta_d W_d.$$

Note that coefficients of the W's are the parameters of the polynomial distributed lags. Schwartz (2000) has included a maximum lag of 5 days for PM_{10}. For comparison purposes he has included the PM_{10} information into the generalized additive model in 4 different ways: (1) PM_{10} values of the same day as the death happens; (2) averaging daily PM_{10} over lag 0 through 1; (3) distributed lag model with coefficients constrained to be quadratic polynomials (constrained DLM); (4) including daily PM_{10} lag 0 through 5 (unconstrained DLM). He showed that both the constrained and unconstrained distributed lag models had substantially greater effect on daily mortality than the model using only a single day's exposure and moderately larger effects than the two day averaged model.

So far we have discussed different parametric approaches for handling the problem of including multiple lags of pollutants in regression models. Recently, statisticians have started looking for better solutions by incorporating nonparametric regression in the regression models. One such approach is discussed in Zanobetti et al. (2000). They have developed a model fitting procedure called generalized additive distributive lag model by combining two established regression techniques: generalized additive models (extend the multiple linear regression by allowing the continuous variables to be modeled as smooth and unspecified functions) and distributed lag models(DLM) [28]. The effectiveness of this approach was demonstrated through weather/pollution and mortality data from Milan, Italy. They have used the daily mortality counts as the response variable. Air pollution was measured by total suspended particles(tiny airborne particles or aerosols that are less than 100 micrometers are collectively referred to as total suspended particulate matter (TSP)) and meteorological data was expressed by mean temperature and relative humidity. By taking y_t as the mortality count of day t, and x_t as the TSP value at day t, the authors

have used a generalized additive DLM with the canonical link for the Poisson distribution. Their model is;

$$g\{E(y_t)\} = \alpha + \gamma^T z_t + \sum_{j=1}^{d} f_j(s_{jt}) + \sum_{l=0}^{q} \beta_l x_{t-l},\ t = q+1, ..., T$$

where g is a link function, z_t is a vector of variables modeled linearly (dummy variables) and s_{jt} is the j^{th} variable at time t modeled as a smooth function. More specifically, they have included temperature, relative humidity and the day number as smooth functions and dummy variables for days of the week, indicators for holidays and influenza epidemics. As you see, they have incorporated TSP values into the regression model through distributed lags given by $\sum_{l=0}^{q} \beta_l x_{t-l}$ and overall impact of a unit change in exposure over past q days was measured by $\sum_{\ell=0}^{q} \beta_\ell$. The β_l's were restricted to be;

$$\beta_l = \sum_{j=0}^{p} \tau_j l^j + \sum_{k=1}^{K} v_k (l - T_k)_+^p$$

where $(l - T_k)_+$ represents $(l - T_k)$ truncated to be positive, τ_j and v_k are real valued coefficients estimated from the data and $T_1, ..., T_k$ are knots. Thus, β_l is a spline written in terms of power basis. But in the estimation they have used the B-spline basis instead of the power basis (see chapter 2 for more details about the power basis and B-splines). They have used *penalized splines*[4] to obtain smooth estimates for the β_l's. By allowing TSP of previous 45 days to be included in the model, they reported that the estimated overall effect ($\sum_{l=0}^{45} \beta_l$) was larger than the results using just a one or two day mean.

Kong et al.(2010)[10] have proposed a semi-parametric time series model, called the functional additive cumulative time series (FACTS) model, in studying the long-term cumulative effects of air pollutants on respiratory diseases in Hong Kong. Instead of a semi-parametric additive model for the cumulative effects given by

$$E\{Y(t_i)|Z(s), X(s), s \leq t_i\} = \beta^T Z(t_i) + g_1(X_1(t_i - \tau_1)) + ... + g_q(X_q(t_i - \tau_q)), \quad (1.4)$$

where $Y(t_i)$ is a discrete response time series, $\boldsymbol{Z}(t) = (1, Z_1(t), ..., Z_p(t))^T$, $\boldsymbol{X}(t) = (X_1(t), ..., X_q(t))^T$, g_k are unknown link functions and τ_k, $k = 1, ...q$ are lags, they proposed to model the cumulative effect of a single covariate X_1 over τ days by $\int_0^\Delta X_1(t - \tau)\theta(\tau)d\tau$ for some $\Delta > 0$, where $\theta(\tau) \geq 0$ is a weight function defined over $[0, \Delta]$. Note that, here τ has been integrated out. Their proposed model with the additive structure is

$$Y(t) = Z(t)^T\beta_0 + \sum_{k=1}^{q} g_k\bigg(\int_0^\Delta X_k(t - \tau)\theta_k(\tau)d\tau\bigg) + \epsilon(t) \quad (1.5)$$

where $g_k(\cdot)$ and $\theta_k(\cdot) > 0$ are unknown smooth functions with θ_k centered and scaled as follows:

$$E\{\int_0^\Delta X_k(t - \tau)\theta_k(\tau)d\tau\} = 0 \quad \text{and} \quad \int_0^\Delta \theta_k(\tau)d\tau = 1.$$

By assuming that $X_k(t)$ has a step sample path, they have further simplified the integral in (1.5) to

$$Y(t) = Z(t)^T\beta_0 + \sum_{k=1}^{q} g_k\bigg(\sum_{l=1}^{D} X_k(t - l)\theta_k(l)\bigg) + \epsilon(t),$$

where D is the largest lag considered. Daily average levels of NO_2, SO_2, PM_{10}, O_3, temperature and humidity were used as the explanatory variables, whereas the response $y(t_i)$ is defined as log (number of hospital admissions of patients suffering from respiratory diseases on day i). They have considered the pollution values of past 300 days on number of hospitalizations on the present day. They used smoothing splines to estimate both the unknown

link functions g_k's and weight functions θ_k's. Moreover, they compare the results obtained from the FACTS (1.5) model and the additive model (1.4) with lags. The closeness of the estimated response from the two approaches to the observed is well explained in Figure 1.3.

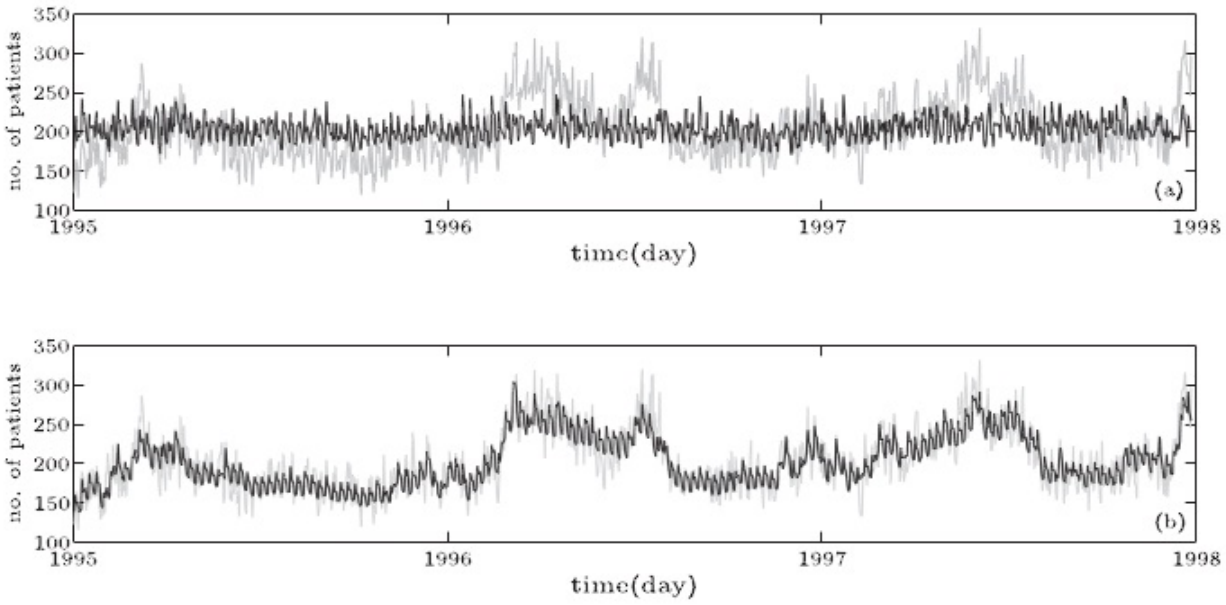


Figure 1.3: Fitted results of the additive model and FACTS model. In panel (a), the gray line is the observed numbers of hospital admissions and the black line is the fitted numbers by the additive model. In panel (b), the gray line is the observed number of hospital admissions and the black line the fitted numbers by the FACTS model. (Taken from Kong et al.(2010)[10])

In addition, all of these existing studies just described have used the daily average to summarize the hourly observations of weather and pollutant variables causing an information loss. Staniswalis et.al(2009) (closely related to our study) explains how a historical functional linear model [16](see Chapter 3 for details) can be used to utilize hourly measurements of pollutants. They illustrate the approach by using a continuous time lag to determine the association between $PM_{2.5}$ hourly levels and daily mortality. Daily mortal-

ity $y(t)$ is modeled by hourly measurements of $PM_{2.5}$, so that $y(t)$ depends only on the behavior of the hourly $PM_{2.5}$, $x(s)$ at past times $s \leq S(t)$; $S(t) = 24t$ was used to convert the daily index $t \in [1, ...T]$ of mortality to the hourly index s of $PM_{2.5}$. Daily mortality $y(t)$ is modeled as a Poisson distributed variable. Their model with the canonical link for the Poisson distribution is

$$\begin{aligned} \eta_t &= \alpha(t) + s_1(t) + \int_0^{min(S(t),S(\tau))} x(S(t)-u)\beta(u)du \\ &\quad + s_2(temperature_t) + \text{terms for day-of-week}_t, \quad \eta_t = ln\mu_t, \quad \mu_t = E[y(t)]. \end{aligned}$$

Here the parameter τ determines the maximum lag in days for which hourly $PM_{2.5}$ measurements are included as a predictor of daily mortality. The intercept function $\alpha(t)$ captures the seasonal changes over time t that influence the daily mortality. The nonparametric function $s_1(t)$ models the non-periodic changes over time t, s_2 models the dependence of mortality on the daily mean temperature and the last term represents indicator variables for day of the week. With this approach, they have been able to find a statistically significant association between morning $PM_{2.5}$ peak at 0.5-3 day lag and daily mortality; the 8 a.m. peak in $PM_{2.5}$ was most highly associated with mortality.

The main goal of our study is to understand the effect of dust and low wind events on hospitalizations for asthma while adjusting for hourly levels of air pollutants. In order to explore this, we utilize the patients' information on hospitalizations due to asthma that was collected by the Texas Health Care Information Council (THCIC), weather data (temperature, dew point, occurrence of dust events) collected at the El Paso International Airport (ELP) and pollution data ($PM_{2.5}$ levels, NO_2) collected by Texas Commission on Environmental Quality, for the time period 2000 through 2005. (see Chapter 6 for more details about data). We use the conditional logistic regression with a case-crossover design to estimate the probability of hospitalization after dust and low wind events while controlling

for pollutants with hourly monitor measurements, and weather. As explained earlier when modeling the relative risk of hospital admission for asthma due to changes in air pollutants, such as particulate matter, a lag-time must be specified between the exposure to the pollutant and the health effect. The historical functional linear model is used to incorporate the hourly pollutant measures into the regression model with a continuous lag, instead of fixed time-lag of daily averages. The historical functional linear model liberates us from specifying this lag-time, by allowing us to study how the exposure to $PM_{2.5}$ in the past days continuously influences the risk of hospitalization. This nonparametric functional linear model in the conditional logistic regression framework is fit by first preprocessing the data, then applying the COXPH function in the R-package for survival analysis with a slight modification. We use the ridge trace to guide the choice of the smoothing parameter in the nonparametric functional linear model.

In the remainder of this thesis we will explain in detail how we used the historical functional linear model within the conditional logistic regression framework in the context of the case-crossover design to achieve our goal. In chapter 2 we will describe how to use the P-splines for nonparametric regression. In chapter 3 we will define the historical functional linear model and use it to setup the model for daily asthma hospitalizations. In chapter 4 we will discuss some of the classical approaches namely case crossover study design and the conditional logistic regression model, ridge regression and use of ridge trace plot. In the last three chapters simulated examples, data analysis and discussion of statistical methods used for this study will be reported.

Chapter 2

Nonparametric Regression

2.1 From Parametric Regression to Nonparametric Regression

Regression analysis is used to determine the relationship between two or more random variables so that the value of the response variable can be predicted with the values of the explanatory variables. For illustration purposes, consider a simple linear regression model, with only one independent variable that has a linear relationship with the response. Suppose the response Y has been observed at the design points $t_1 < t_2 < ... < t_n$ following the regression model

$$y_i = f(t_i) + \epsilon_i, i = 1, ...n, \tag{2.1}$$

where $f(.)$ is the unknown regression function and $\epsilon_1, ..., \epsilon_n$ are unobservable random errors with zero mean and a constant variance σ^2.

2.1.1 Parametric (Classical) Way of Estimating $f(\cdot)$

The parametric approach to estimate the unknown function $f(.)$ is to first assume some functional form. In simple linear regression, we assume that $f(t) = \beta_0 + \beta_1 t$. Thus, the parametric approach assumes the simple linear regression function to be $y_i = \beta_0 + \beta_1 t_i + \epsilon_i$. Then the method of least squares can be used to estimate the regression parameters β_0 and β_1. For the observation pairs $\{(t_i, y_i)\}_{i=1}^n$, the method of least squares considers the

sum of n squared deviations, $Q = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$. Then β_0 and β_1 is estimated by minimizing Q given the sample. The minimum of Q is obtained at $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$, such that

$$\left.\frac{\partial Q}{\partial \beta_0}\right|_{\substack{\beta_0=\hat{\beta}_0 \\ \beta_1=\hat{\beta}_1}} = 0 \qquad \text{and} \qquad \left.\frac{\partial Q}{\partial \beta_1}\right|_{\substack{\beta_0=\hat{\beta}_0 \\ \beta_1=\hat{\beta}_1}} = 0.$$

Here,

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 t_i) \quad \text{and} \quad \frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 t_i)$$

The least squares estimators of β_0 and β_1 are

$$\hat{\beta}_1 = \frac{n\sum y_i t_i - \sum y_i \sum t_i}{n\sum t_i^2 - (\sum t_i)^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \beta_1 \bar{t}$$

But the above explained classical approach is effective only if $f(.)$ is approximately linear. For example consider the data in figure 2.1. When we fit a linear regression line for such data, the classical approach tends to give biased estimators with large standard errors producing wider confidence intervals for the regression coefficients. As a result we may end up with incorrect conclusions. So if we do not have enough confidence in the specified functional form of $f(.)$, it is better to use the nonparametric approach to estimate the unknown function.

2.1.2 Nonparametric Methods of Estimating $f(\cdot)$

Nonparametric regression analysis traces the dependence of a response variable on one or more predictor variables, without specifying any functional form to the relationship. It

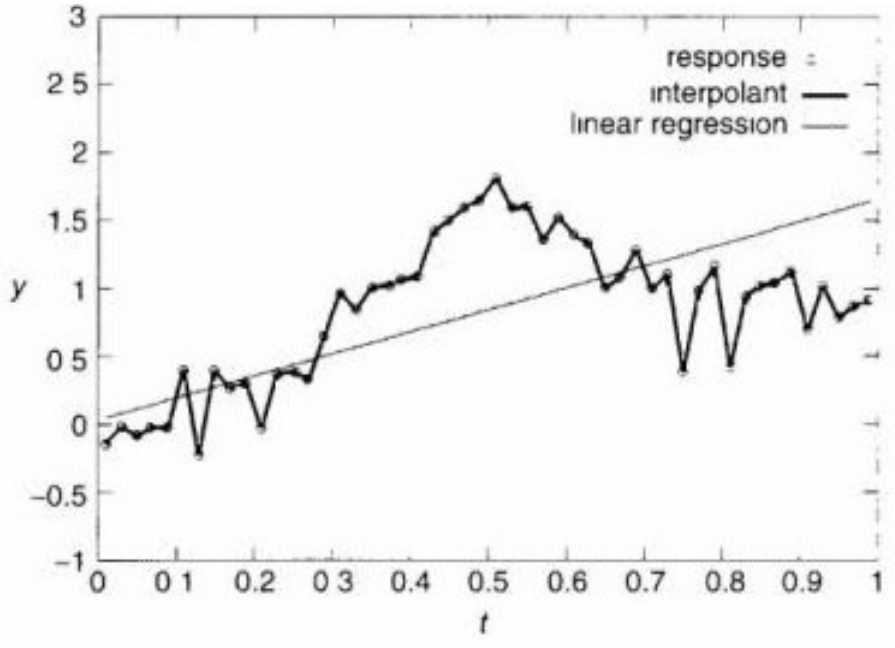


Figure 2.1: Linear regression and interpolation for a data set (Taken from Schimek (2000)[22]).

only assumes that $f(\cdot)$ is a smooth function with k (*usually k=2 or 4*) derivatives on the interval $[a, b]$. Estimation methods in nonparametric regression can be mainly classified as kernel based methods and spline methods. Figure 2.2 illustrates different nonparametric estimation methods and the types of basis functions used in each of those methods.

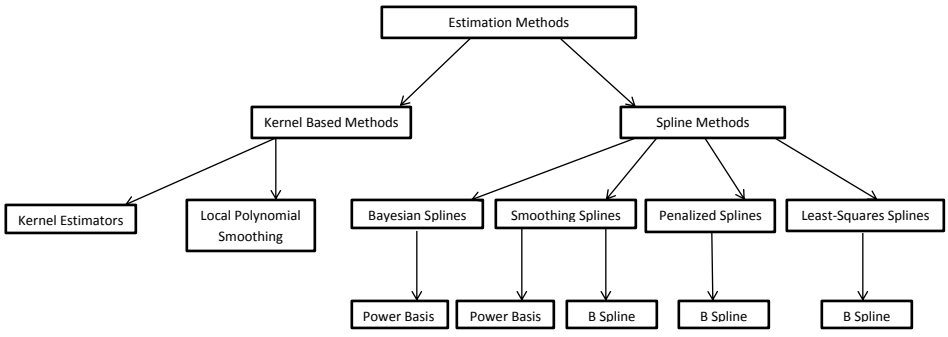


Figure 2.2: Flow chart of the nonparametric methods

2.2 Spline Methods

2.2.1 Smoothing Splines

In this section we discuss smoothing spline estimators for the regression curve $f(\cdot)$. Simply, the concept is to fit a regression function to a data set that reflects the key features of the data while retaining some degree of smoothness[5]. Suppose the response $y_1, y_2, ..., y_n$ have been obtained at the points $t_1 < t_2 < ... < t_n$, where $t_i \in [a, b]$ follows the regression model (2.1), where f is an unknown fixed function with m continuous derivatives that is square integrable. Then the natural measure of smoothness associated with the function $f \in W_2^m[a, b]$ (*Sobolev space of order m*) is given by, $\int_a^b (f^{(m)}(t))^2 dt$. The Sobolev space is defined as $W_2^m[a, b] = \{f | f^{(m)}$is continuous on [a,b] and square integrable$\}$. The goodness of fit is measured by the residual sums of squares $n^{-1} \sum_{i=1}^n (y_i - f(t_i))^2$. Then the overall measure of the goodness of g as the estimator of the function $f(\cdot)$ is given by the quantity called penalized sum of squares,

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - g(t_i))^2 + \lambda \int_a^b g^{(m)}(t)^2 dt \tag{2.2}$$

where,

- λ is the smoothing parameter.
- The first term measures the fidelity to the data.
- The second term (roughness penalty) penalizes for the curvature of the function g.

The unknown regression function $f(t)$ is estimated by minimizing the penalized sums of squares over all $g \in W_2^m[a, b]$. The solution to this minimization problem is a spline of order $2m - 1$. For illustration purposes, consider the case when m=2. Then the solution is a spline of order 3 (*cubic spline*). Schoemberg (1964) has shown that this is equivalent to minimizing $\frac{1}{n} \sum_{i=1}^n (y_i - f(t_i))^2$ subject to $\int_a^b f^{(2)}(t)^2 dt < \rho(\lambda)$ where $\rho(\lambda) \geq 0$. The

solution $\hat{f}(t)$ is a **natural cubic spline** with the knots being placed at the data points [5]. This solution is generally referred to as the smoothing spline estimator of $f(t)$. A natural cubic spline satisfies the following:

- Consists of piecewise cubic polynomials on $[t_i, t_{i+1}]$
- Has two continuous derivatives at the design points $t_1, ..., t_n$
- The third derivative has jumps at the $t_1, \ldots, t_n$
- Linear on $[a, t_1)$ and $(t_n, b]$

Note: Large values of λ give smoother curves while small values tend to track the true data points. As $\lambda \to \infty$, solution is the least squares line. As $\lambda \to 0$, the estimated curve interpolates the data. Figure 2.3 shows the effect of the smoothing parameter λ on the curve estimation.

Definition of the Spline of Order r

Let $t_1 < t_2 < ... < t_n$ be a set of ordered points called knots, in some interval (a, b). Then a spline of order r is any function of the form:

$$S(t) = \sum_{i=0}^{r-1} \theta_i t^i + \sum_{i=1}^{n} \delta_i (t - t_i)_+^{r-1}, \ t \in [a, b] \tag{2.3}$$

for some real valued coefficients $\theta_0, \theta_1, ..., \theta_{r-1}$ and $\delta_1, \delta_2, ..., \delta_n$. From the definition it can be verified that:

- $S(t)$ is a piecewise polynomial of order r (degree at most $(r-1)$) on any subinterval $[t_i, t_{i+1})$.
- $S(t)$ has $(r-2)$ continuous derivatives.
- $S^{(r-1)}(t)$ has jumps at the knots.

