Partially Linear Models and Air Pollution

Isabella Ghement

STAT 526

Smoothing Methods in Regression

24 March, 2003

Mexico City Air Pollution Study

Goal:

Investigate the linear association between air pollution and mortality after adjusting for possible confounders:

- weather

  e.g: mortality higher at low and high temperatures

- seasonality

  e.g: mortality peaks in winter because of influenza epidemics

- long-term trends

  e.g: improvements in medical practice lower mortality over time

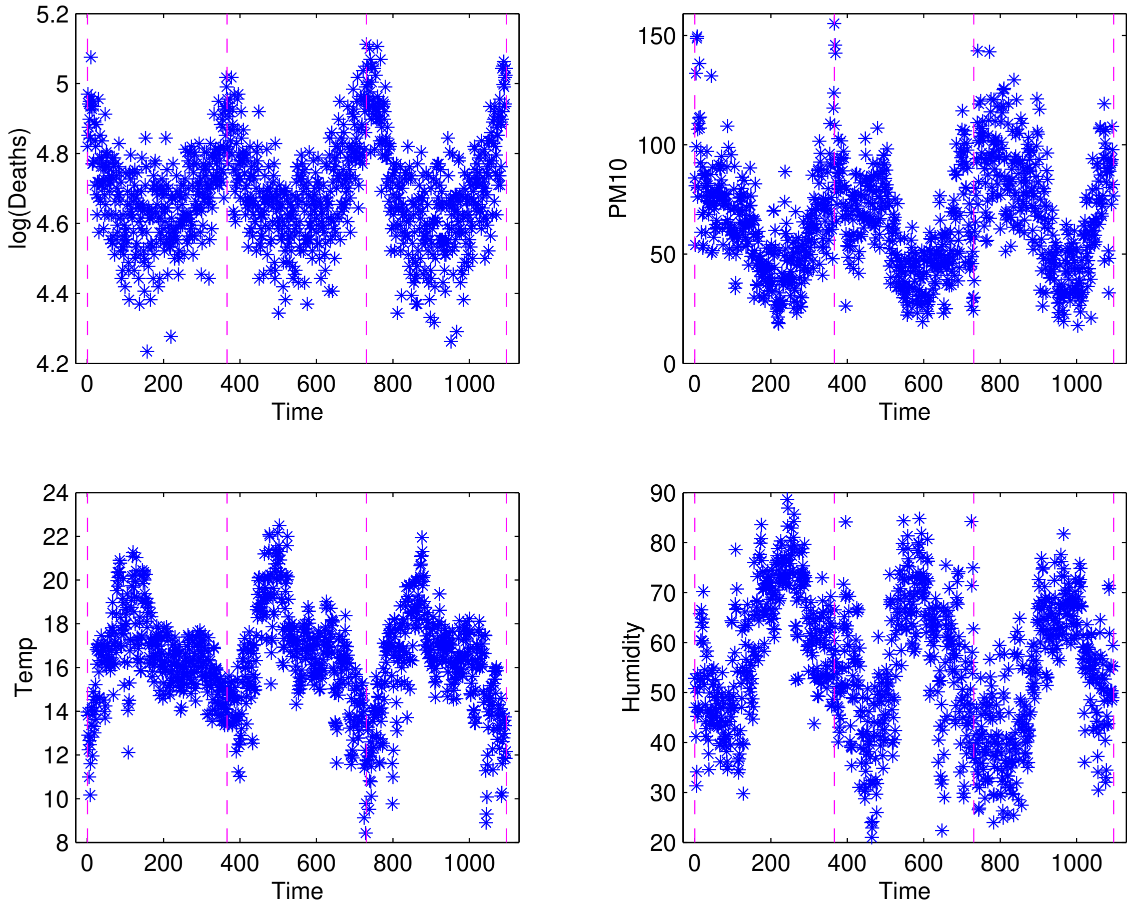
Data:

$Death_i$ = mortality count on day $i \in \{1, \ldots, 1096\}$

$PM10_i$ = level of particulate matter on day i

$Temp_i$ = level of temperature on day i

Hum_i = level of relative humidity on day i



Statistical Model:

$$
\begin{aligned}
Y_i \;=\; & \beta_0 + \overset{\substack{of\ interest \\ \downarrow}}{\beta_1} \cdot PM10_i + \\
& \underbrace{m_1(Temp_i) + m_2(Hum_i) + m_3(Day_i)}_{\substack{\uparrow \\ nuisance(smooth\ or\ linear)}} + \underset{\substack{\uparrow \\ correlated?}}{\epsilon_i}
\end{aligned}
$$

where $i \in \{1, \ldots, 1096\}$ and :

$\beta_1 =$ increase in log mortality per $10\mu g/m^3$ increase in PM10

$m_1, m_2, m_3 =$ unknown smooth functions for dealing with potential weather, seasonal and long-term trend confounding

Notes:

- This statistical model is referred to as a *semiparametric additive model*.