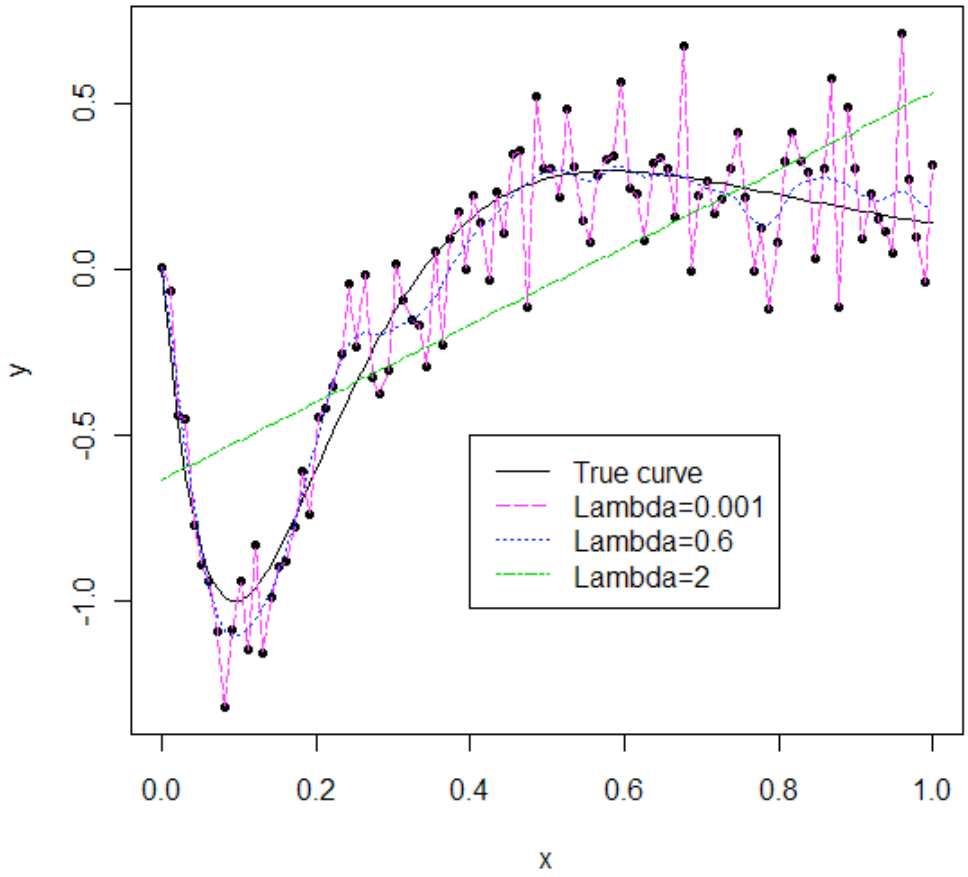


Figure 2.3: Smoothing spline fit for simulated data

Thus a spline is a piecewise polynomial with different polynomial segments combined together at knots $t_1, ..., t_n$, ensuring certain continuity properties. It can be shown that the set of all possible splines of order r denoted by $S^r(t_1, ..., t_n)$, is a vector space with dimension $(n + r)$. Moreover, from (2.3) we can see that the set of functions $\mathscr{T} = \{1, t, t^2, ..., t^{r-1}, (t - t_1)_+^{r-1}, (t - t_2)_+^{r-1}, ..., (t - t_n)_+^{r-1}\}$, are linearly independent and span the vector space of splines of order r. Hence, $\mathscr{T}$ becomes a basis set for the vector space $S^r(t_1, ..., t_n)$; $\mathscr{T}$ is called the *power basis*.

Natural Spline of Order r

Let $NS^r(t_1, ..., t_n)$ with $r = 2m$ denote the subspace of $S^r(t_1, ..., t_n)$ of all natural splines of order r. The spline $S(t)$ in (2.3) is called a *natural spline* if it is a polynomial of order m outside the interval $[t_1, t_n]$. This places restrictions on the m^{th} derivative of the spline for $t \in [a, t_1) \cup (t_n, b]$, and we must have $\theta_m = \theta_{m+1} = ...\theta_{2m-1} = 0$. With those $2m$ restrictions, one can show that $NS^r(t_1, ..., t_n)$ is of dimension n.

Note: When $r = 4$ an element in $S^r(t_1, ..., t_n)$ is called a cubic spline, whereas an element in $NS^r(t_1, ..., t_n)$ is called a natural cubic spline.

B-Spline Basis

In the previous section we discussed computing the smoothing spline estimator by using the truncated power basis. The power basis can be computationally inefficient [5] [3], since it needs quite a large number of operations to compute the smoothing spline estimator. So it is desirable to explore the use of other basis functions that could reduce the number of operations needed to compute the smoothing spline estimator. B-splines are such a set of basis functions. The definition of the B-spline basis is given here by using a recursion formula.

To develop the B-spline basis for $S^r(t_1, ...t_n)$ first we have to define $2r$ additional knots $t_{-(r-1)}, ..., t_{-1}, t_0, t_{n+1}, ..., t_{n+r}$ by letting,

$$t_{-(r-1)} = ... = t_0 = a \quad \text{and} \quad t_{n+1} = ... = t_{n+r} = b$$

Then the recursion is initialized by:

$$N_{i,1}(t) = \begin{cases} 1 & t \in [t_i, t_{i+1}] \\ 0 & \text{otherwise} \end{cases}$$

The B-spline of order r with knots at $t_1, ...t_n$ can be defined recursively by

$$N_{i,r}(t) = \frac{t - t_i}{t_{i+r-1} - t_i} N_{i,r-1}(t) + \frac{t_{i+r} - t}{t_{i+r} - t_{i+1}} N_{i+1,r-1}(t)$$

In particular we note that $N_{i,r}(t)$ has compact support as $N_{i,r}(t) = 0$ if $t \notin [t_i, t_{i+r}]$. Linear, quadratic and cubic B-spline basis obtained from the above recursion formula with interior knots being placed at $0.25, 0.5$ and 0.75 are shown figure 2.4. We can clearly observe that the number of B-spline functions is equal to the order of B-splines plus the number of interior knots. Also we can see that the local support of B-spline functions increases with the order. As a result, the overlapping support of adjacent B-splines also increases with the order of the B-spline functions.

2.2.2 Penalized Splines (P-Splines)

We already discussed the drawback of the truncated power basis and the use of the B-spline basis as an alternative for implementing the smoothing spline estimator. In section 2.2.1 we explained the role of the roughness penalty given by $\lambda \int_a^b f^{(2)}(t)^2 dt$ in controlling the curvature of the estimated function. But depending on the nature of the estimated function, sometimes it is difficult to compute the roughness penalty. To overcome this problem Eilers and Marx (1996) introduced a new concept to approximate the roughness penalty. The basic idea was to impose the penalty on the coefficients of the adjacent B-splines instead of using the integral of the squared second derivative of the estimated function over the interval of interest.

Earlier in this chapter it was explained how that the smoothing spline estimator of the

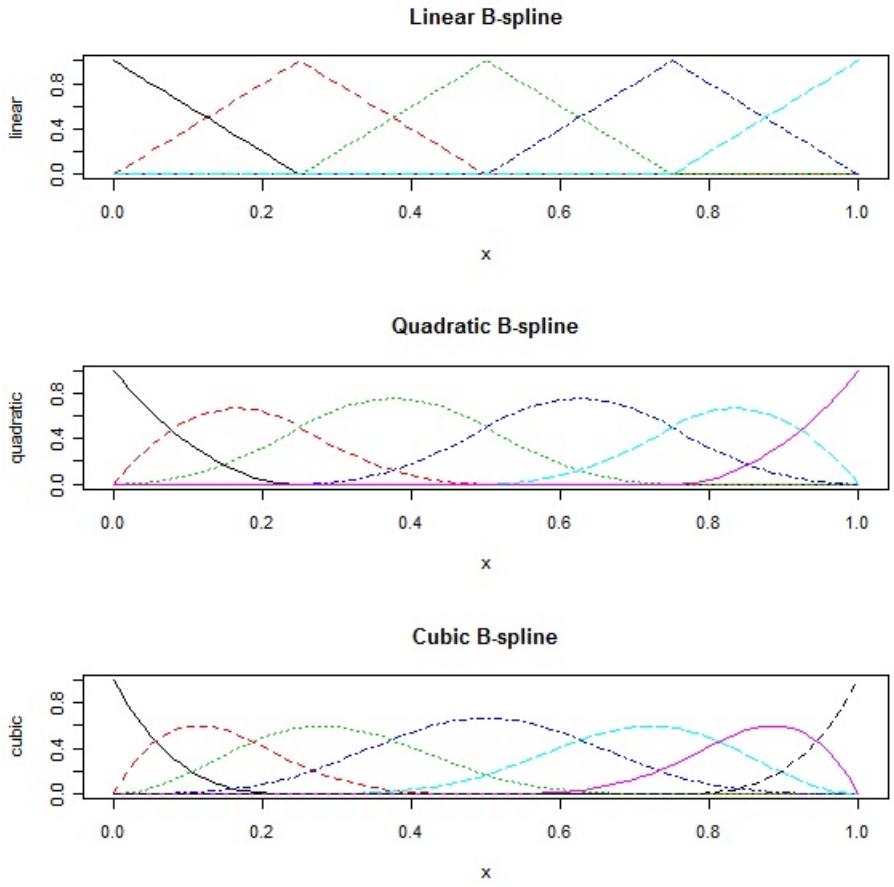


Figure 2.4: Linear, Quadratic and Cubic B-spline functions

unknown function $f(\cdot)$ is a linear combination of the truncated power basis. Since the B-splines were introduced as an alternative set of basis functions to the truncated power basis, we are able to express the fitted curve in terms of B-splines. Let $N_{j,r}(t)$ be the value at t of the j^{th} B-spline of order r. Then the fitted regression function $f(t)$ to the data (t_i, y_i) can be written as a linear combination of the B-spline functions.

$$f(t) = \sum_{j=1}^{K} \beta_j N_{j,r}(t) \tag{2.4}$$

where $K = n + r$ is the dimension of B-spline basis.Therefore, the least squares objective

function to be minimized is;

$$\sum_{i=1}^{n}\{y_i - \sum_{j=1}^{K}\beta_j N_{j,r}(t_i)\}^2.$$

By substituting the expression for f in (2.4), the penalized sum of squares in (2.2) can be rewritten as

$$S = \sum_{i=1}^{n}\{y_i - \sum_{j=1}^{K}\beta_j N_{j,r}(t_i)\}^2 + \lambda \int_a^b \{\sum_{j=1}^{K}\beta_j N''_{j,r}(t)\}^2 dt. \tag{2.5}$$

Eiler and Marx (1996) explained that there is a very strong connection between a penalty on second order differences of the B-spline coefficients and the penalty on the squared second derivative of the fitted function. Using cubic B-splines they have shown

$$\int_a^b \{\sum_{j=1}^{K}\beta_j N''_{j,4}(t)\}^2 dt \approx c_1 \sum_j (\Delta^2\beta_j)^2 + c_2 \sum_j \Delta^2\beta_j \Delta^2\beta_{j-1}$$

where,

$$\Delta\beta_j = \beta_j - \beta_{j-1}$$

$$\Delta^2\beta_j = \Delta(\Delta\beta_j) = \beta_j - 2\beta_{j-1} + \beta_{j-2}$$

$$c_1 = \int_a^b N^2_{j,2}(t)dt;$$

$$c_2 = \int_a^b N_{j,2}(t)N_{j-1,2}(t)dt;$$

It should be noted that the above formula was developed using cubic B-splines and rapidly growing complications in the penalty equation should be expected for the higher degrees of B-splines. To overcome this problem, Eiler and Marx (1996)[4] defined a difference penalty on coefficients of the adjacent B-splines. Instead of the penalty used in (2.5), they used

$$\lambda[\beta^T(D_2^T D_2)\beta]$$

where D_2 is the matrix representation of the difference operator Δ^2 and β is the vector of β_j's

$$D_2 = \begin{bmatrix} 1 & -2 & 1 & \cdots & 0 \\ 0 & 1 & -2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \end{bmatrix}_{(K-2)\times K} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}_{K\times 1}.$$

With this penalty, the penalized sum of squares can be written in matrix notation as

$$(Y - B\beta)^T(Y - B\beta) + \lambda[\beta^T(D_2^T D_2)\beta]$$

where B is the B-spline matrix with elements $(B)_{ij} = N_{j,4}(t_i)$. Minimizing the penalized sum of squares we get

$$\hat{\beta} = (B^T B + \lambda D_2^T D_2)^{-1} B^T Y.$$

Note that $\hat{\beta}$ looks like the results of a ridge regression except that λI has been replaced $\lambda D_2^T D_2$. In chapter 4 we will discuss the ridge regression as a sub-section under classical approaches that will be used in this thesis. In the next chapter we will discuss the historical functional linear model and its estimation with P-splines.

Chapter 3

Historical Functional Linear Model

3.1 Functional Data and FDA

In this chapter we will discuss the historical functional linear model, which refers to a special case of functional linear model when the influence of the covariates on the response is of a feed forward nature. First we will explain the nature of functional data through an example given in Ramsay and Silverman (2005)[20]. The data were originally taken from the "Berkeley Growth Study" done in 1954 by R.D Tuddenham and M.M Snyder. It has heights of 10 girls measured at 31 ages between 1 to 18 years old. Four measurements were taken from each child when they were one year old. Annual measurements were taken from 2 to 8 years and finally from ages 9 through 18, measurements were taken twice a year. In this data even though the heights were obtained at certain discrete points it reflects a smooth variation during the follow up period. Thus, each of 10 records is a height function or in other words it is a smooth curve. So the functional data analysis (FDA) is all about the analysis of information in functions or curves. Figure 3.1 is a plot of the growth data taken from Ramsay and Silverman (2005).

With some understanding about the nature of functional data, let's consider $PM_{2.5}$ data observed for the period 2000 through 2005, that is been used in this study. The data were not continuously sampled over time, but were sampled on an hourly grid. Specifically the available data consists of discrete measurements $\{x_i(t_j)\}_{j=1}^{24}$, where i is the index representing the day. As the first step of the functional data analysis, the discrete measurements need to be converted into a continuous function $x_i(t)$ over time, so that the value at any

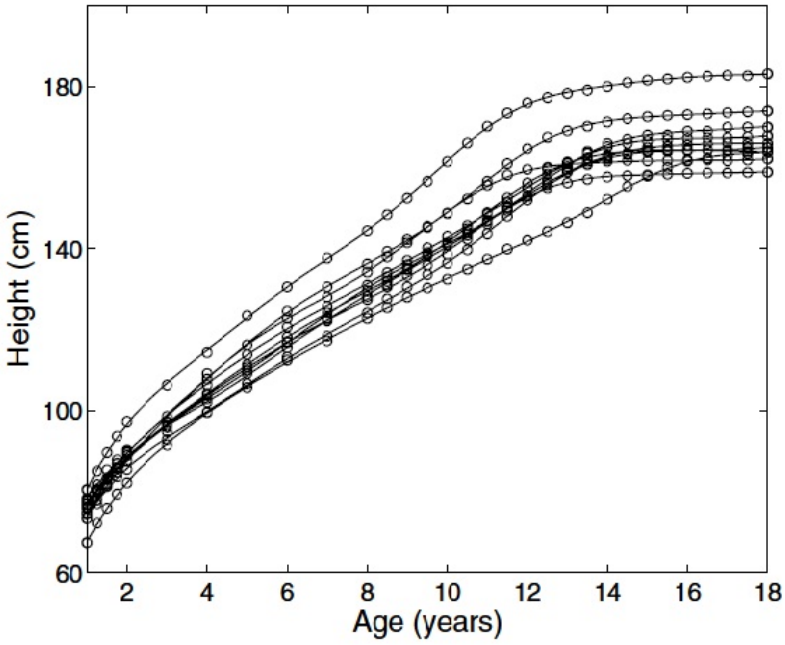


Figure 3.1: The heights of 10 girls measured at 31 ages. The circles indicate the unequally spaced ages of measurement. (Taken from Ramsay and Silverman(2005)).

desired point t within the study period can be obtained. If the observed discrete data is assumed to be errorless, then this process of conversion is called *simple interpolation*: joining each pair of adjacent points by a straight line, but if the data have some noise that needs to be filtered out then the conversion process should involve some smoothing[20]. Either simple interpolation or smooth interpolation allow us to impute the missing observations in the data set, which is very important to this study, since some of the hourly measurements of $PM_{2.5}$ and NO_2 were missing in the original (raw) data. For instance, if day t have $5^{th}, 6^{th}$ and 7^{th} hourly measurements missing and given that rest are nonmissing, then we may predict those missing observations by connecting all the available measurements of that day with a smooth function. Practically it is preferable to use smooth interpolations with basis functions rather than just joining adjacent points with straight lines. Even though great care was taken during the process of data collection, still there might be an

uncertainty or noise in the observed measurements. In smooth interpolations, functional data for the i^{th} day of the observed data is represented as a linear combination of a set of basis functions, $x_i(t) = \sum_{k=1}^{K} c_{ij}\phi_j(t)$. Here, ϕ_j is a set of basis functions and c_{ij} are some real valued coefficients. Fourier basis, B spline basis, wavelets, exponential and power basis and polynomial basis are some of the most common bases that are being used in functional data analysis. The choice of basis set should be done with caution. Because the basis functions should have the features that matches those of the functions being estimated. For example, the Fourier basis system is usually used if the observed data has a periodic nature and B spline basis is used for non-periodic data. In this study we smoothed the hourly measurements of $PM_{2.5}$ and NO_2 with B spline basis. The characteristics of the B-spline basis and the implementation of it was discussed in Chapter 2.

3.1.1 Functional Linear Models

In classical statistics, general linear models such as linear regression and analysis of variance are used to explain the variation in one variable (dependent variable) using one or a set of other variables (covariates). Functional linear models serve the same purpose, but in the context of functional data. Linear models can be functional in either or both of two ways explained below.

- Response variable y with argument t is functional.
- One or more of the covariates are functional.

In this thesis, we focus only on the second case where one or more of the covariates have a functional form with a scalar response. For illustration purposes, we use a functional linear model with only one covariate being functional. But it should be noticed that one can extend this by including as many functional covariates as are needed. Consider the linear regression line with usual matrix notation,

$$Y = X\beta + \epsilon. \tag{3.1}$$

Now, suppose the prediction of scalar values y is based on the function $x(\cdot)$. Then the functional extension to the linear regression model can be written as

$$Y = \alpha + \int_{t_0}^{t_1} x(s)\beta(s)ds + \epsilon. \tag{3.2}$$

Note that the summation given by the matrix multiplication $X\beta$ in (3.1) is being replaced by a definite integral over a continuous index s in (3.2). To give more concrete explanation, let's consider the Canadian weather station example given in Ramsay and Silverman (2005). The original data set has daily temperature and precipitation data for 35 Canadian weather stations. Suppose we want to predict the annual precipitation of a weather station from the pattern of temperature variation throughout the year. Let y_i be the logarithm of total annual precipitation for the i^{th} weather station and $x_i = Temp_i$ be the daily temperature function of the i^{th} weather station. Then the functional linear model for this particular application can be written as

$$\text{Log(Tot.Preci)}_i = \alpha + \int_0^T Temp_i(s)\beta(s)ds + \epsilon.$$

3.2 Historical Functional Linear Model

In this section we consider a special case of functional linear models, where the influence of the covariates on the response is of a feed-forward nature. This means the response at time t, $y(t)$ depends only on the behavior of covariate $x(s)$ at past times $s \leq t$. Consider a functional regression model in which a function $x_i(s)$ with a hourly index s, $s \leq S(t)$ is used as the independent variable to explain the variation in the response $y_i(t)$ with daily index t, $t \in \{1, ...T\}$, where $i = 1, ...N$. Here $S(t) = 24t$ is used to convert the daily index t in the response to hourly index s of the covariate. Then the historical functional linear model can be written as

$$\eta_t = \alpha(t) + \int_0^{S(t)} x(S(t) - s)\beta(s)ds, \quad \eta_t = g[\mu_t], \quad \mu_t = E[Y(t)] \tag{3.3}$$

where g is the link function determined from the distribution of $Y|X$. For example, for the Poisson $g(\mu) = log(\mu)$; for the binomial $g(\mu) = logit(\mu)$. The intercept function $\alpha(t)$ is the overall summary of the influence of unspecified covariates on the response at day t. The slope function $\beta(s)$ reveals the strength of the association between x at lag s hours and the response y_t.