- The model is *identifiable* provided:

$$\sum_{i=1}^{1096} m_1(Temp_i) = 0$$
$$\sum_{i=1}^{1096} m_2(Hum_i) = 0$$
$$\sum_{i=1}^{1096} m_3(Day_i) = 0$$

- If weather effects are assumed *linear*, that is:

$$m_1(Temp_i) = \beta_2 \cdot Temp_i$$
$$m_2(Hum_i) = \beta_3 \cdot Hum_i$$

  while temporal effects are assumed *nonlinear*, then model is termed *partially linear*.

Goals of Statistical Analyses:

- if proposed model provides a good fit for these data, carry out inferences about the linear PM10 effect on log mortality, β_1
  - is this effect statistically significant?
  - what is the magnitude of this effect?

- assess whether the proposed model provides a good fit for these data

Challenges:

- PM10 signal has order of magnitude *smaller* than that of weather and temporal confounders

- degree of adjustment for confounders controlled by amount of smoothing performed on nonparametric terms in the model

- intuitively, amount of smoothing must be chosen so that

  – weather and temporal confounding are removed

  – PM10 signal is not entirely washed out

- theoretically, amount of smoothing must be chosen so that
  - estimate of linear PM10 effect, $\widehat{\beta}_1$, has *optimal* statistical properties

- choice of amount of smoothing complicated by:
  - high correlation between the linear and nonlinear covariates in the model (*concurvity*)
  - possible presence of serial correlation amongst model error

Partially Linear Model with I.I.D. Errors

Data:

$$(Y_i, X_{i1}, \ldots, X_{ip}, Z_i),\ i \in \{1, \ldots, n\}$$

Model:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \cdots + \beta_p X_{ip} + m(Z_i) + \epsilon_i$$

Identifiability Constraint:

$$\sum_{i=1}^{n} m(Z_i) = 0$$

Assumptions:

Z_i's fixed design points on $[0,1]$ with design density f

$X_{ij} = g_j(Z_i) + \eta_{ij},\ i \in \overline{1,n},\ j \in \overline{1,p}$

g_j's and m smooth, unknown

η_{ij}'s and ϵ_i's independent, unobserved

$\epsilon_i \sim (0, \sigma^2)$

$(\eta_{i1}, \ldots, \eta_{ip})^T \sim (\mathbf{0}, \mathbf{\Sigma})$

Notation:

$\boldsymbol{Y} = (Y_1, \ldots, Y_n) \quad X = (X_{ij})$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$

$\boldsymbol{m} = (m(Z_1), \ldots, m(Z_n))^T$

Backfitting Algorithm:

Estimates $\widehat{\boldsymbol{\beta}}(h)$ and $\widehat{\boldsymbol{m}}$ of $\boldsymbol{\beta}$ and $\boldsymbol{m}$ solve:

$$\begin{cases} \widehat{\boldsymbol{\beta}}(h) & = & (X^T X)^{-1} X^T (\boldsymbol{Y} - \widehat{\boldsymbol{m}}) \\ \widehat{\boldsymbol{m}} & = & S^*(\boldsymbol{Y} - X\widehat{\boldsymbol{\beta}}(h)) \end{cases}$$

$$\begin{array}{c} \text{centered} \\ \downarrow \end{array}$$

$$S^* = (I - \frac{\mathbf{1}\mathbf{1}^T}{n})S$$

$$\begin{array}{c} \uparrow \\ \text{local linear smoother matrix} \end{array}$$

$I = n \times n$ identity matrix

$\mathbf{1} = n \times 1$ vector of 1's

$S = (S_{ij}) \quad n \times n$ matrix with:

$S_{ij} = \frac{w_j^{(i)}}{\sum_{k=1}^n w_k^{(i)}}$

$w_j^{(i)} = K\left(\frac{Z_i - Z_j}{h}\right)[\mathcal{S}_{n,2}(Z_i) - (Z_i - Z_j)\mathcal{S}_{n,1}(Z_i)]$

$\mathcal{S}_{n,l}(Z_i) = \sum_{k=1}^n K\left(\frac{Z_i - Z_j}{h}\right)(Z_i - Z_k)^l, \ l = 1, 2$

Explicit expression of $\widehat{\boldsymbol{\beta}}(h)$:

$$\widehat{\boldsymbol{\beta}}(h) = \left(X^T(I - S^*)X\right)^{-1} X^T(I - S^*)\boldsymbol{Y}$$

Exact conditional bias and variance of $\widehat{\boldsymbol{\beta}}(h)$:

If

$$A = \left(X^T(I - S^*)X\right)^{-1} X^T(I - S^*)$$

then:

$$Bias(\widehat{\boldsymbol{\beta}}(h)|X) = A\boldsymbol{m}$$

$$Var(\widehat{\boldsymbol{\beta}}(h)|X) = \sigma^2 AA^T$$

Asymptotic conditional bias and variance of $\widehat{\beta}(h)$ when $h \to 0$ and $nh^3 \to \infty$:

$$Bias(\widehat{\boldsymbol{\beta}}(h)|X) = h^2 \cdot C^{-1}D + \boldsymbol{o}_p(h^2)$$
$$Var(\widehat{\boldsymbol{\beta}}(h)|X) = \frac{\sigma^2}{n+1} \cdot C^{-1} + \boldsymbol{o}_P\left(\frac{1}{n}\right)$$

where

$$C = \tilde{\Sigma} + \int_0^1 \tilde{\boldsymbol{g}}(z)f(z)dz \cdot \int_0^1 \tilde{\boldsymbol{g}}^T(z)f(z)dz$$

$$D = -\frac{\nu_2(K)}{2}\int_0^1 \tilde{\boldsymbol{g}}(z)m''(z)f(z)dz +$$

$$\frac{\nu_2(K)}{2}\int_0^1 \tilde{\boldsymbol{g}}(z)f(z)dz \cdot \int_0^1 m''(z)f(z)dz$$

and $\tilde{g}(z) = (g_1(z), \ldots, g_p(z))^T$, $\tilde{\Sigma}$ obtained from Σ by adding a first row and column of 0's

Asymptotic conditional mean square error of $\widehat{\beta}$:

$$MSE(\widehat{\boldsymbol{\beta}}|X) = \boldsymbol{\mathcal{O}}_P(h^4) + \boldsymbol{\mathcal{O}}_P\left(\frac{1}{n}\right)$$

What h would yield a $\sqrt{n}-$ consistent estimate of $\widehat{\boldsymbol{\beta}}$?

$\widehat{\boldsymbol{\beta}}$ is $\sqrt{n}-$ consistent if $MSE(\widehat{\boldsymbol{\beta}}|X) = \boldsymbol{\mathcal{O}}_P\left(\frac{1}{n}\right)$

so h must satisfy:

$$h^4 << \frac{1}{n}$$

$$nh^3 \to \infty$$

that is

$$h \sim n^r, \;\; r \in (-\frac{1}{3}, -\frac{1}{4}]$$

Note that $r_0 = -\frac{1}{5} \notin (-\frac{1}{3}, -\frac{1}{4}]$ where $h_0 \sim n^{r_0}$ is the optimal bandwidth for estimating m consistently in a model without β

In fact, $n^r << n^{r_0}$ for $r \in (-\frac{1}{3}, -\frac{1}{4}]$, which means that we have to *undersmooth* m when model includes β!

Partially Linear Model for Mexico City Data:

$$\begin{aligned} log(Death_i) &= \beta_0 + \beta_1 \cdot PM10 + \\ &\quad \beta_2 \cdot Temp + \beta_3 \cdot Hum + m(Day_i) + \\ &\quad \epsilon_i \end{aligned}$$

Data-driven choice of h:

- usually motivated by accurate prediction of response given the covariate values:

$$\widehat{h}_{CV} = argmin_h \frac{1}{n} \sum_{i=1}^{n} \left(log(Death_i)^{(-i)} - log(Death_i) \right)^2$$

- when temperature, humidity and time are a nuisance, choice should be motivated by accurate estimation of β_1:

$$\widehat{h}_{EBBS} = argmin_h \widehat{MSE} \left(\widehat{\beta_1}(h) \right)$$

Notes:

$$\widehat{h}_{EBBS} = argmin_h \left\{ \widehat{Bias}(\widehat{\beta_1}(h)|X)^2 + \widehat{Var}(\widehat{\beta_1}(h)|X) \right\}$$

where

- $\widehat{Bias}(\widehat{\beta_1}(h)|X)$ is estimated empirically making minimal use of asymptotic theory

- $\widehat{Var}(\widehat{\beta_1}(h)|X)$ is computed from exact formula, replacing σ by an estimate $\widehat{\sigma}$

Empirical conditional bias estimation

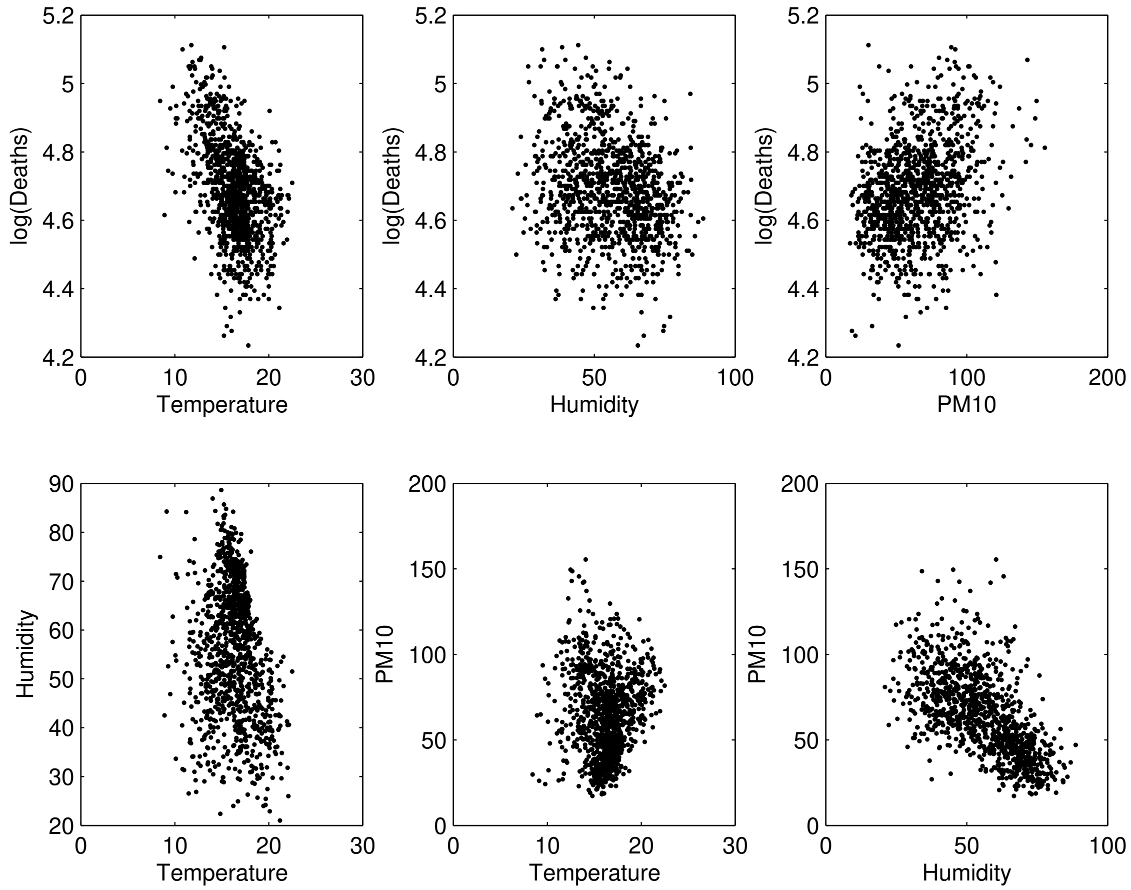
- compute $\widehat{\beta}_1(h)$ over a given range of values for h, namely $h_1, \ldots, h_M$

- to estimate $Bias(\widehat{\beta_1}(h_j)|X)$, fit the quadratic regression model
$$\widehat{\beta_1}(h) = b_0 + b_1 \cdot h^2 + \text{error}$$
to the data $(h_k, \widehat{\beta_1}(h_k))$, with $k \in \{j-j_1, \ldots, j-1, j, j+1, \ldots, j+j_1\}$