3.2.1 Estimating the Slope Function $\beta(\cdot)$

Here the method of estimation is discussed step by step. It should be noted that we only focus on estimating β. Because in this study we use the historical functional linear model in the context of conditional logistic regression. In the conditional logistic regression model, the regression parameters are estimated based on a likelihood function conditional on a sufficient statistic for $\alpha(\cdot)$. So the nuisance parameter $\alpha(t)$ is ignored without being estimated (see Chapter 4 for more details). Consider the model in (3.3). Let $\{N_j(t), j = 1, ...K\}$ be cubic B-spline basis for a given knot sequence on $[0, S(t)]$. The slope function is approximated by

$$\beta(u) = \sum_{j=1}^{K} b_j N_j(u)$$

Once the expression for the $\beta(u)$ is substituted in (3.3) we get

$$\begin{aligned}
\eta_t = \int_0^{S(t)} x(S(t)-u)\beta(u)du &= \int_0^{S(t)} x(S(t)-u)\Big\{\sum_{j=1}^{K} b_j N_j(u)\Big\}du \\
&= \sum_{j=1}^{K} b_j \Big\{\int_0^{S(t)} N_j(u)x(S(t)-u)du\Big\} \\
&= \sum_{j=1}^{K} b_j \phi_j(t;x). \qquad (3.4)
\end{aligned}$$

Then $\phi_j(t;x)$ is approximated by a quadrature rule defined below. Let $x_+(S(t)-u)$ be a function such that;

$$x_+(S(t)-u) = \begin{cases} x(S(t)-u) & \text{when } u \le S(t) \\ 0 & \text{otherwise} \end{cases}$$

where $u \in [0, 24T]$ and $t \in [0, T]$. Then $\phi_j(t;x)$ can be rewritten as;

$$\begin{aligned}
\phi_j(t;x) &= \int_0^{S(t)} x(S(t)-u)N_j(u)du \\
&= \int_0^{24T} x_+(S(t)-u)N_j(u)du.
\end{aligned}$$

Then approximate the integral in $\phi_j(t;x)$ with a quadrature rule using Q equally spaced time points $\{g_q\}_{q=1}^{Q}$ in [0,24T]:

$$\phi_j(t;x) = \frac{24T}{Q}\sum_{q=1}^{Q} x_+(S(t)-g_q)N_j(g_q)$$

We can easily derive matrix notation for η_t, by letting $X'(t) = (x_+(S(t)-g_1), ...x_+(S(t)-$

$g_Q))$, $h = \frac{24T}{Q}$ and

$$N_b = \begin{bmatrix} N_1(g_1) & N_2(g_1) \cdots & N_K(g_1) \\ N_1(g_2) & N_2(g_2) \cdots & N_K(g_2) \\ \vdots & \vdots & \vdots \\ N_1(g_Q) & N_2(g_Q) \cdots & N_K(g_Q) \end{bmatrix}_{Q \times K} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}_{K \times 1}.$$

Then (3.4) can be written in matrix form:

$$g(t) = hX'(t)N_b b.$$

Once the $\phi_j(t; x)$ are approximated by a Riemann sum, one can use standard statistical packages like R or SAS to estimate the b_j's in (3.4). We refer to this approximation process as *preprocessing of the data*. In the next chapter we will discuss some of the classical regression approaches that used in this study.

Chapter 4

Classical Modeling Approaches

4.1 Case-Crossover Study Design and Conditional Logistic Regression Model

We use conditional logistic regression with a case-crossover design to estimate the probability of hospitalization for asthma after a dust event. It is desired for one to be well-informed regarding the nature of the case-crossover study design and the conditional logistic regression framework, before understanding how we exploited existing statistical packages. Therefore, first we'll take a look at the characteristics of case-crossover study design and then we will explain maximum likelihood estimation of a conditional logistic regression model applied to a case-crossover study design.

4.1.1 Case-Crossover Study Design

Before understanding the case-crossover study design, one should become familiar with the case-control study design, because case-crossover design is a special case of case-control study design. Case control studies are widely used in epidemiological studies to investigate the cause of disease. In these types of studies individuals having a particular disease (cases) are compared with the individuals who do not have the disease (controls) [23]. Once the cases and the controls are identified it looks back in time to see what characteristics of cases are differ from the controls. The purpose of these studies is to identify the degree of association between the risk of disease and the exposure factors under study [18]. The case control studies are appropriate for studying rare diseases because it starts with a set of

individuals who have the disease, rather than starting with a population free of disease and looking for the development of the disease. This makes it easy to find enough patients with a rare disease. Because of that the case control studies are considered to be a quick, easy and inexpensive method of finding risk factors for diseases compared to other methods [12]. Usually in case control studies, the cases and the controls having similar characteristics, such as age, gender and socioeconomic status matched to lessen the confounding effects. A *confounding variable* refers to a variable associated with the outcome as well as associated with the other exposure factors in the study [12]. In the case-control study design setup, confounding variables are ideally identified and used as *matching variables* or *stratum variables*. Matching cases and controls increases the efficiency of the study. Matched case control studies can be either $1:1, 1:M$ or $M:N$. If one case is matched with a one control within each stratum, then it is called a 1:1(one to one) matched case control study. Similarly $1:M$ (one to many) refers to a matching of one case with M controls in each stratum, and $M:N$ (many to many) refers to a matching of M cases to N controls. In practice, the most common is the $1:M$ matched case-control study design.

Recall that, a control is an individual free of disease and similar in characteristics to the case. In practice, it may be difficult to find individuals who have similar characteristics as the cases. One approach would be to choose relatives, friends and neighbors of cases as the control groups, which is considered to be a good method for controlling for the possible differences in socioeconomic and education status. Another method to eliminate possible confounding effects is by having a case serve as their own control. The method was first developed by Maclure (1991)[13], as a hybrid of the case-control design in which controls are sampled from the cases' own history, and the cross-over design in which the exposure is switched according to the interest of the study hypothesis [18]. Over the past decade, the case-crossover study design design has become a popular tool in pollution studies [7], [18], [1]. In these studies the exposure of an individual to a pollutant immediately prior to the health outcome (asthma/bronchitis, acute coronary syndrome) is compared with

the exposure of the same individual to the same pollutant at control times. Therefore, in case-crossover design we don't have to match the cases and controls by age, gender or any other confounding variables, since the controls were drawn from the same individuals that are considered as cases. For instance in our study, if Tom is going to the hospital on day t, then we define his controls, 7 days before and after the hospitalization. This method of selecting controls is called symmetric bi-directional reference selection. More specifically, we are using not only 7 days before and after the hospitalization as our controls, but also 14, 21, and 28 days before and after the hospitalization as well. When selecting controls for a particular patient, we restricted them to fall within the same month and year of his/her hospitalization. This restricted control selection procedure can be considered more suitable, because the weather in El Paso varies greatly from month to month, and also it controls for the day of the week.

4.1.2 Logistic Regression Model and Conditional Logistic Regression Model Applied for Case-Crossover Studies

The conditional logistic regression model is similar to the logistic regression model except that it takes into account the matching in the analysis of the observed data in the process of estimating the parameters [9]. The general method of parameter estimation for the logistic regression model is the maximum likelihood approach. The method of maximum likelihood (ML) yields values of the unknown parameters which maximize the probability of obtaining the observed set of data.

Let x be an indicator variable representing the exposure status to a given pollutant ($dust$). Then the logistic regression model can be written in general as

$$logit[Pr(Y_{ik} = 1)] = \alpha_i + \beta x_{ik}, \quad k = 1, 2, \tag{4.1}$$

where

- $Pr(Y_{ik} = 1)$ denotes the probability that the i^{th} person under the the k^{th} exposure status is having the outcome of interest (disease).
- Here α_i is the i^{th} individual effect (stratum effect).
- β is the increase of log odds of the disease for the exposure group compared to the unexposed group.

The likelihood function to be maximized can be written as

$$\mathcal{L}(\alpha_i, \beta) = \prod_{i=1}^{n} \pi_{ik}^{Y_{ik}} (1 - \pi_{ik})^{1-Y_{ik}}$$

where

$$\pi_{ik} = P(Y_{ik} = 1) = \frac{e^{\alpha_i + \beta x_{ik}}}{(1 + e^{\alpha_i + \beta x_{ik}})}.$$

The ordinary maximum likelihood estimators of the logistic regression model gives the best estimators when the sample size n is large compared to the numbers of parameters to be estimated. When the sample size is not large compared to the number of parameters, the ML approach might result inconsistent parameter estimates. This situation typically arises when data are stratified and the intercepts for each stratum are fixed effects. For instance if we have n strata, then $n+1$ parameters (α_i's and β) in model (4.1) need to be estimated. In such a situation, we can avoid estimating a large number of parameters by fitting the *conditional logistic regression model* instead of logistic regression model. The conditional logistic regression model is also written as (4.1). Even though two models are the same, the likelihood functions used to estimate the parameters are different. The conditional likelihood function allows us to get rid of the nuisance stratum parameters (α_i) in the estimation. Here, the rationale is to find sufficient statistics for α_i's and use a likelihood

function conditioned on those sufficient statistics, where the sufficient statistic for α_i is the stratum total and the total number of cases observed[9]. So, once we condition on the stratum total and the total number of cases observed, according to sufficiency theory, the likelihood function will not depend on the α_i's and will depend only on the slope parameters. Therefore the number of parameters we need to estimate will be limited to the number of slope parameters.

Let x_{it} be an exposure series defined at times $t = 1, 2, ...T$ common to all $i = 1, 2, ...n$ subjects, t_i and W_{t_i} be the index time and referent window for the i^{th} subject respectively. The referent window includes the index time t_i (time at which the case is observed) and referent times (times at which controls are observed). Here, x_{ij} is the exposure of the i^{th} person at time j. Suppose each stratum has 1 case and m controls. Then there are $\binom{m+1}{1}$ possible ways of assigning case status among $m + 1$ subjects in the the stratum.

Let

$$\rho_1 = P(S_i = 1|Y_i = 1) \quad \text{and} \quad \rho_0 = P(S_i = 1|Y_i = 0)$$

where s_i is defined as

$$S_i = \begin{cases} 1 & \text{if the } i^{th} \text{ subject is selected in the sample} \\ 0 & \text{if the } i^{th} \text{ subject is not selected in the sample} \end{cases}$$

The likelihood function for the i^{th} stratum can be expressed as [9];

$$\begin{aligned} l_i(\beta) &= \frac{P(x_{i1}|Y_1 = 1) \prod_{j \in W_{t_i}, j \neq 1} P(x_{ij}|Y_j = 0)}{\sum_{j_0 \in W_{t_i}} [P(x_{ij_0}|Y_{j_0} = 1) \prod_{j \in W_{t_i}, j \neq j_0} P(x_{ij}|Y_j = 0)]} \\ &= \frac{\pi^*(x_{i1}) \prod_{j \in W_{t_i}, j \neq 1} [1 - (x_{ij})]}{\sum_{j_0 \in W_{t_i}} [\pi^*(x_{ij_0}) \prod_{j \in W_{t_i}, j \neq j_0} [1 - \pi^*(x_{ij})]]} \end{aligned} \tag{4.2}$$

where

$$\pi^*(x_{ij}) = \frac{e^{\alpha_i^* + \beta x_{ij}}}{1 + e^{\alpha_i^* + \beta x_{ij}}} \quad \text{and} \quad \alpha_i^* = ln(\rho_1/\rho_0) + \alpha_i.$$

Replacing the expression for $\pi^*(x_{ij})$ in (4.2) we get

$$\begin{aligned} l_i(\beta) &= \frac{e^{\alpha_i^* + \beta x_{i1}}}{\sum_{j_0 \in W_{t_i}} (e^{\alpha_i^* + \beta x_{ij_0}})} \\ &= \frac{e^{\beta x_{i1}}}{\sum_{j_0 \in W_{t_i}} e^{\beta x_{ij_o}}}. \end{aligned}$$

The full conditional likelihood function of the case-crossover design is

$$l(\beta) = \prod_{i=1}^{n} l_i(\beta)$$

4.2 Ridge Regression and Ridge Trace

Recall that, in Chapter 2 we described the relationship between the ridge regression and the estimated coefficients of the adjacent B-splines in smoothing spline estimator. Based on this theoretical relationship we modify the *ridge* function provided in *coxph* of the R package to add the penalty needed for the historical functional linear model. Also we use the ridge trace to guide the choice the of smoothing parameter in the historical functional linear model. In this section we discuss the concept of ridge regression and the use of ridge trace in the context of classical regression.

4.2.1 Ridge Regression

It is well known that in multiple linear regression the method of least squares provides minimum variance unbiased estimates for the regression coefficients. It is also known that collinearity among the predictor variables (multicollinearity) can severely reduce the precision of least squares estimates of the regression coefficients[14]. More specifically, in the

presence of multicollinearity, the model parameters become unstable with large variances and complicate the interpretations of the regression coefficients. Ridge regression is one of the methods that have been proposed to reduce the variance of estimated regression coefficients, while introducing some bias. When an estimator has small bias, but is more precise than the unbiased estimator, then it may be the preferred estimator, since it has a larger probability of being closer to the true value of the parameter. Figure 4.1 taken from Kutner et al.(2005)[11] best explains this situation. By looking at the two sampling distributions, we can see that the probability of b_R (biased estimator) falling near the true β is greater that that of b (unbiased estimator).

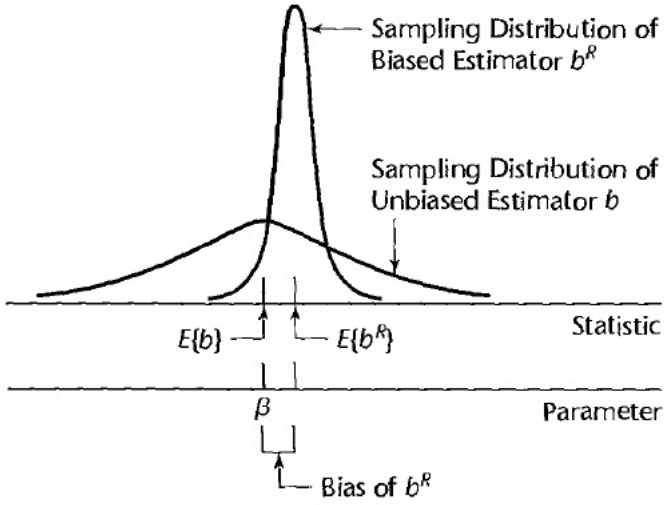


Figure 4.1: Biased estimators with small variance may be preferable to unbiased estimator with large variance (Taken from Kutner et al.(2005)[11])

For the purpose of illustration, consider the multiple linear regression model,

$$Y = X\beta + \epsilon$$

where Y is the $n \times 1$ vector of standardized independent variable, X represents the p

standardized explanatory variables and ϵ is the vector of random error. Recall that the least squares estimator of the regression coefficient vector β is given by;

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{4.3}$$

From a mathematical standpoint, collinearity among predictor variables makes the $X'X$ matrix near singular: the value of it's determinant is nearly zero. The ridge regression was developed to overcome this singularity by adding a small amount to the diagonal of $X'X$. This modification makes the determinant of $X'X$ different from zero. With this adjustment for the singularity, the ridge estimator β^* is defined as;

$$\hat{\beta^*} = (X'X + kI)^{-1}X'Y \tag{4.4}$$

where k is a small positive value. The $\hat{\beta^*}$ minimizes $\sum(Y_i - \beta^T X_i)^2 + k\beta^T I^T I\beta$. By assuming X is fixed and substituting $X\beta + \epsilon$ for Y we get the expected value of β^*

$$\begin{aligned} \mathrm{E}(\hat{\beta^*}) &= (X'X + kI)^{-1}(X'X)\beta \text{ since } E(\epsilon) = 0 \\ &= (X'X + kI)^{-1}(X'X + kI - kI)\beta \\ &= \beta - k(X'X + kI)^{-1}\beta. \end{aligned}$$

So the bias $(\hat{\beta^*}) = -k(X'X + kI)^{-1}$. If we assume independent and uncorrelated error terms, then

$$\mathrm{Var}(\hat{\beta^*}) = \mathrm{Var}[(X'X + kI)^{-1}X'Y] = \sigma^2(X'X + kI)^{-1}X'X(X'X + kI)^{-1}$$

4.2.2 Ridge Trace

The constant k (biasing constant) in (4.4) reflects the amount of bias in the estimator. When $k = 0$ it reduces to the unbiased least squares estimator in (4.3). The mean square error (MSE) of an estimator is equal to the variance of the estimator plus the squared bias of that estimator. It can be shown that the bias component in MSE$(\hat{\beta^*})$ increases in k

(with all regression coefficients in vector $\hat{\beta}^*$ tending towards zero) while variance component becomes smaller as k increases[11]. But there exists some value k for which the MSE of $\hat{\beta}^*$ is smaller than the mean squared error of the least squares estimator of β. It is difficult to find the optimal value of k since the MSE cannot be calculated practically. Ridge trace is one of the commonly used methods for estimating the optimal value of k. The ridge trace is a plot of estimated ridge coefficients of the regression model with variables transformed with the correlation transformation verses k(usually between 0 and 1). Here, the correlation transformation refers to a simple modification of a usual standardization of a variable. More specifically, it is a simple function of the standardized variables given by, $Y_i^* = \frac{1}{\sqrt{n-1}}(\frac{Y_i - \bar{Y}}{s_Y})$ and $X_{ik}^* = \frac{1}{\sqrt{n-1}}(\frac{X_{ik} - \bar{X_k}}{s_k})$ with s_Y and s_k are the respective standard deviations defined as usual. These estimated ridge coefficients may vary drastically as k is changed slightly from zero. As k increases further, the estimated regression coefficients stabilize and move towards zero. The ridge trace is used to detect the smallest value of k where the the regression coefficients first stabilize. Thus, the choice of optimal value of biasing constant k based on ridge trace is quite subjective. Figure 4.2 is a picture of a ridge trace.

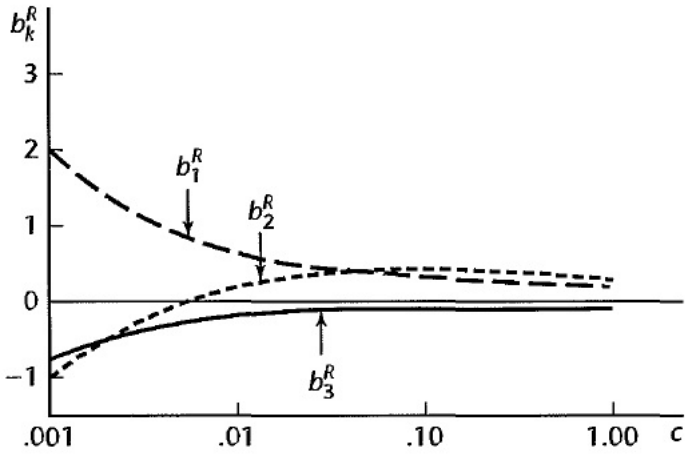


Figure 4.2: Ridge Trace (Taken from Kutner et al.(2005)[11])

Having the theoretical background required for this study, in the next chapter we will discuss some simulated examples to explore how to use the ridge trace concept to choose the smoothing parameter in a nonparametric regression model.

Chapter 5

Simulated Examples

5.1 Adapt the Ridge Trace to Choose the Smoothing Parameter in a Nonparametric Regression Model

We propose to use the ridge trace used in ridge regression to guide the choice of smoothing parameter in the nonparametric functional linear model. In chapters 2-4 we studied the smoothing spline estimator and came to understand the P-spline as a special case of ridge regression. In this chapter we will discuss some simulated examples to explain how we adapted the ridge trace to choose the smoothing parameter of the P-spline estimator in the context of semi-parametric regression.

Consider the regression model with $Y|X_1, X_2$ as the response and ϵ the normally distributed random error with zero mean and variance σ^2:

$$y_i = \theta x_{i1} + g(x_{2i}) + \epsilon_i, \quad i = 1, 2, ...n, \tag{5.1}$$

where

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim MVN\Big(\mathbf{0}, \mathbf{\Sigma}\Big), \tag{5.2}$$

and $g(\cdot)$ is a unknown but smooth function. Suppose $g(x_2)$ is approximated by $g(x_2) = \sum_{j=1}^{K} \beta_j N_{j,4}(x)$, where $N_{j,4}(x)$ is the j^{th} cubic B-spline function evaluated at x and β_j's are the corresponding coefficients. Then the regression model (5.1) can be written in matrix

notation as

$$Y = \theta X_1 + B\beta + \epsilon$$

where B is the B-spline design matrix and β is the corresponding coefficient vector. The penalized sum of squares to be minimized is

$$Q(\theta, \beta) = (Y - \theta X_1 - B\beta)^T (Y - \theta X_1 - B\beta) + \lambda[\beta^T (D_2^T D_2)\beta].$$

This is minimized using the profile likelihood approach. First, fix β and solve for θ_β. The partial derivative of Q with respect to θ is

$$\frac{\partial Q}{\partial \theta} = 2X_1'(Y - \theta X_1 - B\beta)$$

By letting $\frac{\partial Q}{\partial \theta} = 0$, we get

$$\theta_\beta = (X_1' X_1)^{-1} X_1'(Y - B\beta)$$

Second, consider the profile likelihood $PL(\beta) = Q(\theta_\beta, \beta)$ where,

$$Q(\theta_\beta, \beta) = (Y - \theta_\beta X_1 - B\beta)^T (Y - \theta_\beta X_1 - B\beta) + \lambda[\beta^T (D_2^T D_2)\beta].$$

By setting $\frac{\partial PL(\beta)}{\partial \beta} = 0$, we can derive an expression for the $\hat{\beta}$:

$$\hat{\beta} = [(B'B + \lambda D_2' D_2) - (B' X_1 (X_1' X_1)^{-1} X_1' B)]^{-1} B' [I - X_1 (X_1' X_1)^{-1} X_1'] Y.$$

The minimizer of $Q(\theta, \beta)$ is given by $(\theta_{\hat{\beta}}, \hat{\beta})$.

Let $A = [(B'B+\lambda D_2'D_2)-(B'X_1(X_1'X_1)^{-1}X_1'B)]^{-1}B'[I-X_1(X_1'X_1)^{-1}X_1']$. Then the above expression for $\hat{\beta}$ can be rewritten as

$$\hat{\beta} = AY.$$

Then $\mathrm{E}[\hat{\beta}]$ and $\mathrm{Var}[\hat{\beta}]$ are given by

$$\mathrm{E}[\hat{\beta}] = A(\theta X_1 + g(x_2)) \quad \text{and} \quad \mathrm{Var}[\hat{\beta}] = \sigma^2 AA'$$

Our objective is to find a suitable smoothing parameter (λ) in the P-spline estimator of $g(x_2)$. Usually, the value of λ is chosen to minimize the averaged mean squared error (AMSE) of the estimator. The mean squared error (MSE) of an estimator is given by squared bias of the estimator plus variance of the estimator. Thus, the $MSE[\hat{g}(x_2)]$ is given by

$$\mathrm{MSE}[\hat{g}(x_2)] = \mathrm{Bias}^2[\hat{g}(x_2)] + \mathrm{Var}[\hat{g}(x_2)]$$

where

$$\mathrm{Bias}^2[\hat{g}(x_2)] = BE[\hat{\beta}] - g(x_2) \quad \text{and} \quad \mathrm{Var}[\hat{g}(x_2)] = \sigma^2 BAA'B',$$

and the averaged mean squared error, AMSE is given by

$$\mathrm{AMSE}[\hat{g}(x_2)] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}[\hat{g}(x_{2i})].$$

Note that B is the B-spline design matrix evaluated at $x = x_2$. For an example, we will examine the optimal value of λ (value at which the AMSE is minimum) on the adapted

ridge trace plot, in order to determine whether visual inspection of the adapted ridge trace can help find a suitable value of the smoothing parameter. Define

$$t(x) = \frac{x - min(x)}{max(x) - min(x)} \qquad \text{and} \qquad f(x) = 4.26(e^{-3.25x} - 4e^{-6.50x} + 3e^{-9.75x}).$$

We simulated data from the regression model

$$y_i = \theta x_{i1} + g(x_2 i) + \epsilon_i, \quad i = 1, 2, ...n$$

in which $g(x_2) = 50f \circ t(x_2)$, with $\theta = 1$, $n = 50$ and $\epsilon \sim N(0, 400)$. Here $f \circ t(\cdot)$ is the usual composition of two functions f and t.