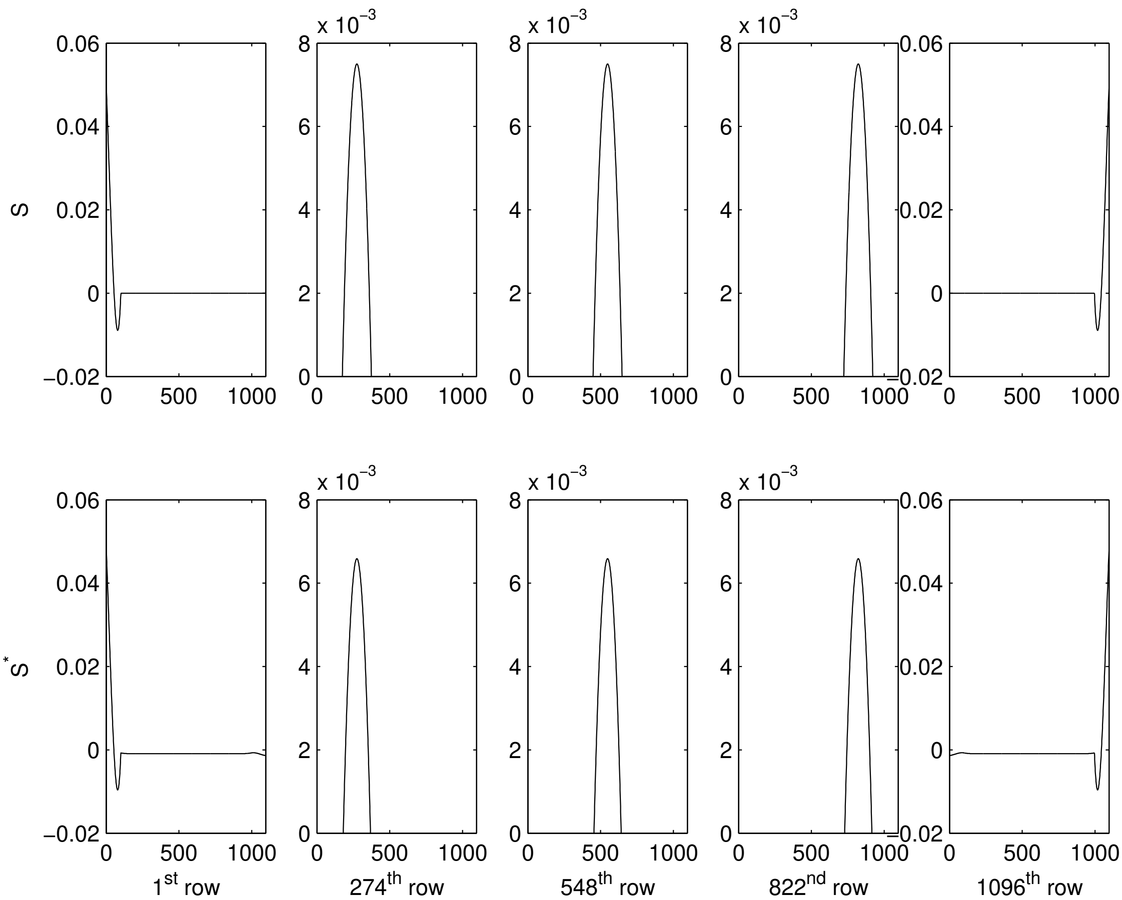
- set:
$$Bias\widehat{(\widehat{\beta_1}(h_j)}|X) = \widehat{b}_0 \cdot h^2$$

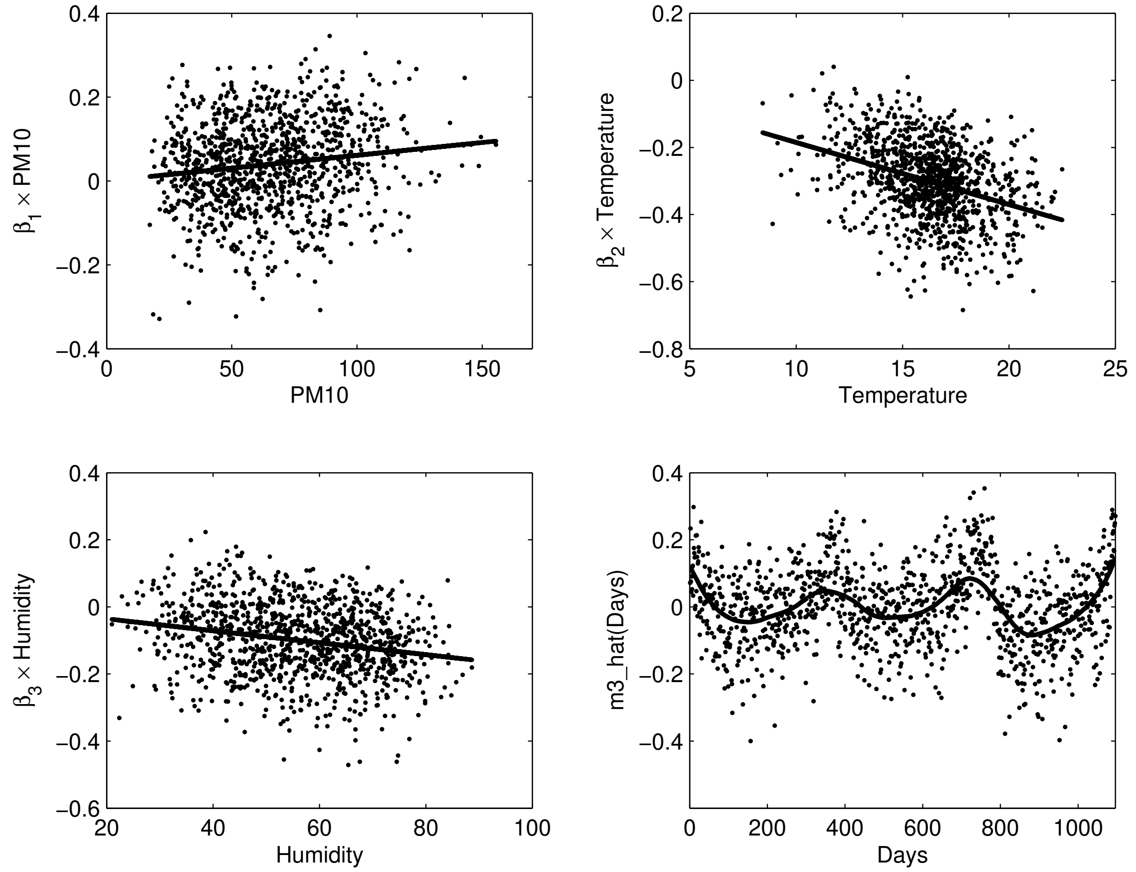
- j_1, j_2 tuning parameters

- could use $b_1 \cdot h^2 + \cdots b_t \cdot + h^t$ in the above with t being a tuning parameter

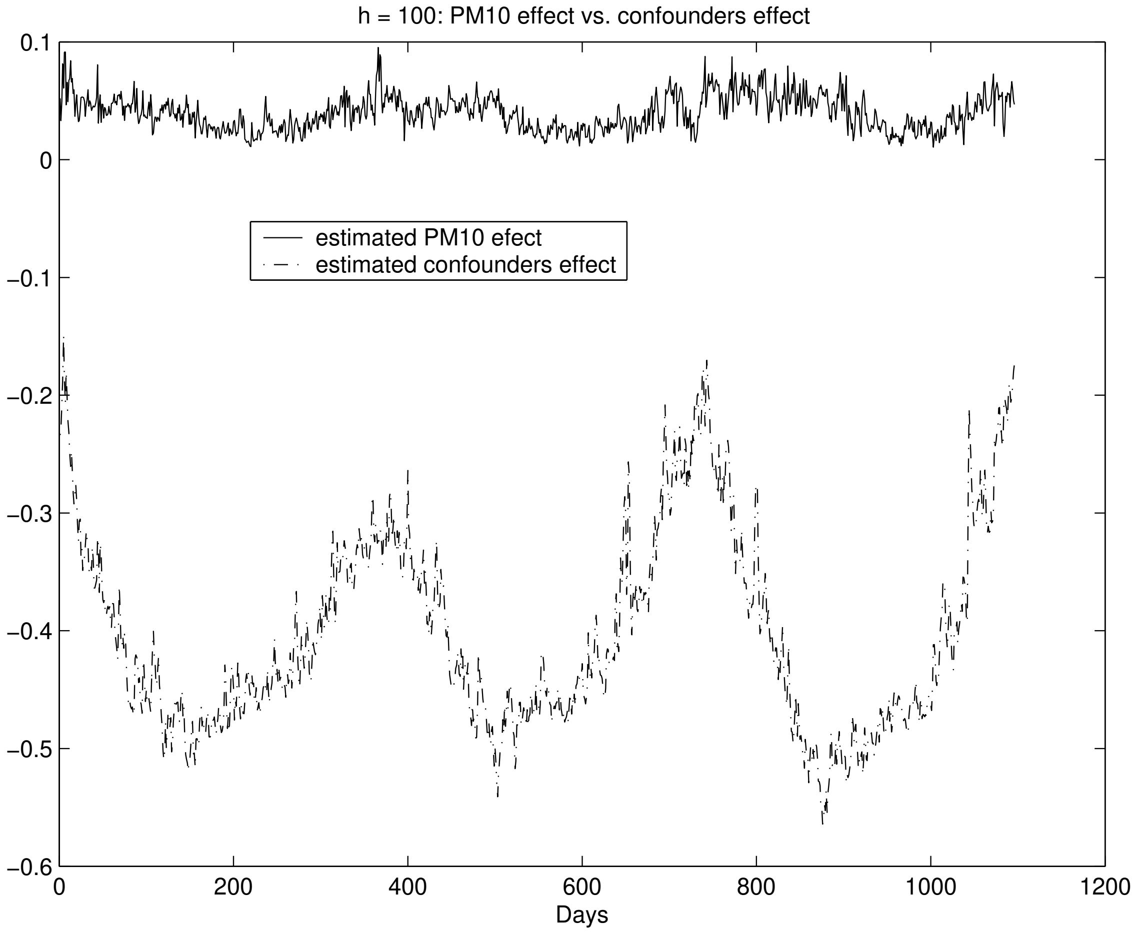


log(Deaths)
Temperature
Humidity
PM10

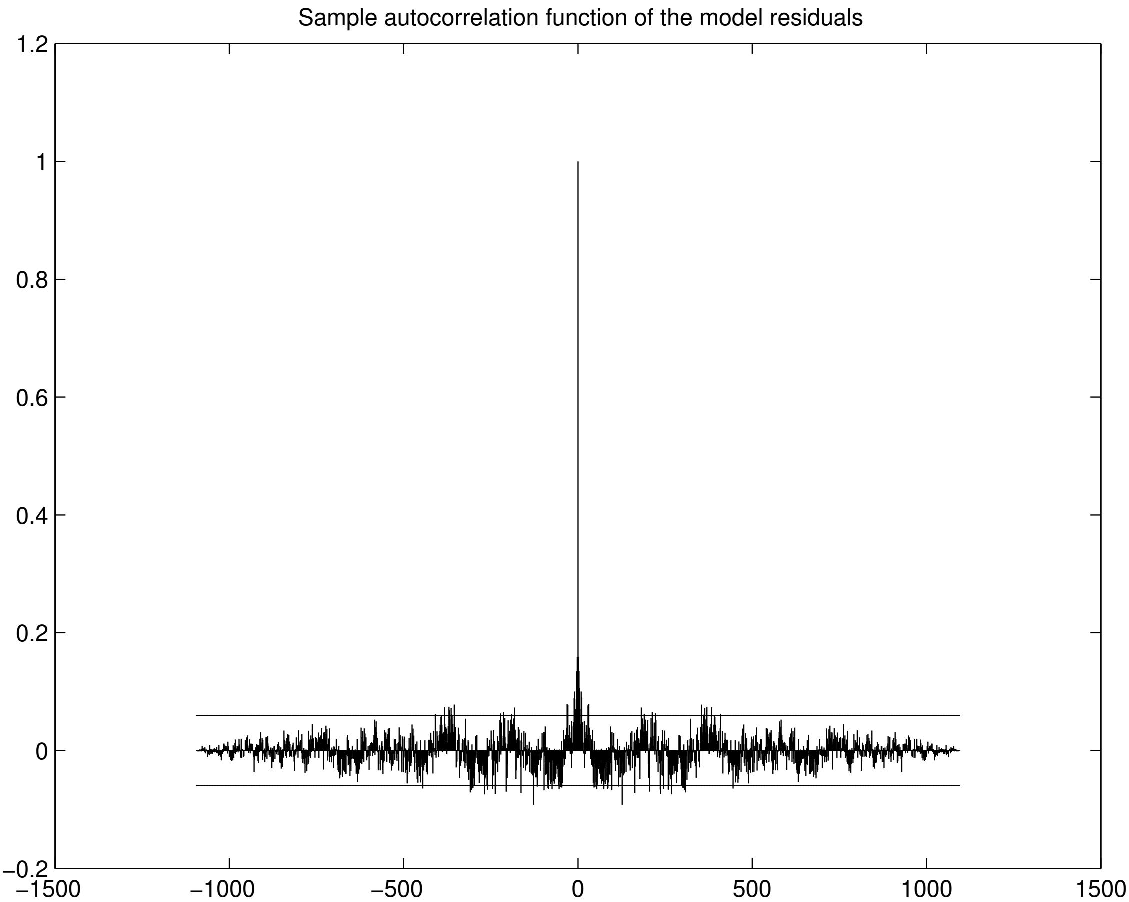

S
S*
x 10^-3
1st row