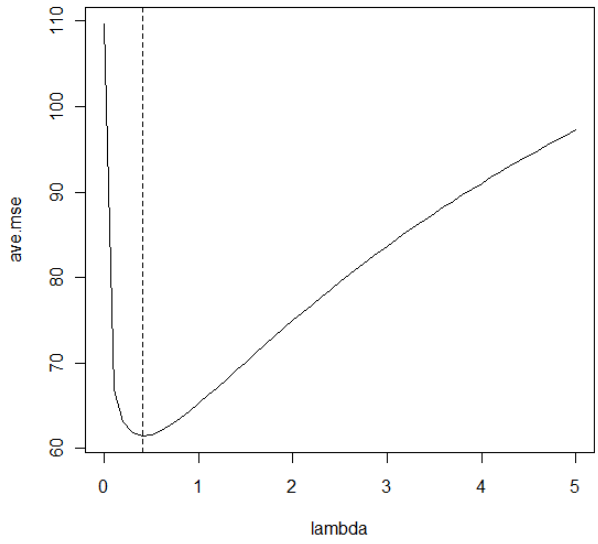
Example 1

The $\{(\ X_{1i},\ X_{2i})\}_{i=1}^{n}$ were generated from a multivariate normal distribution with correlation $\rho = 0.8$:

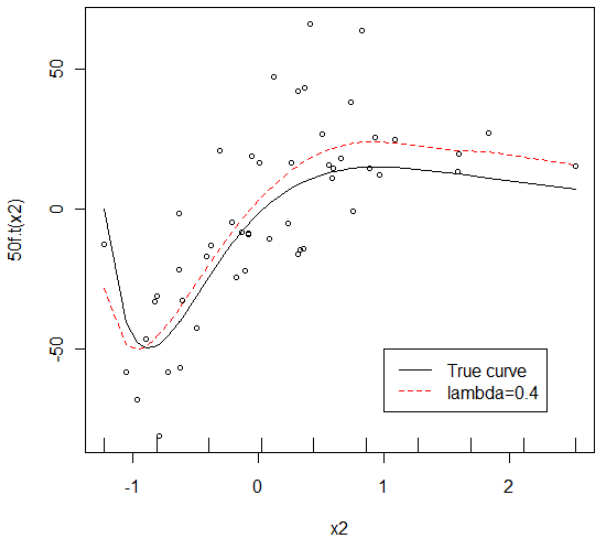
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 625 & 20 \\ 20 & 1 \end{bmatrix} \right).$$

We plot the average mean squared error (AMSE) of $g(x_2)$ against the λ. The optimal value of λ where the AMSE is minimum is at 0.4. Figure 5.1(a) shows the behavior of AMSE as λ increases and figure 5.1(b) is the true and the estimated curve with the optimal λ.

We suggest visual examination of a plot of $\hat{\beta}/SE(\hat{\beta})$ (standardized coefficients) verses λ for selection of a suitable value λ for smoothing data, where $\hat{\beta}$ represents an estimate of a coefficient of the B-spline design matrix and the denominator is the corresponding standard error. For instance, in this example we used cubic B-splines with 8 interior knots. Thus, we have 12 (order plus no. of interior knots) B-spline functions with a corresponding



(a) AMSE as a function of λ



(b) True and estimated curve with the optimal λ. Points in the graph denote Y_i's. The placement of knots were indicated by inner grids.

Figure 5.1: Fitted function with the value of λ in at which AMSE is minimum (Example 1).

coefficient for each. We use the name *adapted ridge trace* for the plot of $\hat{\beta}/SE(\hat{\beta})$ verses λ. The adapted ridge trace was plotted for each of the 50 data sets to get more insight into the behavior of the standardized coefficients as λ increases. Figures 5.2, 5.3 and 5.4 show the variation of the each standardized coefficients across the 50 trials.

Note that even though we varied λ from 0 to 5, when plotting the standardized β coefficients we restrict the x scale of the plots from 0 to 1, so that the behavior of the coefficients around the optimal λ (at 0.4) can be seen clearly. Recall that, in ridge regression we choose the optimal biasing constant using the ridge trace when the coefficients first start to stabilize. Similarly, by looking at the adapted ridge trace plots for example 1, we can roughly say that the standardized coefficients of B-spline functions start to stabilize around 0.4, which is the value of the optimal λ in this case. Before confidently suggesting that the adapted ridge trace plot can be used to capture a neighborhood of the optimal λ, we explored how sensitive this is to the correlation among X_1 and X_2, by changing the variance covariance matrix of X_1 and X_2.

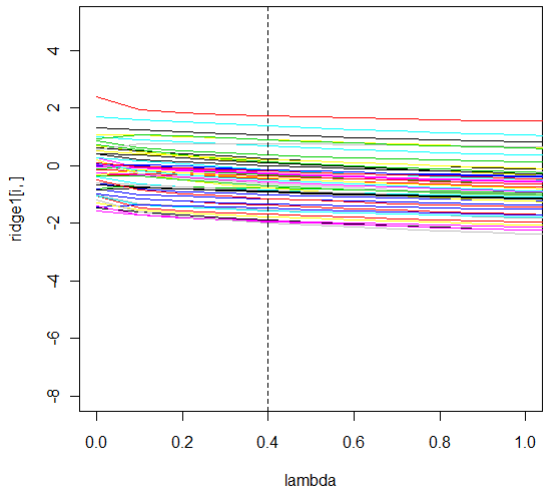
Example 2

Next we generate X_1 and X_2 from a multivariate normal distribution so that the correlation is $\rho = 0.9$, while everything else remained the same

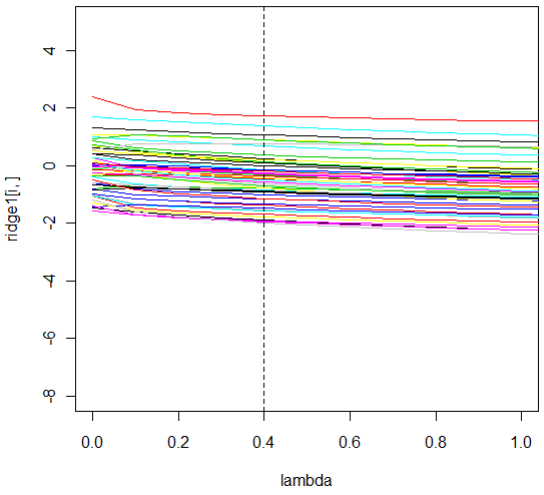
$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 400 & 180 \\ 180 & 100 \end{bmatrix} \right).$$

The AMSE was still minimum at $\lambda = 0.4$; the plot of AMSE verses λ and the fitted function with $\lambda = 0.4$ is shown in figure 5.5. Similar to example 1, we plot the adapted ridge trace for 50 data sets and the corresponding plots of standardized β coefficients verses λ is given in figures 5.6, 5.7 and 5.8.

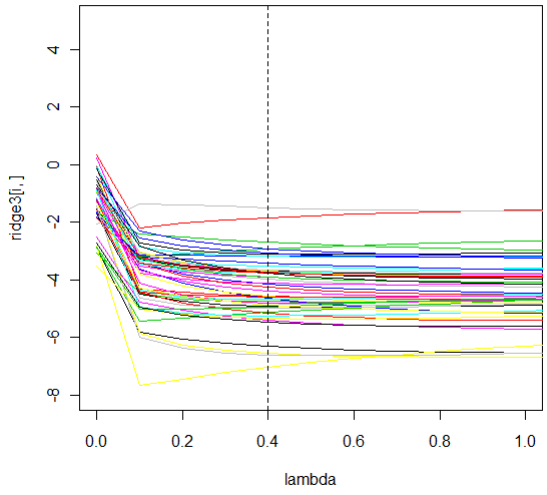
Even for this example we can see that the standardized β coefficients start stabilizing



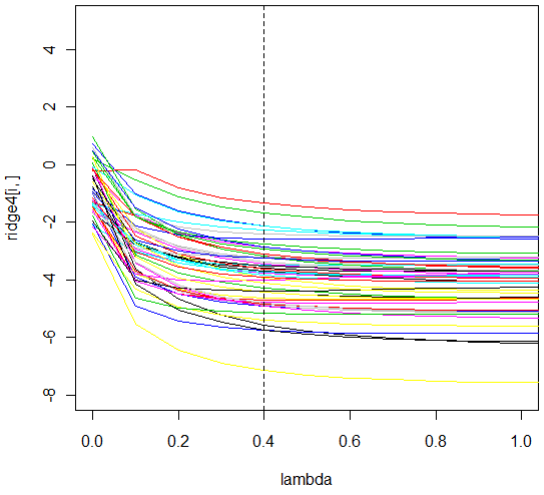
(a) Standardized β_1 across the 50 trials



(b) Standardized β_2 across the 50 trials

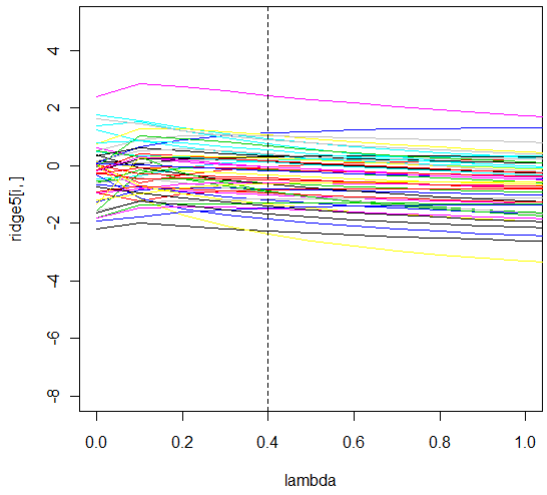


(c) Standardized β_3 across the 50 trials

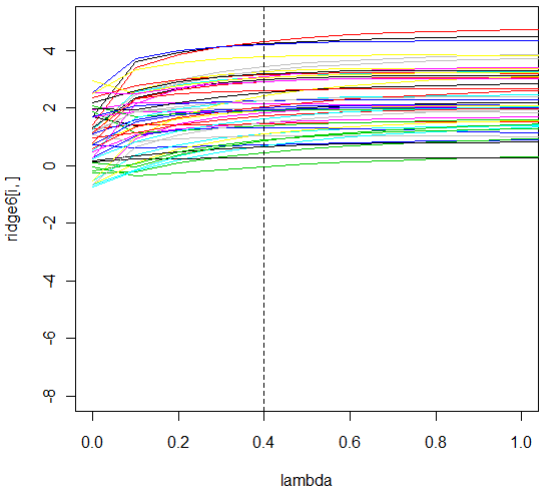


(d) Standardized β_4 across the 50 trials

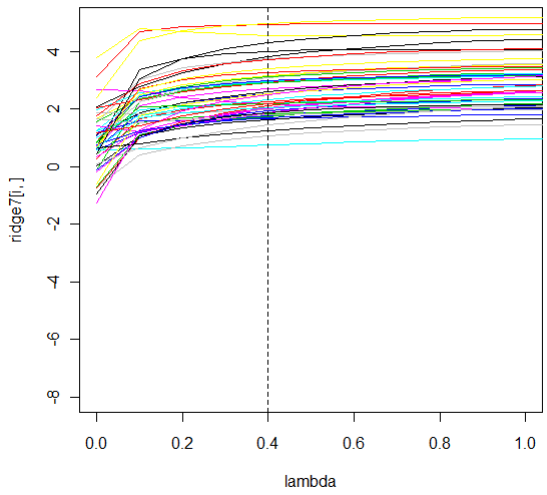
Figure 5.2: First four standardized β coefficients across the 50 trials (Example 1).



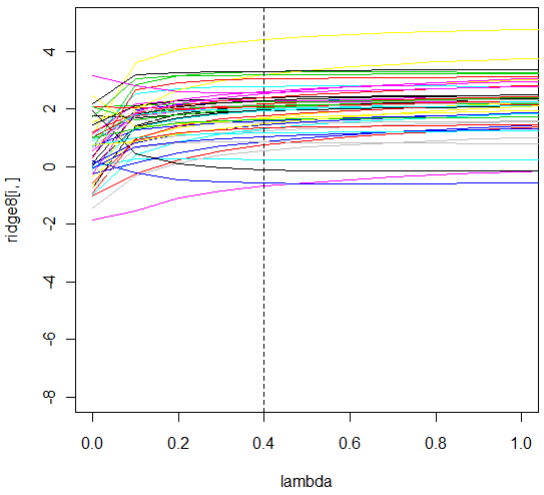
(a) Standardized β_5 across the 50 trials



(b) Standardized β_6 across the 50 trials

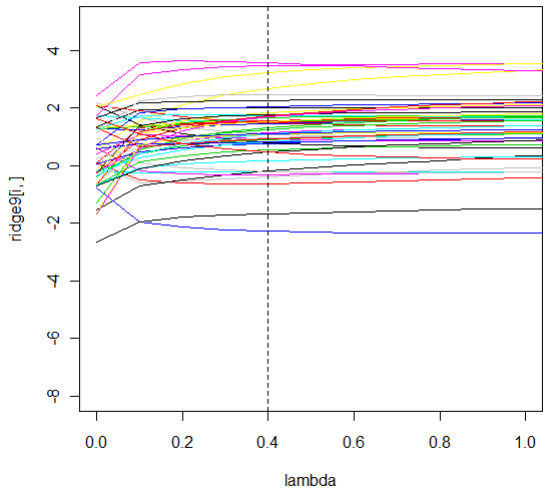


(c) Standardized β_7 across the 50 trials

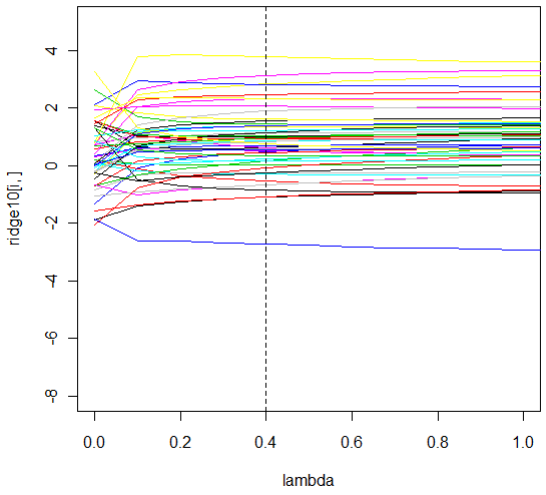


(d) Standardized β_8 across the 50 trials

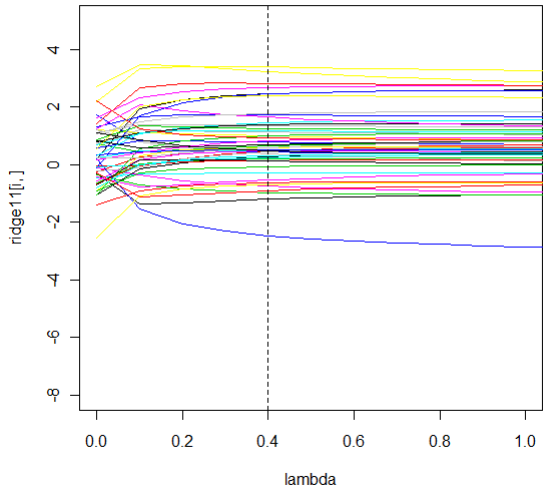
Figure 5.3: Second four standardized β coefficients across the 50 trials (Example 1).



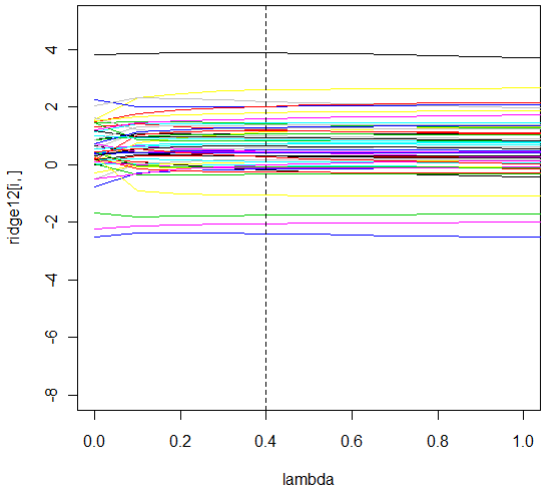
(a) Standardized β_9 across the 50 trials



(b) Standardized β_{10} across the 50 trials



(c) Standardized β_{11} across the 50 trials



(d) Standardized β_{12} across the 50 trials

Figure 5.4: Last four standardized β coefficients across the 50 trials (Example 1).

around the optimal λ, which is 0.4. In summary, we can observe that when X_1 and X_2 are highly correlated the adapted ridge trace identifies a neighborhood of the optimal value of λ. Next we will examine the behavior of adapted ridge trace plot when the correlation among X_1 and X_2 is comparatively low.

Example 3

For this example, X_1 and X_2 were generated from a multivariate normal distribution with $\rho = 0.5$, while everything else remained the same

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 64 & 20 \\ 20 & 25 \end{bmatrix} \right).$$

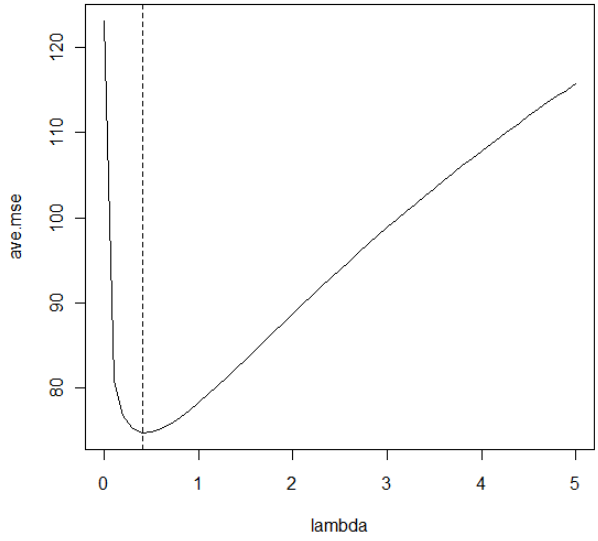
The AMSE is minimum at $\lambda = 0.5$. The plot of AMSE verses λ and the fitted function with $\lambda = 0.5$ is shown in figure 5.9. Corresponding plots of standardized β coefficients verses λ across 50 trials are shown in figures 5.10, 5.11 and 5.12. Still we can see that the standardized β coefficients start stabilizing in the neighborhood of the optimal λ.

Example 4

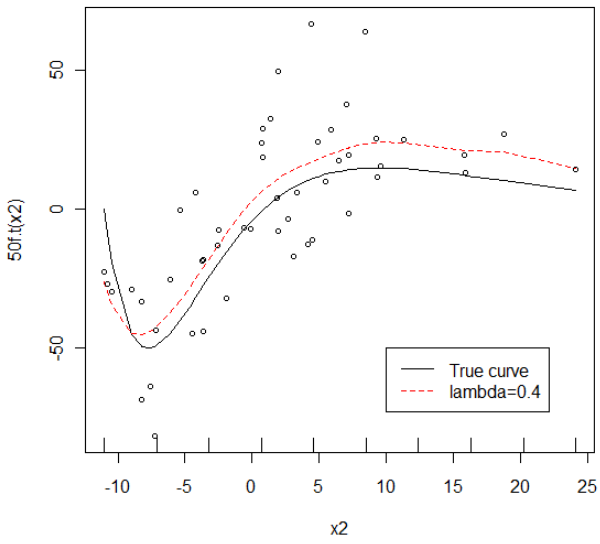
For this example we used even a smaller correlation. X_1 and X_2 were generated from a multivariate normal distribution with $\rho = 0.3$, while everything else remain the same.