274th row
548th row
822nd row
1096th row

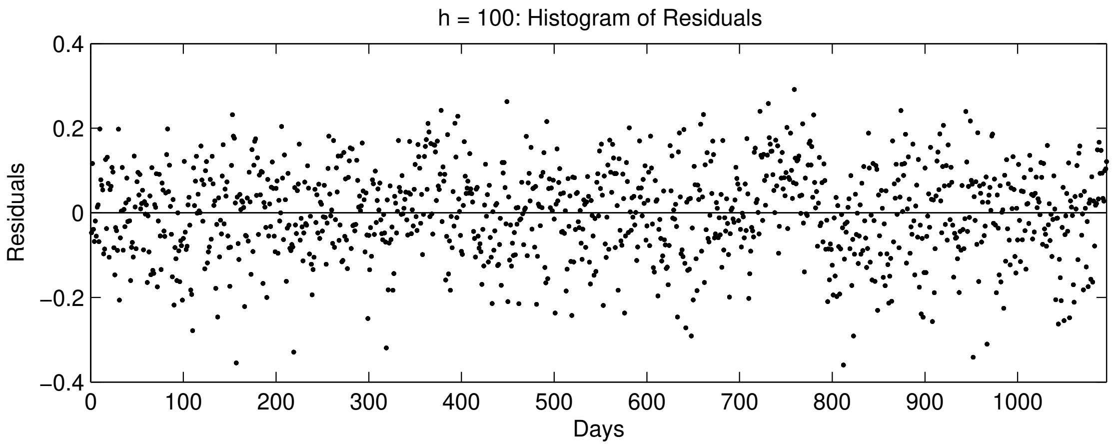


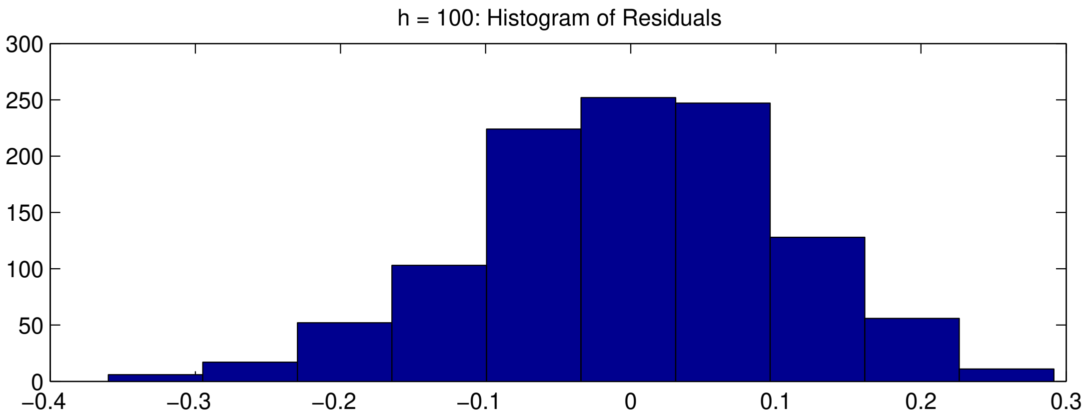
$\beta_1 \times$ PM10