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 25 & 6 \\ 6 & 16 \end{bmatrix} \right).$$

The AMSE is still minimum at $\lambda = 0.5$. The plot of AMSE verses λ and the fitted function with $\lambda = 0.5$ is shown in figure 5.13. Corresponding plots of standardized β coefficients verses λ across 50 trials are shown in figures 5.14, 5.15 and 5.16. Even for this example we can see that the standardized β coefficients start stabilizing around the optimal λ, which

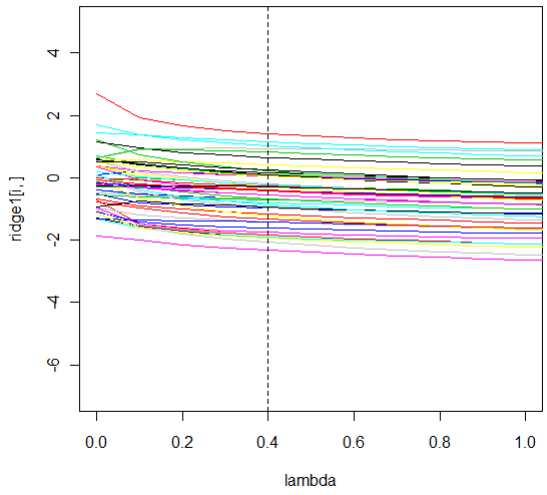


(a) AMSE as a function of λ

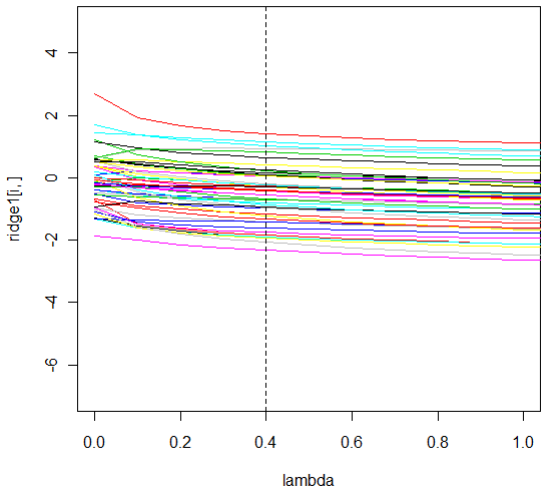


(b) True and estimated curve with the optimal λ. Points in the graph denote Y_i's. The placement of knots were indicated by inner grids.

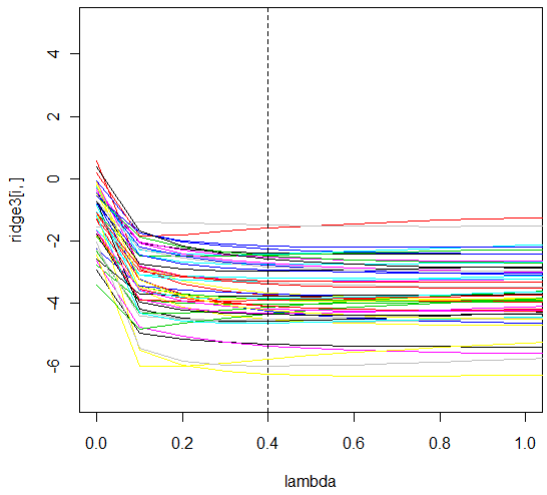
Figure 5.5: Fitted function with the value of λ in at which AMSE is minimum (Example 2).



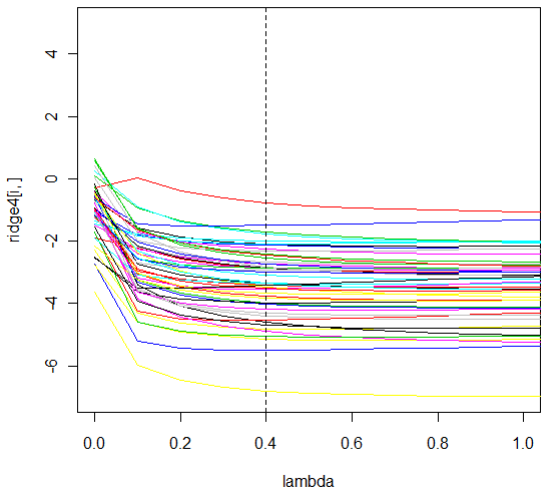
(a) Standardized β_1 across the 50 trials



(b) Standardized β_2 across the 50 trials

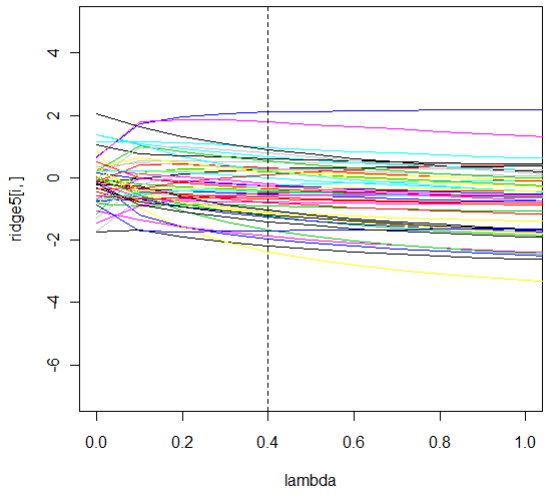


(c) Standardized β_3 across the 50 trials

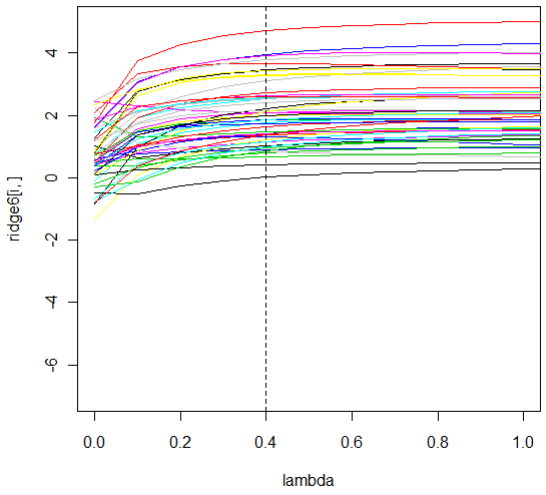


(d) Standardized β_4 across the 50 trials

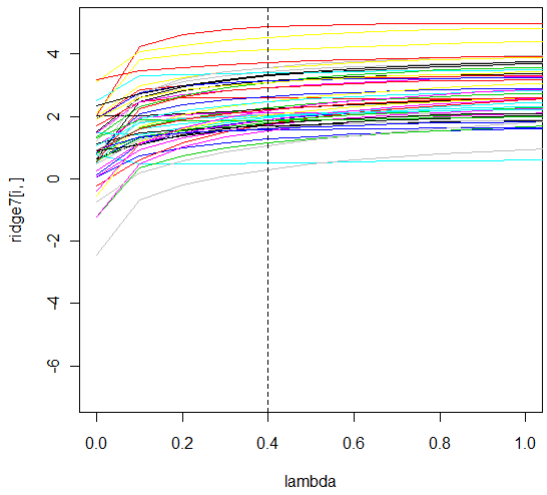
Figure 5.6: First four standardized β coefficients across the 50 trials (Example 2).



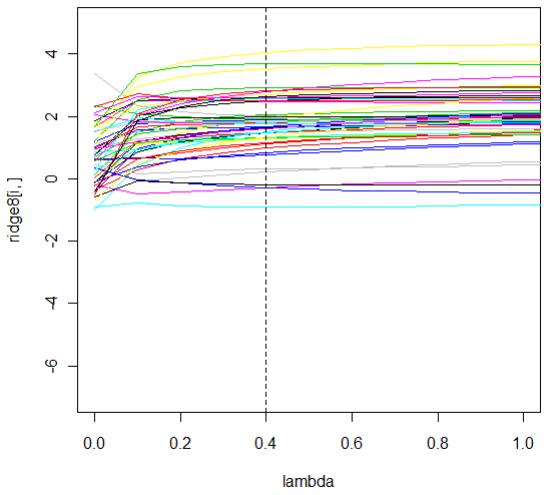
(a) Standardized β_5 across the 50 trials



(b) Standardized β_6 across the 50 trials

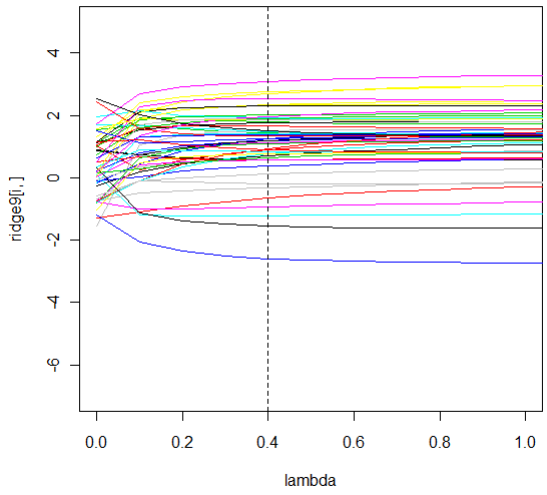


(c) Standardized β_7 across the 50 trials

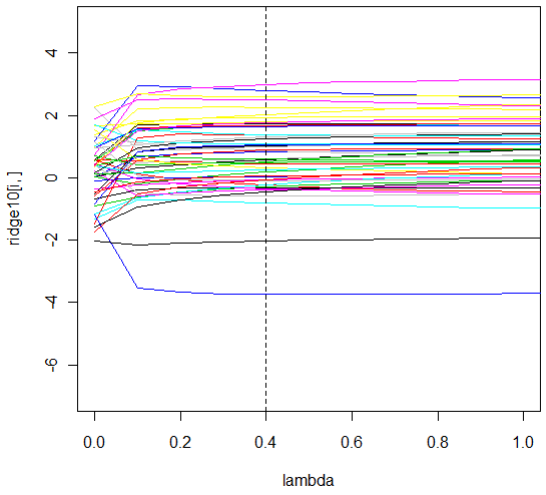


(d) Standardized β_8 across the 50 trials

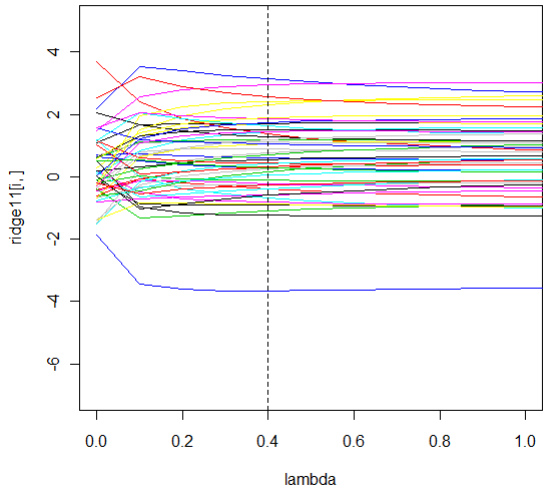
Figure 5.7: Second four standardized β coefficients across the 50 trials (Example 2).



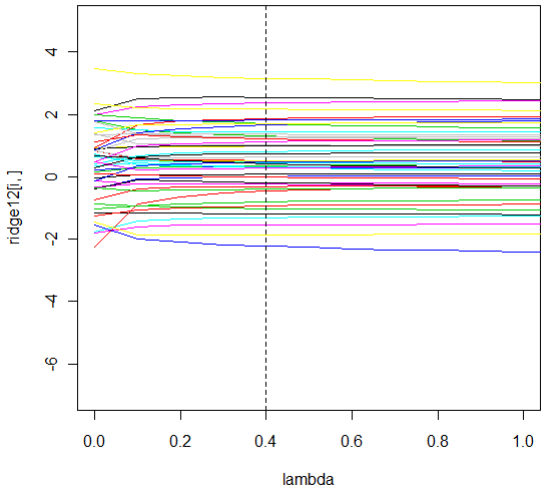
(a) Standardized β_9 across the 50 trials



(b) Standardized β_{10} across the 50 trials



(c) Standardized β_{11} across the 50 trials

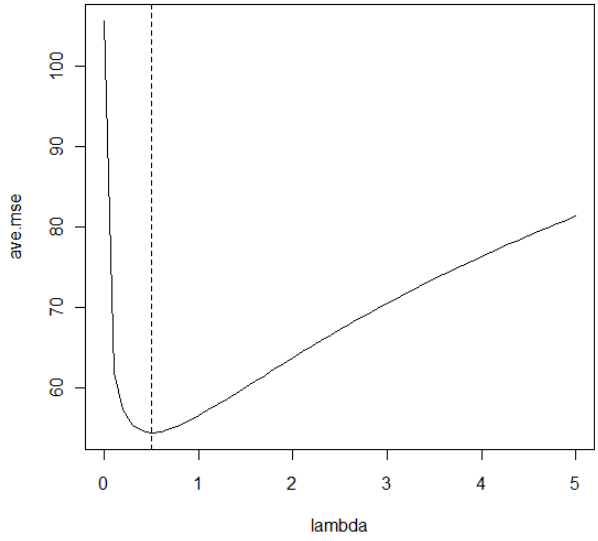


(d) Standardized β_{12} across the 50 trials

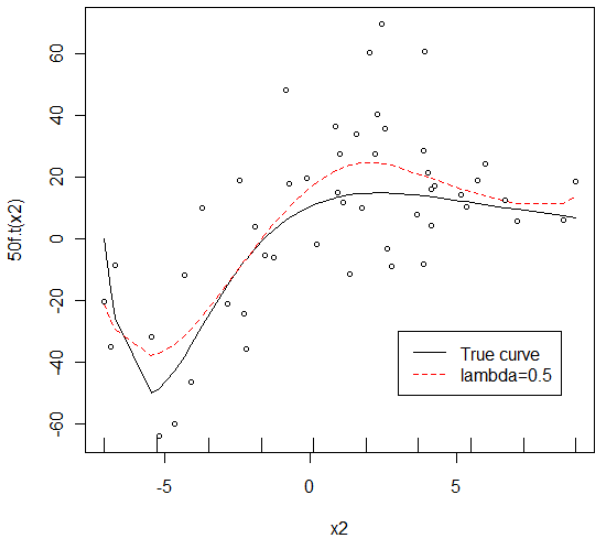
Figure 5.8: Last four standardized β coefficients across the 50 trials (Example 2).

is 0.5. From example 3 and 4, we can see that even for small correlations between X_1 and X_2 , the adapted ridge trace roughly captures the optimal value of λ.

By looking at the adapted ridge trace plots that we obtained for all four examples, we suggest that it can be used to identify the optimal λ that minimizes the AMSE of a P-spline estimator of a regression function. Also, the correlation among two variables does not seem to affect the usefulness of the adapted ridge trace plot for these purposes. This may not be a perfect way to identify the suitable smoothing parameter, but practically MSE can not be calculated and common suggested methods to approximate the MSE are complicated. Even with those complicated methods we can not find the exact value of λ that minimizes the AMSE of a regression function. Also it make sense to use a plot of standardized coefficients verses λ to capture the smoothing parameter because of the similarity between the estimated coefficients of B-spline functions of the smoothing spline estimator and the ridge regression. With the support of all these arguments we suggest visual examination of the adapted ridge trace plot to identify a suitable smoothing parameter for the P-spline estimator.

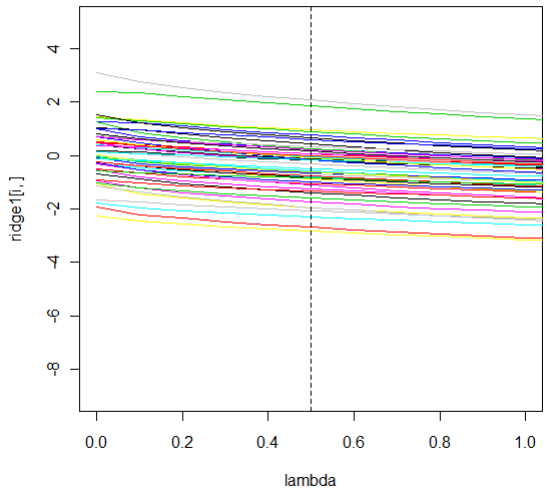


(a) AMSE as a function of λ

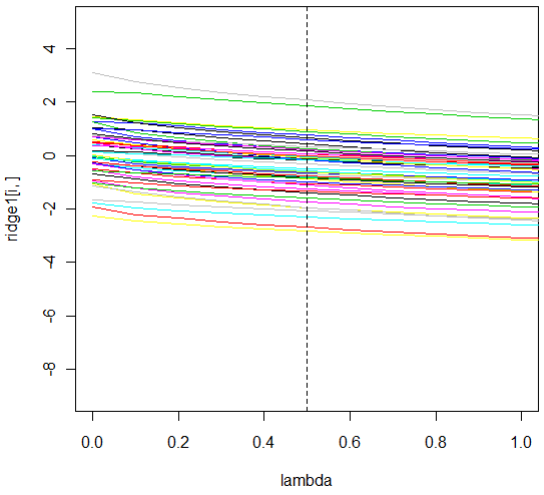


(b) True and estimated curve with the optimal λ. Points in the graph denote Y_i's. The placement of knots were indicated by inner grids.

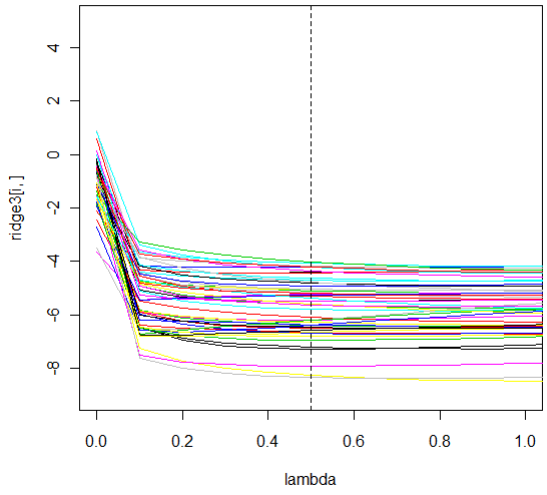
Figure 5.9: Fitted function with the value of λ in at which AMSE is minimum (Example 3).



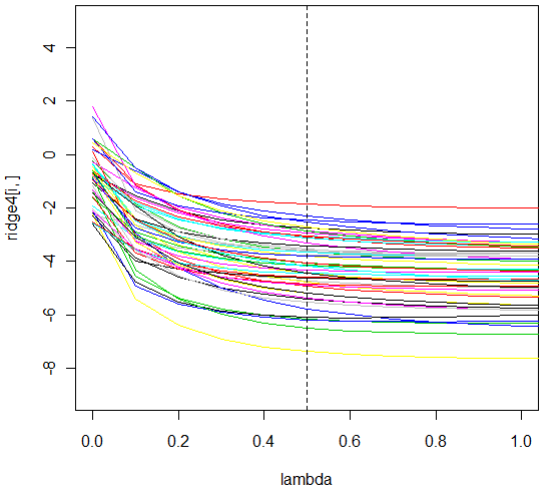
(a) Standardized β_1 across the 50 trials



(b) Standardized β_2 across the 50 trials

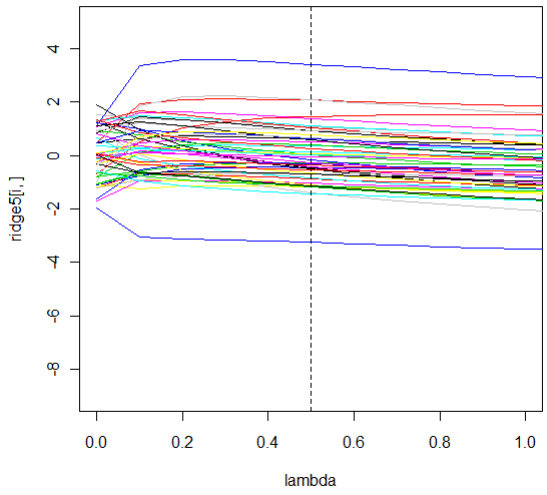


(c) Standardized β_3 across the 50 trials

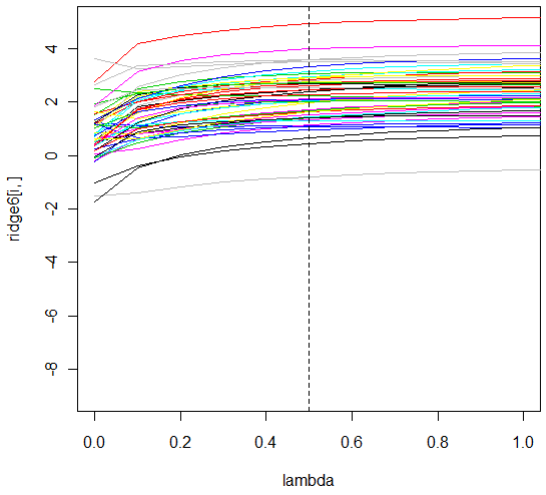


(d) Standardized β_4 across the 50 trials

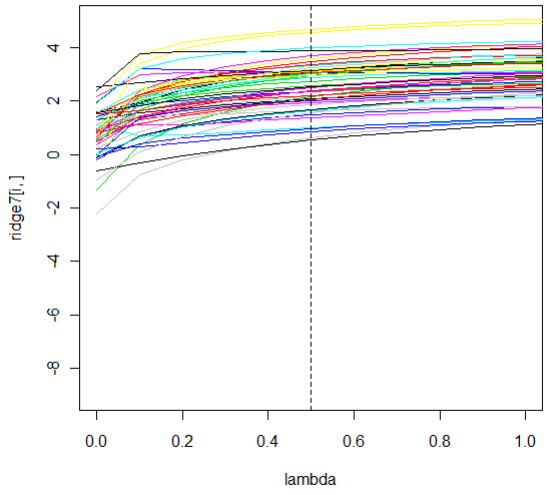
Figure 5.10: First four standardized β coefficients across the 50 trials (Example 3).



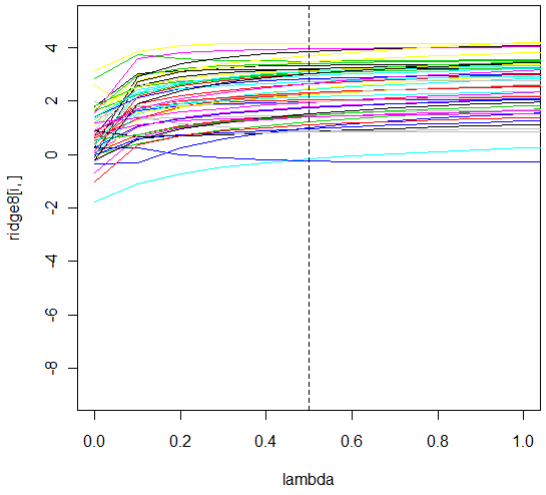
(a) Standardized β_5 across the 50 trials



(b) Standardized β_6 across the 50 trials

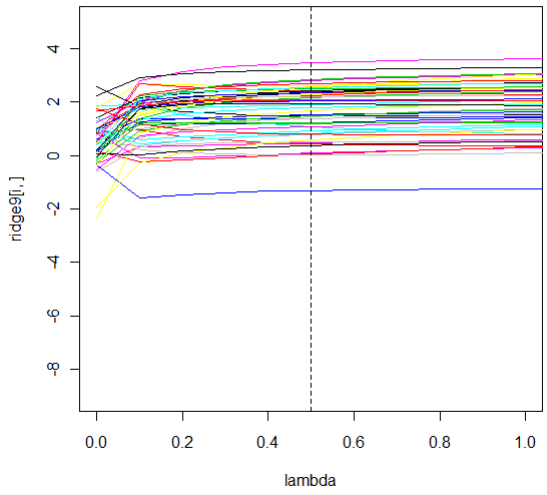


(c) Standardized β_7 across the 50 trials

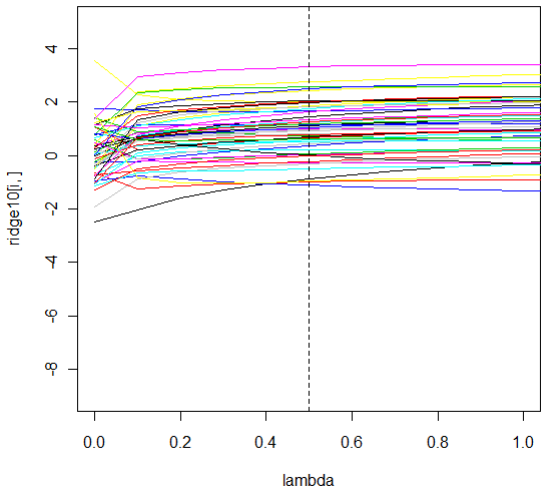


(d) Standardized β_8 across the 50 trials

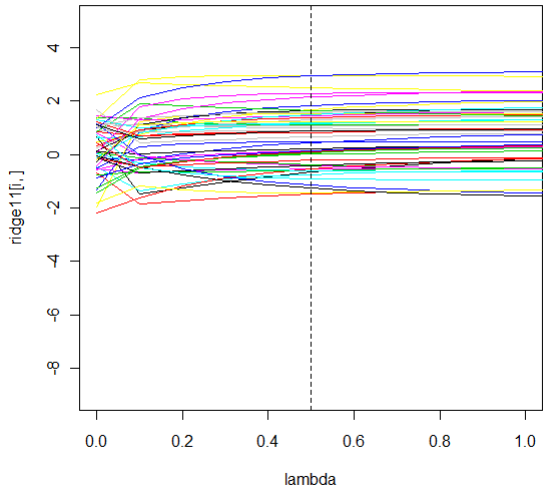
Figure 5.11: Second four standardized β coefficients across the 50 trials (Example 3).



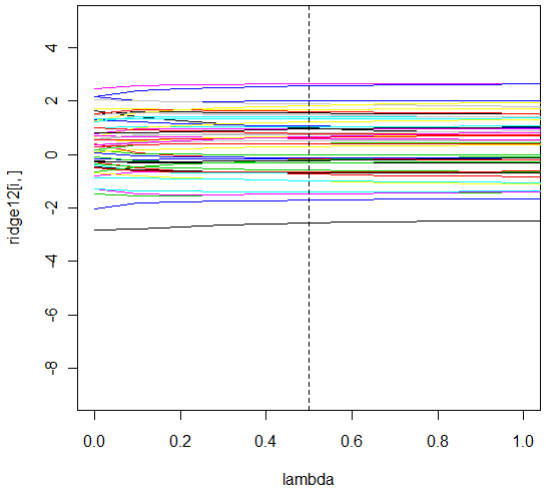
(a) Standardized β_9 across the 50 trials



(b) Standardized β_{10} across the 50 trials

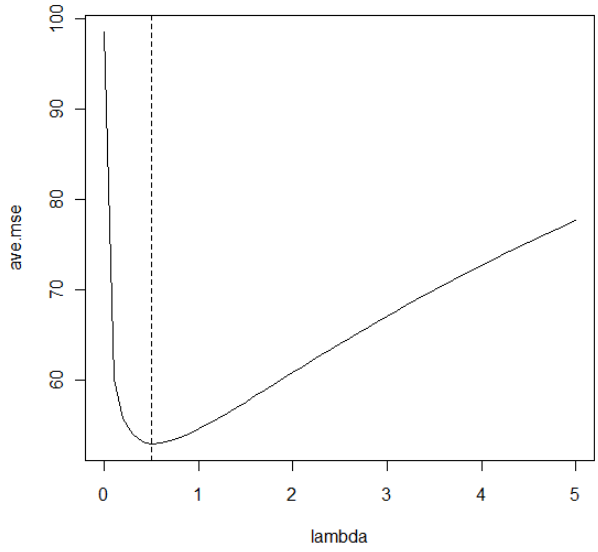


(c) Standardized β_{11} across the 50 trials

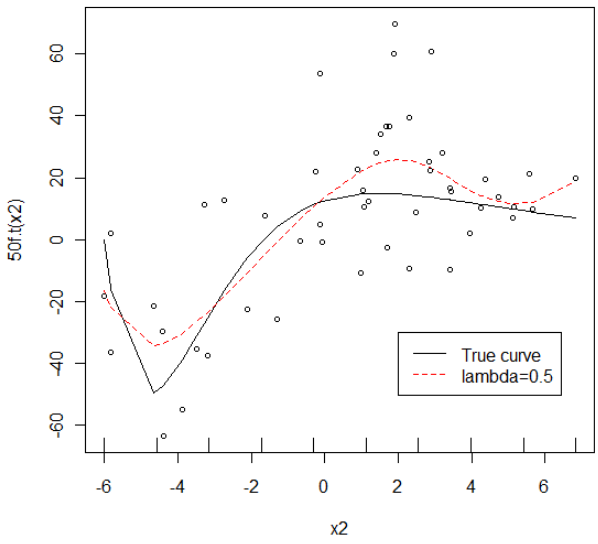


(d) Standardized β_{12} across the 50 trials

Figure 5.12: Last four standardized β coefficients across the 50 trials (Example 3).

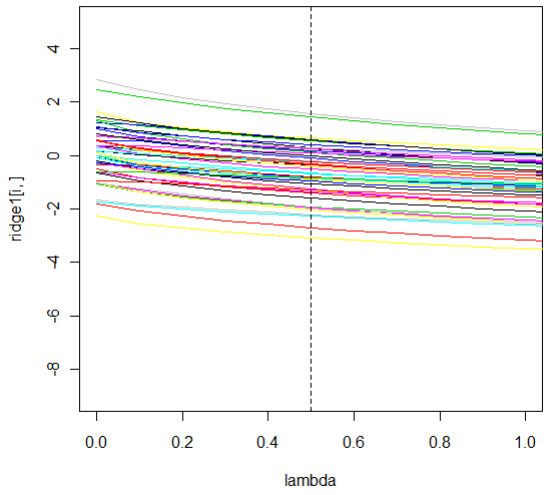


(a) AMSE as a function of λ

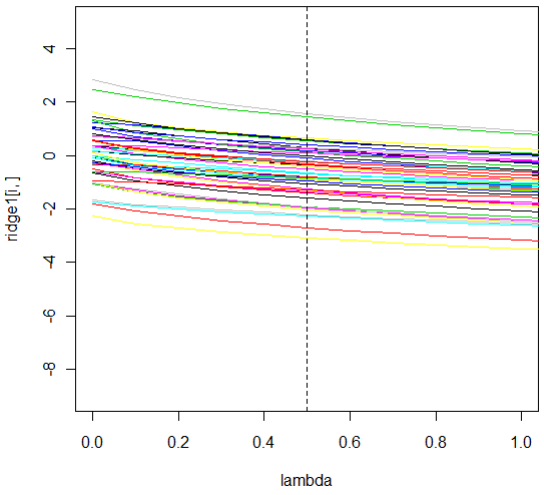


(b) True and estimated curve with the optimal λ. Points in the graph denote Y_i's. The placement of knots were indicated by inner grids.

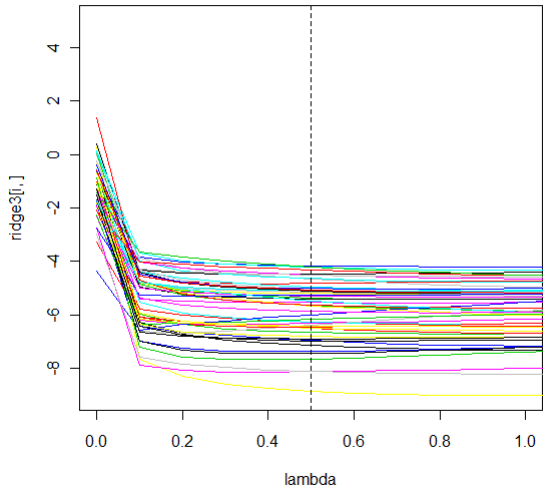
Figure 5.13: Fitted function with the value of λ in at which AMSE is minimum (Example 4).



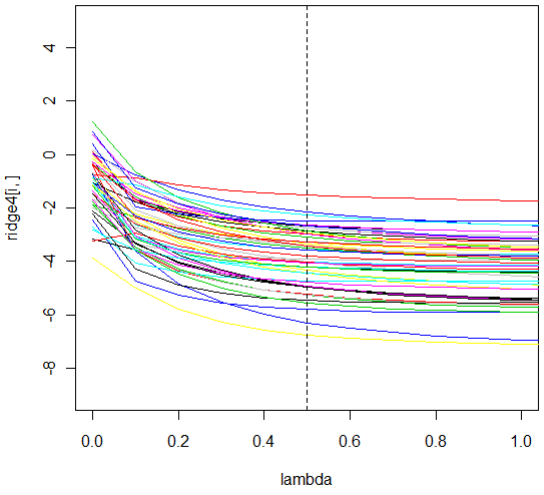
(a) Standardized β_1 across the 50 trials



(b) Standardized β_2 across the 50 trials

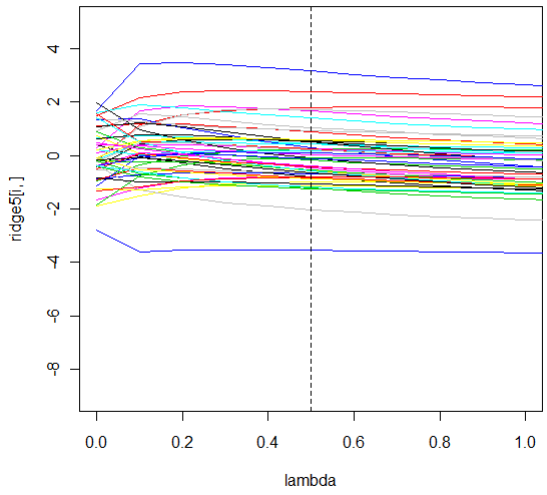


(c) Standardized β_3 across the 50 trials

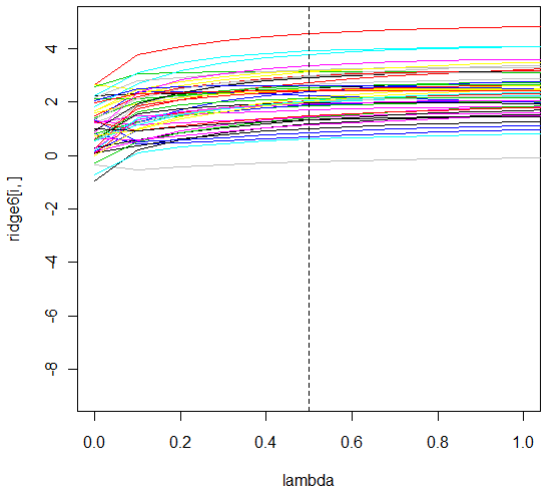


(d) Standardized β_4 across the 50 trials

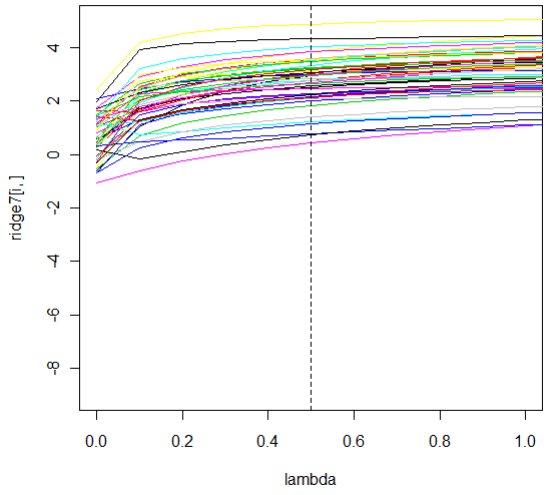
Figure 5.14: First four standardized β coefficients across the 50 trials (Example 4).



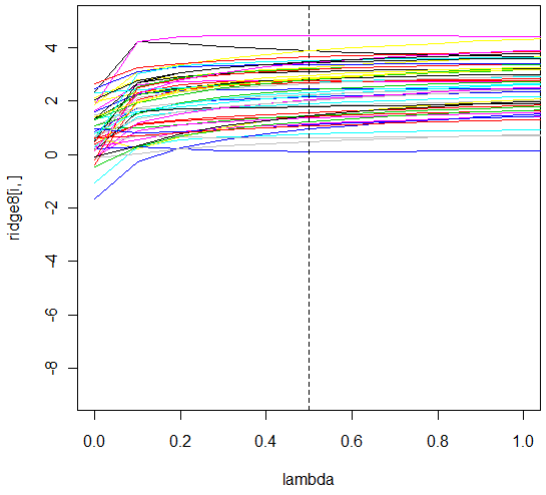
(a) Standardized β_5 across the 50 trials



(b) Standardized β_6 across the 50 trials

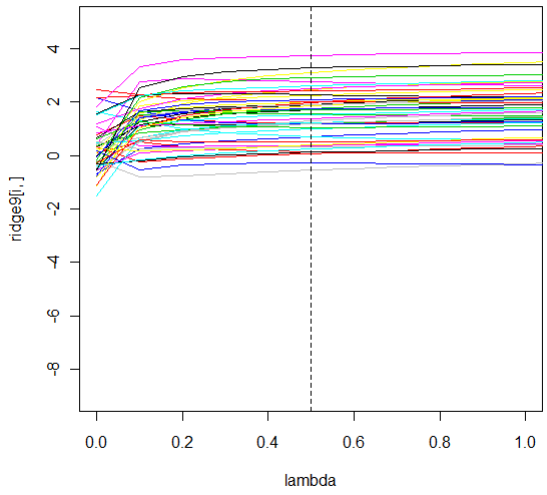


(c) Standardized β_7 across the 50 trials

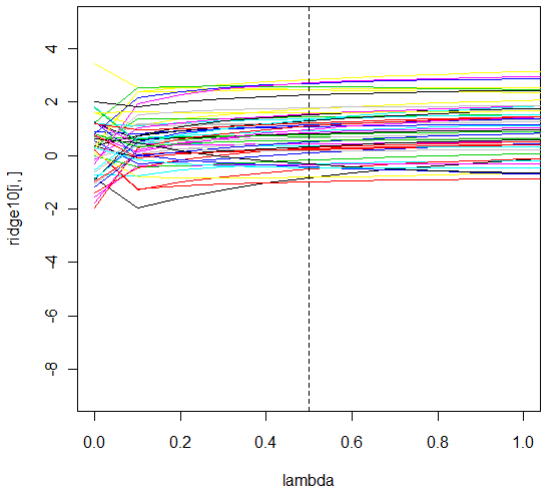


(d) Standardized β_8 across the 50 trials

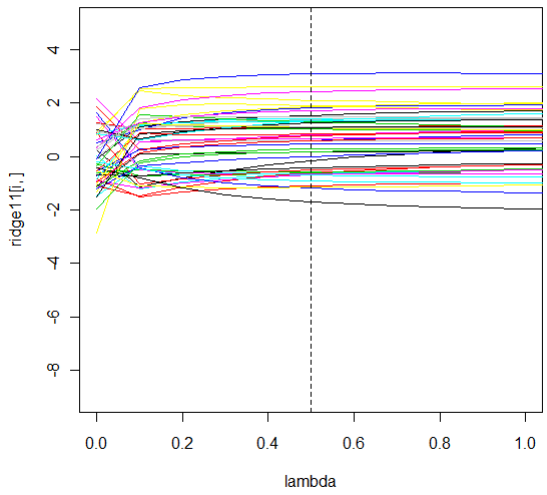
Figure 5.15: Second four standardized β coefficients across the 50 trials (Example 4).



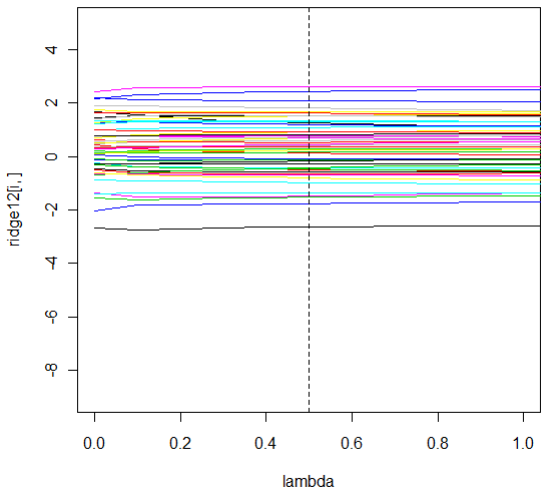
(a) Standardized β_9 across the 50 trials



(b) Standardized β_{10} across the 50 trials



(c) Standardized β_{11} across the 50 trials



(d) Standardized β_{12} across the 50 trials

Figure 5.16: Last four standardized β coefficients across the 50 trials (Example 4).

Chapter 6

Data Analysis

6.1 Description of Data

As mentioned in Chapter 1, our goal is to understand the effect of dust and low wind events on hospitalizations for asthma while adjusting for hourly levels of air pollutants. Data from three different sources are used in our study: hospitalization data due to asthma, weather, and pollution data. First, we describe our data in detail, then results are reported.

6.1.1 Hospitalizations Data Due to Asthma

The patients' information on hospitalizations due to asthma was obtained from the Texas Health Care Information Council (*THCIC*) in Austin, Texas for the time period 2000 through 2005. The original dataset contained patients information for four diseases: asthma, bronchitis, sinusitis and upper respiratory infections, but did not include the visits to the emergency room or to a primary care provider. From this data we extracted the patients living in El Paso county that were hospitalized due to asthma (*ICD-9 code 493.X*) during the study period. Each admission was given a unique identification number (*PAT ID*) so that even if the same patient was admitted to the hospital twice during the study period, they were considered as two different patients. Babies (patients less than 1 year old) were excluded from our study, since it is difficult to diagnose asthma in babies. Once the relevant patients were extracted from the original dataset, just the patient ID and the admission date were retained for our study. There were total of 5437 asthma admissions in the final dataset for the time period 2000-2005.

6.1.2 Weather Data

The weather data (average daily temperature, dew point, average wind speed and occurrence of dust events) were obtained from the US National Weather Service collected at the El Paso International Airport. Instead of using the average daily temperature, we used apparent temperature created based on the Celsius scaled values of average daily temperature and dew point of the day. The standard formula for the apparent temperature is

$$Apparent\ Temperature = -2.653 + 0.994\ Temperature(^0C) + 0.0153\ Dew\ Point\ ^2(^0C)$$

In the raw data there were three types of storm conditions: no dust storms ("0"), convectivly driven events or haboob dust storms ("C") and non-convective dust storms ("D"). As mentioned in Chapter 1, we ignore the convective/haboob storms and focus only on non-convective dust storms. More specifically, we collapse the the first two categories ("0" and "C") and defined them as days with no dust storms and only non-convective dust storms ("D") were considered as dust storm days. We create a new indicator variable to specify the dust storm condition of each day in the study period such that

$$storm = \begin{cases} 1 & \text{if the original code is "D"} \\ 0 & \text{if the original code is "0" or "C"} \end{cases}$$

There were 145 dust storm days reported in El Paso county during the six year study period. In addition, we also consider another weather variable: low wind. We convert the average wind speed into an indicator variable of low wind events. We use the 10^{th} percentile (4.5 mph) of the daily average wind speed for the six year period as the cutoff point for the low wind:

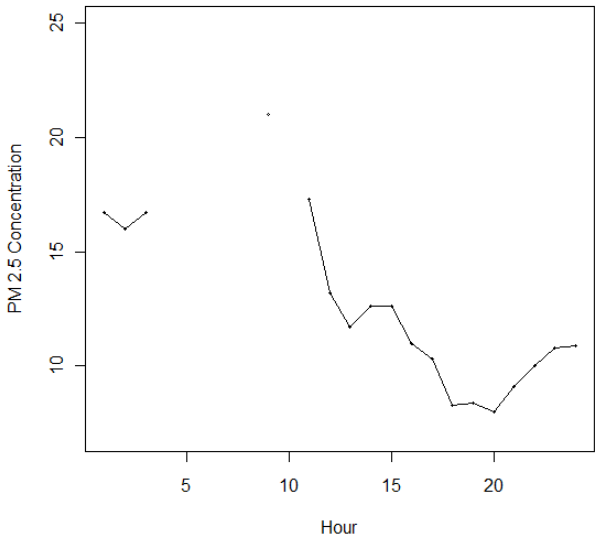
$$lowwind = \begin{cases} 1 & \text{if average wind speed} \leq 4.5 \text{ mph} \\ 0 & \text{Otherwise.} \end{cases}$$

Based on this definition there were 221 low wind days in El Paso county for the time period 2000 through 2005.

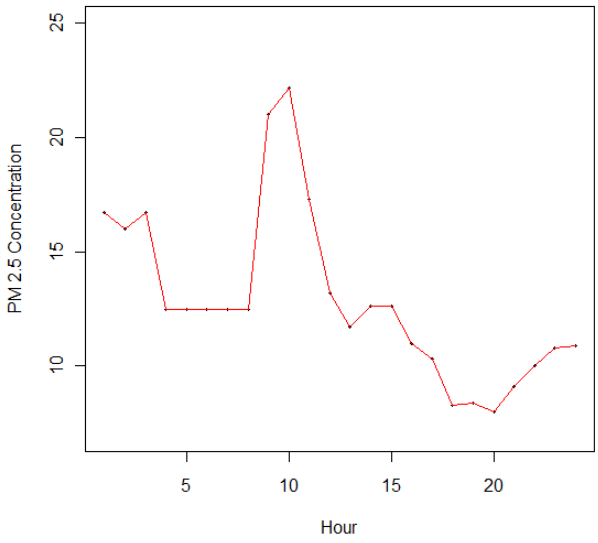
6.1.3 Air Pollution Data

Hourly measurements of $PM_{2.5}$ and NO_2 for the study period were obtained from the Aerometric Information Retrieval System for the site located at the edge of the University of Texas at El Paso (CAMS 12). There are two stations (CAMS 12 and CAMS 40) in El Paso providing hourly $PM_{2.5}$ data for the study period. CAMS 40 is located in an industrial region of El Paso, which is not heavily populated. Since the $PM_{2.5}$ levels from CAMS40 do not sufficiently represent the general population in El Paso, only $PM_{2.5}$ data collected at CAMS12 was used for this study. $PM_{2.5}$ measurements from CAMS12 may not be a perfect representation of exposure for the general population in El Paso and also it may underestimate the $PM_{2.5}$ in the region as well. But CAMS12 data is the only complete data available for the six year study period as the CAMS40 data is available only for the time period 2001-2005. Due to the mechanical failures of the monitoring station or automatic shutdowns due to high wind speed in the area, some of the hourly measurements of $PM_{2.5}$ and NO_2 were missing during the study period. The procedure discussed below was followed to impute the missing measurement of both $PM_{2.5}$ and NO_2.

- Five or more consecutive missing hourly levels in a given day were replaced by the the mean of the available hourly observations of that day.
- Days with all hourly measurements missing were replaced with the mean of all available hourly data over the six year period.
- B- spline interpolation was used to impute any missing observations spanning fewer than 5 hours.
- Once the interpolation was done, the negative values were replaced by zero.



(a) With missing data



(b) Imputed data

Figure 6.1: $PM_{2.5}$ Data August 2 2005

Figure 6.1 is a display of missing and imputed $PM_{2.5}$ data on August 2, 2005. On this day $PM_{2.5}$ measurements of $4^{th}, 5^{th}, 6^{th}, 7^{th}$ and 8^{th} hours were consecutively missing as well as the 10^{th} hour. The data are shown in 6.1(a) and the imputation is shown in 6.1(b).

6.2 Results

In the remainder of the report we explain the model fitting procedure using the statistical methods explained in the previous chapters.

6.2.1 Truncated Historical Functional Linear Model

Recall that, we have a binary response with a case-crossover design where each patient serves as their own control. Thus, each patient is a stratum with admission date taken to be as the case and reference days within a symmetric bi-directional referent window serve as his/her controls. In Chapter 3, the historical functional linear model was introduced for a daily response given hourly measurements of pollutants in a feed-forward nature. This means that the probability of hospitalization on day t depends only on the behavior of hourly $PM_{2.5}$ at past times. Daily hospitalization is modeled as a binomially distributed random variable with mean denoted by $\mu_t = E[Y(t)]$. The *logit* link for the binomial distribution is $\eta_t = log(\mu_t)$. Using $PM_{2.5}$ as the only covariate, the model can be written as

$$\eta_t = \alpha(t) + \int_0^{24t} x(24t - u)\beta(u)du \tag{6.1}$$

where $\alpha(\cdot)$ is the time dependent intercept function and $\beta(u)$ is the slope function which reflects the strength of association between $PM_{2.5}$ and log odds of hospitalization at lag u hours. For instance, $\beta(u)_{|u=lag12\ in\ hours}$ describes the influence of the noon reading of hourly $PM_{2.5}$ on today's hospitalizations. It is unreasonable to assume that the hospitalizations on day t is affected by all past exposures, so we used a truncated historical functional

linear model so that the dependence of daily hospitalizations is limited to pollutant exposures in recent history. That is, we assumed that only pollution exposures corresponding to time lag $u \leq \tau$ have an impact on hospitalizations on a given day. This reduces the length of the interval of the integral in (6.1). Thus, the truncated historical function linear model for $\eta_t = log(\mu_t)$ is written as

$$\eta_t = \alpha(t) + \int_0^{min(24t,24\tau)} x(24t-u)\beta(u)du.$$

Even though the truncation introduces one more parameter (τ), it simplifies the computer implementation of the historical functional linear model in predicting the probability of hospitalization on a current day. Yang (2005) [27] explains how the truncation of past influence to τ days affected the estimate of $\beta(\cdot)$. Following the same simplification procedure explained in Chapter 3, the truncated historical functional linear model with the logit link can be written as

$$\eta_t = \alpha(t) + \sum_{j=1}^{K} b_j \phi_j(t).$$

where

$$\phi_j(t;x) = \int_0^{24\tau} x_+(24t-u)N_j(u)du.$$

with the same previous definitions used in Chapter 4. Then the integral in $\phi_j(t;x)$ is approximated using Q equally spaced time points $\{g_q\}_{q=1}^{Q}$ in $[0, 24\tau]$:

$$\phi_j(t;x) = \frac{24\tau}{Q} \sum_{q=1}^{Q} x_+(S(t)-g_q)N_j(g_q).$$

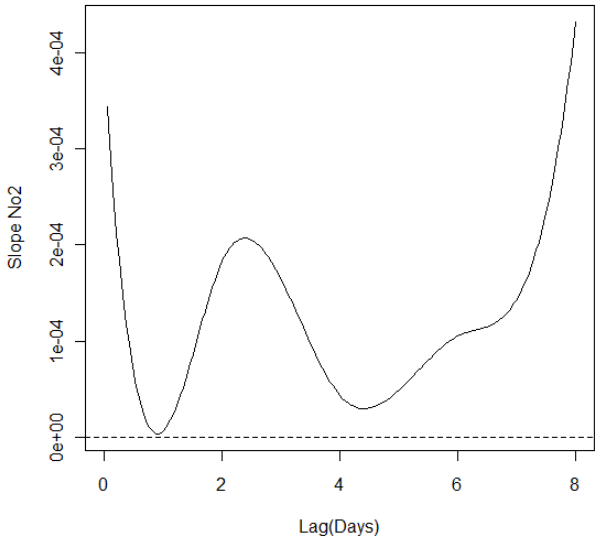
We set $\frac{24\tau}{Q} = 1$ by choosing time points (g_q) on an hourly grid. Recall that, this approximation was referred to as a preprocessing of the data. Similarly, we incorporate NO_2 to the regression model as well. The apparent temperature of the previous day (at_{t-1}) was included in the model as a piecewise linear function with two equally spaced interior knots

(no smoothing). Indicator variables for the storm and low wind of the current day were also included in the model. With all the above covariates, the semi-parametric model is written as

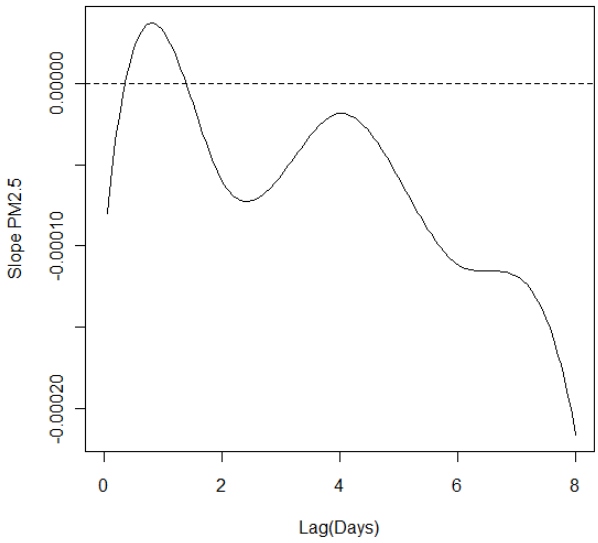
$$\begin{aligned} logit[Pr(Y(t)=1)] &= \alpha(t) + \text{piecewise linear function in } at_{t-1} + \\ &\quad \beta_1(lowwind_t) + \beta_2(storm_t) + \\ &\quad \int_0^{24min(t,\tau_1)} x(24t-u)\beta_x(u)du + \int_0^{24min(t,\tau_2)} z(24t-u)\beta_z(u)du. \end{aligned} \tag{6.2}$$

We included all the past hourly measurements of both $PM_{2.5}$ and NO_2 up to 8 days prior to the hospitalizations ($\tau_1 = \tau_2 = 8$). The truncation point was chosen based on the simulations in Yang (2005) [27]. Yang (2005) explains the unusual behavior of the slope function at both early and late truncations. After preprocessing both $PM_{2.5}$ and NO_2, we used the *COXPH* function in the R- package to fit the conditional logistic regression model. The estimated slope function in the historical functional linear model is given by $\beta(u) = \sum_{j=1}^{K} \hat{b}_j N_j(u)$, where $\{N_j(t), j = 1, ...K\}$ are cubic B-spline basis with knots placed at 2 day intervals and K is given by the number of interior knots plus the order of B-splines. The unpenalized slope functions of $PM_{2.5}$ and NO_2 obtained from model (6.2) are shown in Figure 6.2

Next, we used P-splines to get smooth estimates for the the slope functions of both $PM_{2.5}$ and NO_2. The *ridge* function provided in *COXPH* was modified to add the penalty needed for smoothing. Recall that, the estimated coefficients of P-spline fitting are given by $\hat{\beta} = (B^T B + \lambda D_2^T D_2)^{-1} B^T Y$ and the estimated coefficients in ridge regression $\hat{\beta}_{ridge} = (X'X + kI)^{-1} X'Y$, where B is the B-spline design matrix, D_2 is the matrix representation of the difference operator Δ^2, X is the design matrix of the data, and λ is the smoothing parameter



(a) NO_2 slope function



(b) $PM_{2.5}$ slope function

Figure 6.2: Unpenalized slope function of NO_2 and $PM_{2.5}$

in P-splines; k is the biasing constant in ridge ridge regression. Thus, the *ridge* function uses $k[\beta^T(I^T I)\beta]$ as the penalty in the likelihood function. We replaced it with $\lambda[\beta^T(D_2^T D_2)\beta]$. Next we will explain how a visual inspection of the adapted the ridge trace plot aided in the choice of smoothing parameter in the truncated historical functional linear model.

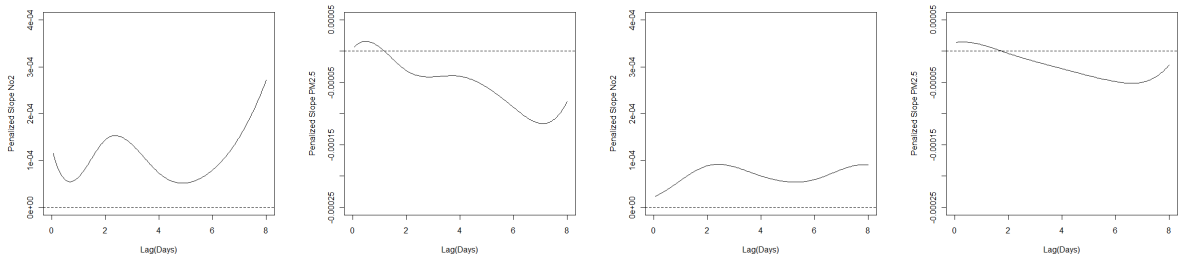
6.2.2 Choosing the Smoothing Parameters for the Hospitalization Data

We already obtained the unpenalized slope functions for both $PM_{2.5}$ and NO_2 in the regression model 6.2. Then we explained how we changed the *ridge* function to add the penalty in the nonparametric functional linear model. In Chapter 5 we discussed how to adapt the ridge trace to choose a suitable smoothing parameter in a nonparametric regression model. Now we will explain how to use the adapted ridge trace concept to find suitable smoothing parameters that reflect the relationship between the two pollutants: NO_2 and $PM_{2.5}$ and the probability of daily hospitalizations. To start with, we increased the penalty in both NO_2, $PM_{2.5}$ simultaneously until the slope functions became straight lines. At $\lambda = 10^{10}$ the slope functions of both NO_2 and $PM_{2.5}$ were straight lines, in fact, they were flattened to the horizontal zero line. Figure 6.3 shows how the slope functions slowly move towards the horizontal zero line as λ increases.

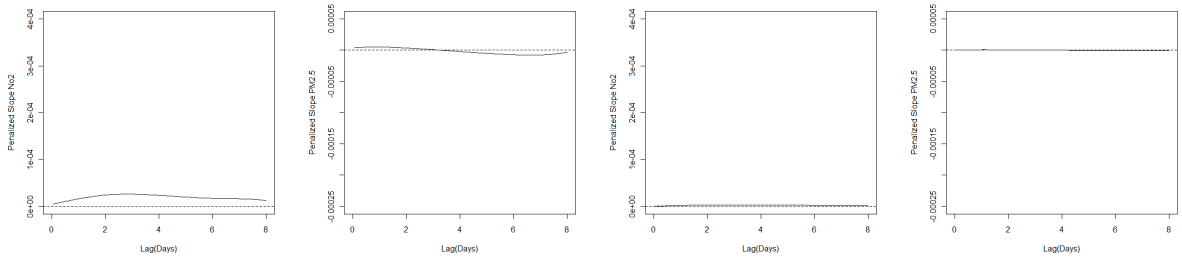
In ridge trace the biasing constant usually varies from 0 to 1. In order to keep the axis scale for the biasing constant of the adapted ridge trace plot within 0 and 1, let

$$c = \frac{\lambda}{\lambda_{max}}$$

where λ_{max} is the value of λ at which the slope function becomes a straight line ($\lambda_{max} = 10^{10}$ for both NO_2, $PM_{2.5}$). Table 6.1 gives the c values and the corresponding value of λ. The estimated coefficients and standard errors of the B-spline coefficients were obtained by fitting the regression model (6.2) in *COXPH*.



(a) NO_2 slope function when $\lambda = 10^7$ (b) $PM_{2.5}$ slope function when $\lambda = 10^7$ (c) NO_2 slope function when $\lambda = 10^8$ (d) $PM_{2.5}$ slope function when $\lambda = 10^8$



(e) NO_2 slope function when $\lambda = 10^9$ (f) $PM_{2.5}$ slope function when $\lambda = 10^9$ (g) NO_2 slope function when $\lambda = 10^{10}$ (h) $PM_{2.5}$ slope function when $\lambda = 10^{10}$

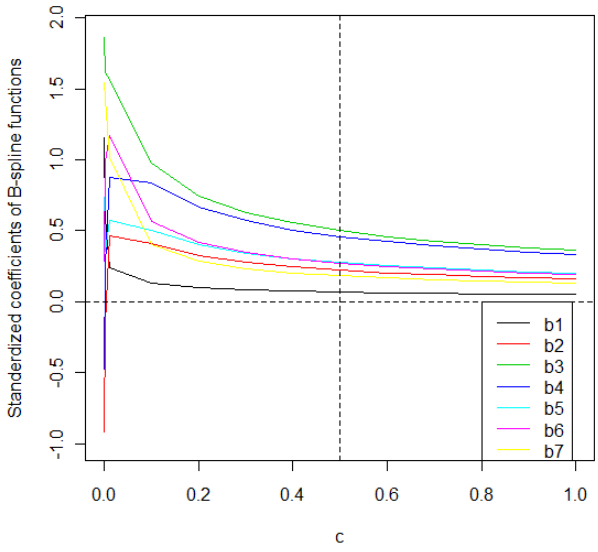
Figure 6.3: Behavior of the slope functions as λ increases

Table 6.1: c and corresponding λ values.

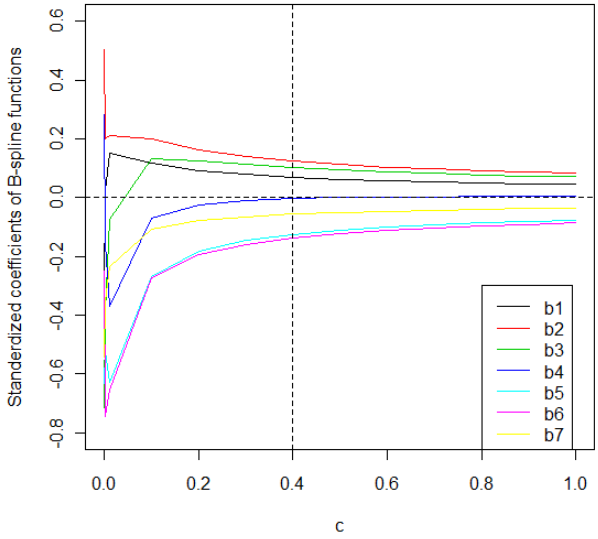
λ	c
0	0
10^7	0.001
10^8	0.01
10^9	0.1
2×10^9	0.2
3×10^9	0.3
4×10^9	0.4
5×10^9	0.5
6×10^9	0.6
7×10^9	0.7
8×10^9	0.8
9×10^9	0.9
10^{10}	1

Adapted ridge trace plots for NO_2 and $PM_{2.5}$ obtained by plotting $\hat{b_j}/SE(b_j)$ (standardized coefficients) versus c are shown in Figure 6.4. By visual inspection of the adapted ridge trace plot for NO_2, we can see that the standardized coefficients start to stabilize around $c = 0.5$. Thus, we picked the corresponding value of λ (5×10^9) for smoothing with respect to NO_2. Similarly, the smoothing parameter value of 4×10^9 for $PM_{2.5}$ was chosen. The slope functions with the chosen smoothing parameters obtained by fitting the penalized conditional logistic regression model in (6.2) are shown in Figure 6.5.

The fitted slope function for NO_2 has a broad global maximum at 1:00 a.m two days prior to the hospitalizations, suggesting that the hospitalizations on today are most highly as-

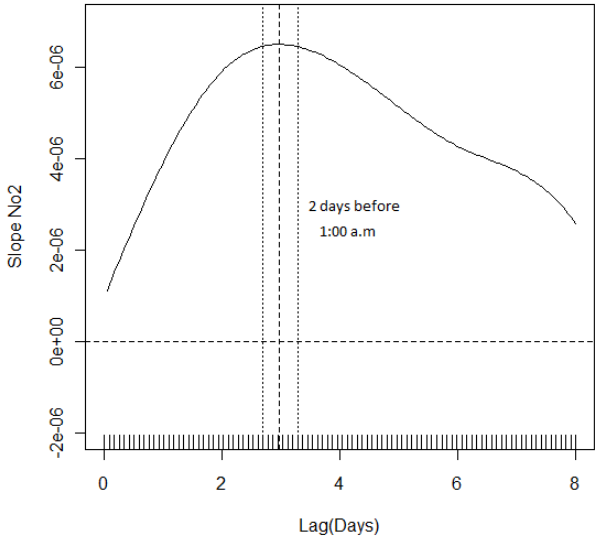


(a) Adapted ridge trace for choosing smoothing parameter for NO_2.

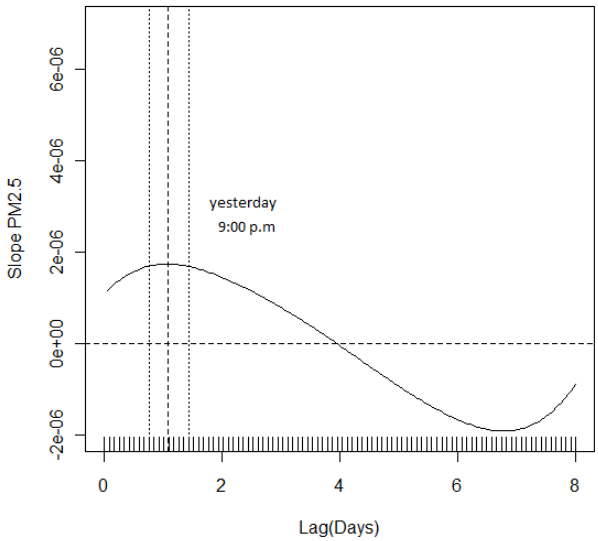


(b) Adapted ridge trace for choosing smoothing parameter for $PM_{2.5}$

Figure 6.4: Adapted ridge trace plots for the hospitalization data



(a) Estimated slope function for NO_2 with the chosen smoothing parameter (5×10^9).



(b) Estimated slope function for $PM_{2.5}$ with the chosen smoothing parameter (4×10^9).

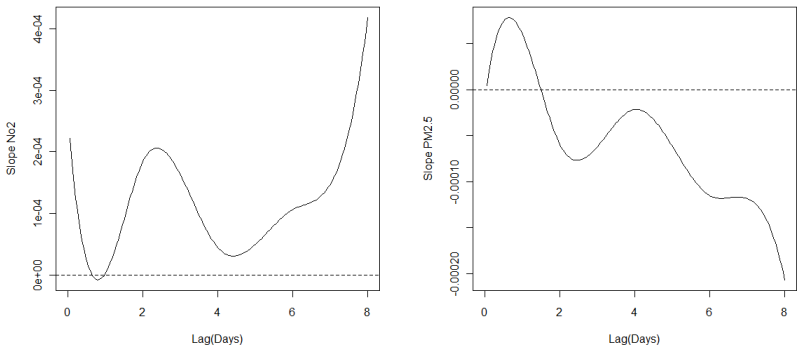
Figure 6.5: Penalized slope functions with chosen smoothing parameters

sociated with NO_2 concentrations 3 days ago during late night and during early morning 2 days ago. Todays hospitalizations are most highly associated with the $PM_{2.5}$ concentrations on yesterday around 9:00 p.m. The fitted slopes suggests that it takes 2-3 days to mount an immune response to NO_2, whereas only 0-1 day for $PM_{2.5}$. The shape of the $PM_{2.5}$ slope function suggests that a contrast between the $PM_{2.5}$ levels at lag 0-3 with lag 3-7 is the predictor for hospital admissions.

Recall that, our goal was to understand the effect of dust and low wind events on asthma hospitalizations while adjusting for hourly levels of pollutants. Note that, in the regression model (6.2), we have indicator variables for the storm and low wind conditions on the same day as the hospitalizations. In Chapter 1, we also mentioned it is not reasonable to assume that the hospitalizations on a given date is affected only by the weather and pollution conditions on the same day. Thus, we ran separate models not only for storm and low wind on the same day, but also at lags 1, 2 and 3 days as well. In general, the models can be expressed as

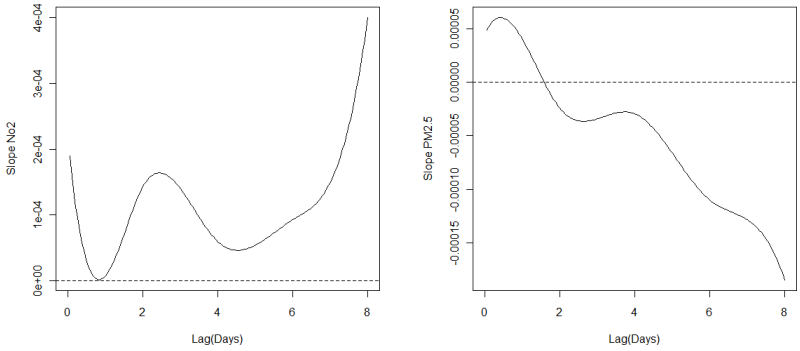
$$\begin{aligned} logit[Pr(Y(t)=1)] &= \alpha(t) + \text{piecewise linear function in } at_{t-1} + \\ &\quad \beta_1(lowwind_{t-l}) + \beta_2(storm_{t-l}) + \\ &\quad \int_0^{24min(t,\tau_1)} x(24t-u)\beta_x(u)du + \int_0^{24min(t,\tau_2)} z(24t-u)\beta_z(u)du \end{aligned} \tag{6.3}$$

where l takes values 0,1,2 and 3. The unpenalized slope functions of NO_2 and $PM_{2.5}$ obtained for each of the above models, did not change much as l varied from 0 to 3, so the values of the smoothing parameters already selected for lag $l = 0$ were also used for $l = 1, 2, 3$. Figure (6.6) shows the unpenalized slope functions obtained each time.



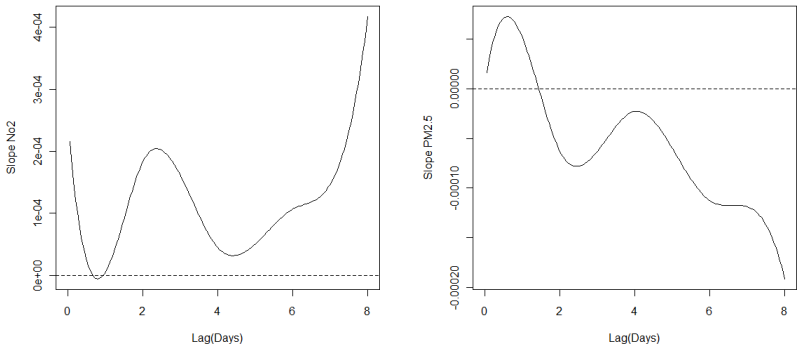
(a) NO_2 slope function with lag 1 storm and low wind in the model.

(b) $PM_{2.5}$ slope function with lag 1 storm and low wind in the model.



(c) NO_2 slope function with lag 2 storm and low wind in the model.

(d) $PM_{2.5}$ slope function with lag 2 storm and low wind in the model.



(e) NO_2 slope function with lag 3 storm and low wind in the model.

(f) $PM_{2.5}$ slope function with lag 3 storm and low wind in the model.

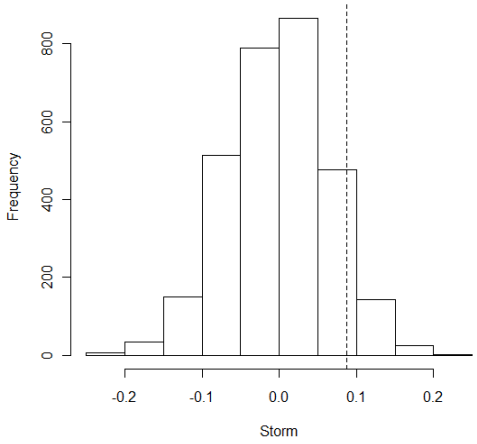
Figure 6.6: Unpenalized slope functions of NO_2 and $PM_{2.5}$: sensitivity for the lag of storm and low wind.

In order to find the significance of storm and low wind, we derived the bootstrap distributions of the coefficients separately for storm and low wind. As mentioned earlier there were 145 storm days and 221 low wind days during the study period. For instance, to find the bootstrap distribution of the coefficient of the storm indicator variable, we generate a random sequence of 0's and 1's with 145 1's and the rest zeros. Then we shuffled and assigned 1's and 0's to each day of the study period. Storm lag 1, 2 and 3 days were also replaced correspondingly. Each time, the regression models (6.3) were fitted and the coefficients stored. Finally p values for the one tailed test were calculated by counting the number of bootstrap storm coefficients greater than the coefficient estimated from the data. Similarly, we derived the bootstrap distribution of the low wind indicator variable as well. The frequency histograms of the bootstrap distributions for storm and low wind are shown in Figure 6.7 and Figure 6.8, respectively. Table 6.2 displays the fitted values of the regression coefficients and the corresponding p-values obtained from the bootstrap distributions.

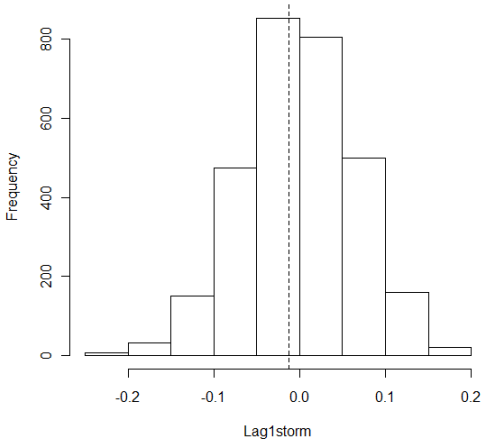
Table 6.2: Estimated regression coefficients for storm and low wind indicator variables. The corresponding 1-tail p-values obtained from the bootstrap distributions are given within brackets.

Lag	Storm	Low wind
0	0.0876 (0.084)	0.0045 (0.467)
1	-0.0128 (0.576)	0.0051 (0.461)
2	-0.113 (0.95)	0.0625 (0.113)
3	-0.0233 (0.629)	0.0508 (0.183)

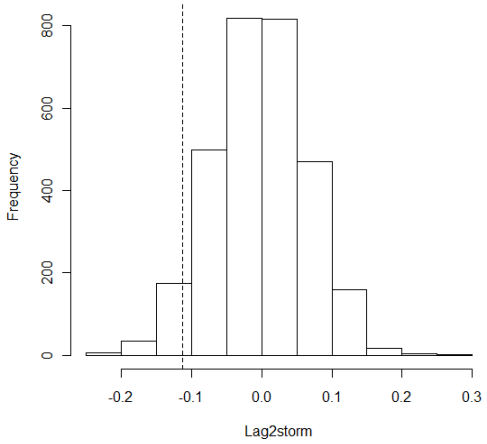
We picked the lags of storm and low wind with the lowest p-value/ highest odds ratio (lag 0 storm and lag 2 low wind) and ran a separate model



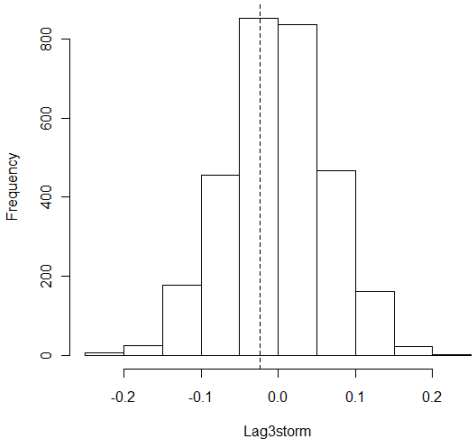
(a) Bootstrap distribution of lag 0 storm



(b) Bootstrap distribution of lag 1 storm

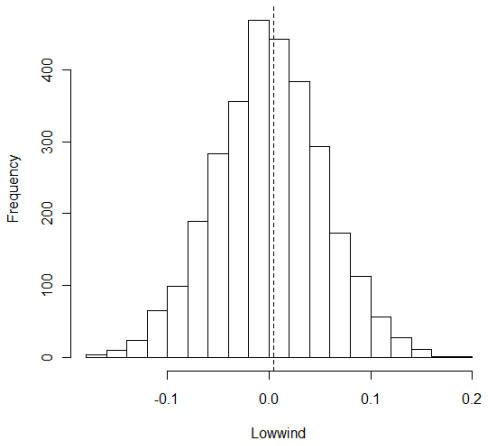


(c) Bootstrap distribution of lag 2 storm

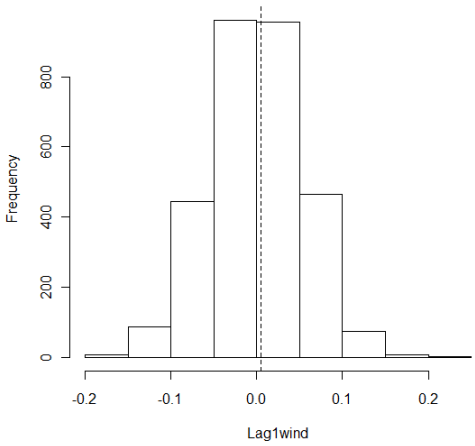


(d) Bootstrap distribution of lag 3 storm

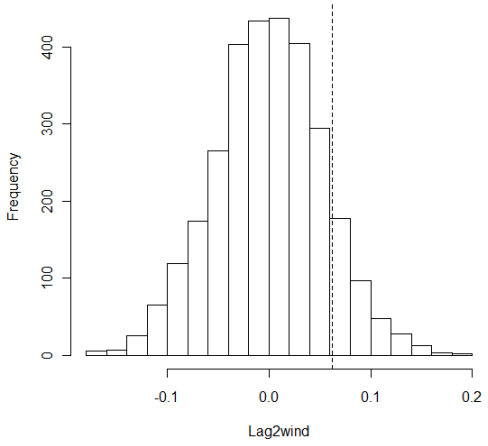
Figure 6.7: Bootstrap distributions of storm coefficient under no effect. The fitted value is indicated by the dashed line



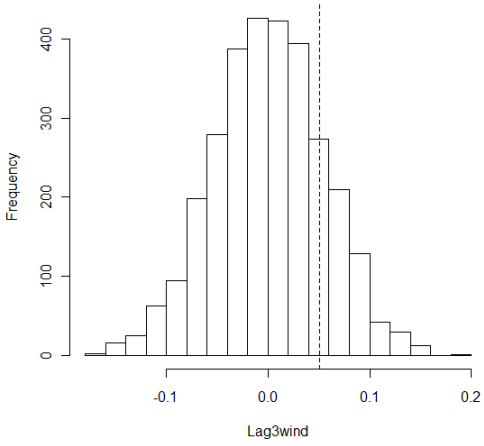
(a) Bootstrap distribution of lag 0 low wind



(b) Bootstrap distribution of lag 1 low wind



(c) Bootstrap distribution of lag 2 low wind

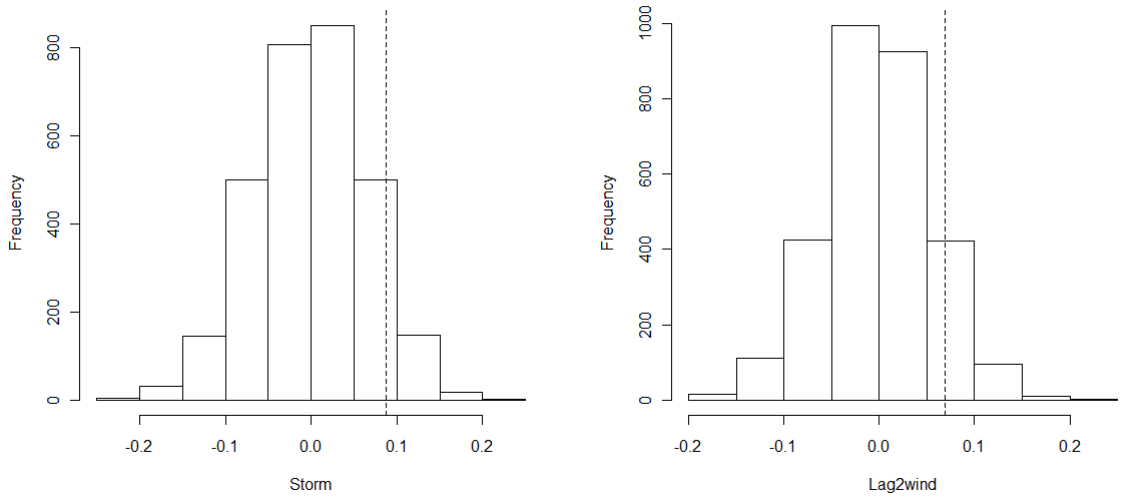


(d) Bootstrap distribution of lag 3 low wind

Figure 6.8: Bootstrap distributions of Low wind coefficient under no effect. The fitted value is indicated by the dashed line

$$
\begin{aligned}
logit[Pr(Y(t)=1)] &= \alpha(t) + \text{piecewise linear function in } at_{t-1} + \\
&\quad \beta_1(lowwind_{t-2}) + \beta_2(storm_t) + \\
&\quad \int_0^{24min(t,\tau_1)} x(24t-u)\beta_x(u)du + \int_0^{24min(t,\tau_2)} z(24t-u)\beta_z(u)du.
\end{aligned} \tag{6.4}
$$

The frequency histograms of the bootstrap distributions of the coefficient of lag 0 storm and lag 2 low wind are shown in Figure 6.9. Corresponding 1-tail p-values are 0.080 for storm ($\hat{\beta}_2 = 0.0875$) and 0.098 for lag 2 low wind ($\hat{\beta}_1 = 0.0685$). We can see that both lag 0 storm and lag 2 low wind are significant at the 10% level of significance suggesting that the probability of asthma hospitalization on a given day is increased by occurrence of dust storms on the same day and low wind events on two days prior to the admissions.



(a) Bootstrap distribution of lag 0 storm (p-value=0.080)

(b) Bootstrap distribution of lag 2 low wind (p-value=0.098)

Figure 6.9: Bootstrap distributions of coefficients of the picked lags of storm and low wind under no effect. The fitted value is indicated by the dashed line

Chapter 7

Discussion

Our goal was to understand the effect of dust and low wind events on hospitalizations for asthma while adjusting for hourly levels of air pollutants. In this chapter we will discuss the strengths and the limitations of the methods we used and possible future work related to this study. We explored how to use the truncated historical functional linear model for a daily response with hourly measurements of predictors that lessen the information loss. The great advantage of the historical functional linear model was demonstrated, namely, the pollutant past exposure was included without having to choose a (short-term) lag. Based on the theoretical relationship between P-splines and ridge regression we modified the available functions in the R-package (*COXPH*) to allow for a penalty in the nonparametric functional linear model, which was a very time saving approach for us. We introduced a simple, but meaningful, method to choose a suitable smoothing parameter for the P-spline estimator. As mentioned earlier, the commonly suggested methods to choose the smoothing parameter are complicated in practice, since those are based on approximated mean squared error. The results obtained from the simulated examples suggest that the adapted ridge trace plot can be used to choose a suitable smoothing parameter. We use a case-crossover study design that sets up patients as their own controls. We could have modeled the daily counts of hospitalizations for asthma by fitting a Poisson regression model instead of a conditional logistic regression model. But the case-crossover study design has some advantages over a study based on daily counts of hospitalizations : (1) the confounding effects of seasonal variables are reduced since the cases serve as their own controls; (2) it is possible to expand the study to test for the effect modification of patient characteristics such as age, gender, insurance plan by including interaction terms. One limitation in our

study is that we had to assume that the each hospitalization has occurred at 11:00 p.m due to the unavailability of the time of hospitalizations.

In this study we were not interested in testing for the significance of NO_2 and $PM_{2.5}$. The association between the a pollutant and the probability of hospitalizations for asthma was given by a function instead of one regression coefficient. Thus, it remains to explore methods for hypothesis testing of $H_0 : \beta(t) = 0$ for $t \in [0, \tau]$.

References

[1] V. Belleudi, A. Faustini, M. Stafoggia, G. Cattani, A. Marconi, C. A. Perucci, and F. Forastierea. Impact of fine and ultrafine particles on emergency hospital admissions for cardiac and respiratory diseases. *Epidemiology*, 21:414–423, 2010.

[2] S. Chatterjee and A. S. Hadi. *Regression Analysis by Example*. Joan Wiley and sons, Inc, 2006.

[3] Carl de Boor. *A practical Guide to Splines*. Springer-Verlag, 1978.

[4] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–121, 1996.

[5] R.L Eubank. *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, 1999.

[6] S.E. Grineski, J. G. Staniswalis, Y. Peng, and C. Atkinson-Palombo. Childrens asthma hospitalizations and relative risk due to nitrogen dioxide (no2): Effect modification by race, ethnicity and insurance status. *Environmental Research*, 110:178–188, 2011.

[7] S.E. Grineski, J. G. Staniswalis, Y. Peng, P. Bulathsinhala, and T.E Gill. Hospital admissions for asthma and acute bronchitis in el paso, texas: Do age, sex, and insurance status modify the effects of dust and low wind events? Submitted to Environmental Research.

[8] A.E. Horel and R.W. Kennard. Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12:69–82, 1970.

[9] D. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, 1989.

[10] Efang Kong, Howell Tong, and Yingcun Xia. Statistical modeling of nonlinear long-term cumulative effects. *Statistica Sinica*, 20:1097–1123, 2010.

[11] M.H. Kutner, C.J. Nachtsheim, Joan Neter, and William Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 2005.

[12] S. Lewallen and P. Courtright. Epidemiology in practice: Case-control studies. *Journal of Community Eye Health, International Centre for Eye Health, London*, 11, 1998.

[13] M. Maclure. The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, 133:144–153, 1991.

[14] R. Mason and W.G. Brown. Multicollinearity problems and ridge regression in sociology models. *Social Science Research*, 4:135–149, 1975.

[15] G.C. McDonald and R.C. Schwing. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15:463–481, 1973.

[16] N.Malfait and J.Ramsay. The historical functional linear model. *The Canadian Journal of Statistics*, 31(2):115–128, 2003.

[17] D.J. Novlan, M. Hardiman, and T.E. Gill. "a synoptic climatology of blowing dust events in El Paso, texas from 1932-2000". *16th Conference on Applied Climatology, American Meteorological Society*, J3.12, 2007.

[18] Y. Peng. A retrospective study of dust storms and respiratory hospitalizations in el paso, texas using a case- crossover study design. Master's thesis, University of Texas at El Paso, 2009.

[19] L. Perez, A. Tobias, X. Querol, N. Kunzli, J. Pey, A. Alastuey, M. Viana, N. Valero, and M. Gonzalez-Cabreand J. Sunyera. Coarse particles rrom saharan dust and daily mortality. *Epidemiology*, 19:800–807, 2008.

[20] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, 2005.

[21] S.Almon. The distributed lag between capital appropriations and expenditures. *Econometrica*, 33:178–196, 1965.

[22] M.G Schimek. *Smoothing and Regression*. John Wiley and Sons, 2000.

[23] J. J. Schlesselman. *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York, 1982.

[24] J. Schwartz. Distributed lag between air pollution and daily deaths. *Epidemiology*, 11(3):320–326, 2000.

[25] J.G. Staniswalis, N.J. Parks, J.O. Bader, and Y. M. Maldonado. Temporal analysis of airborne particulate matter reveals a dose-rate effect on mortality in el paso: Indications of differential toxicity for different particle mixtures. *Journal of the Air & Waste Management Association*, 55:893–902, 2005.

[26] J.G. Staniswalis, H. Yang, W. Li, and K.E. Kelly. Using a continuous time lag to determine the association between ambient pm2.5 hourly levels and daily mortality. *Journal of the Air & Waste Management Association*, 59:1173–1185, 2009.

[27] H. Yang. A further study of the relationship between pm10 level and daily mortality in el paso, tx using a historical functional linear model. Master's thesis, University of Texas at El Paso, 2005.

[28] A. Zanobetti, M.P Wand, J. Schwartz, and L.M Ryan. Generalized additive distributed lag models:quantifying mortality displacement. *Biostatistics*, 1:279–292, 2000.

Curriculum Vitae

Priyangi Kanchana Bulathsinhala was born on March 1, 1983, in Colombo, Sri Lanka. She received her primary and secondary education at Anula Vidyalaya, Nugegoda, Sri Lanka, and was qualified to enter the college to pursue her bachelor's degree. Priyangi joined University of Colombo in 2004 and completed her bachelor's degree in Statistics in 2008. Right after her graduation she joined the Department of Statistics, University of Colombo, Sri Lanka as a temporary instructor where she received the encouragement and determination to pursue her studies.

She joined The University of Texas at El Paso in 2009 as a Statistics master's student at the Mathematical Sciences department. While pursuing her graduate studies she worked as teaching assistant, research assistant, and research technician at the department of Mathematical Sciences, University of Texas at El Paso. Based on her performances she was elected as The Outstanding Graduate Student in Statistics in 2011. She will continue her graduate studies at Souther Methodist University, Dallas, Texas.

Permanent address: 2/217, Egodawatte Road,
Boralesgamuwa, Sri Lanka