PM10
$\beta_2 \times$ Temperature
Temperature
$\beta_3 \times$ Humidity
Humidity
m3_hat(Days)
Days

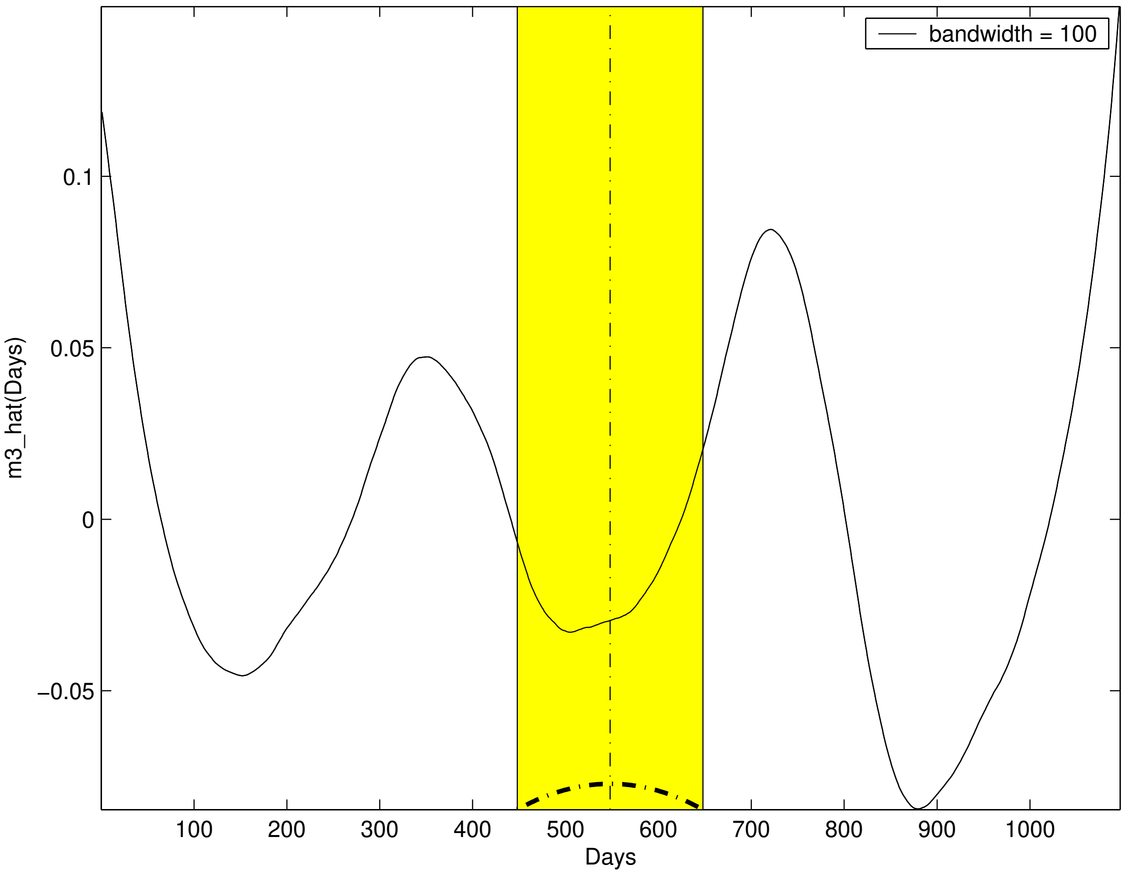


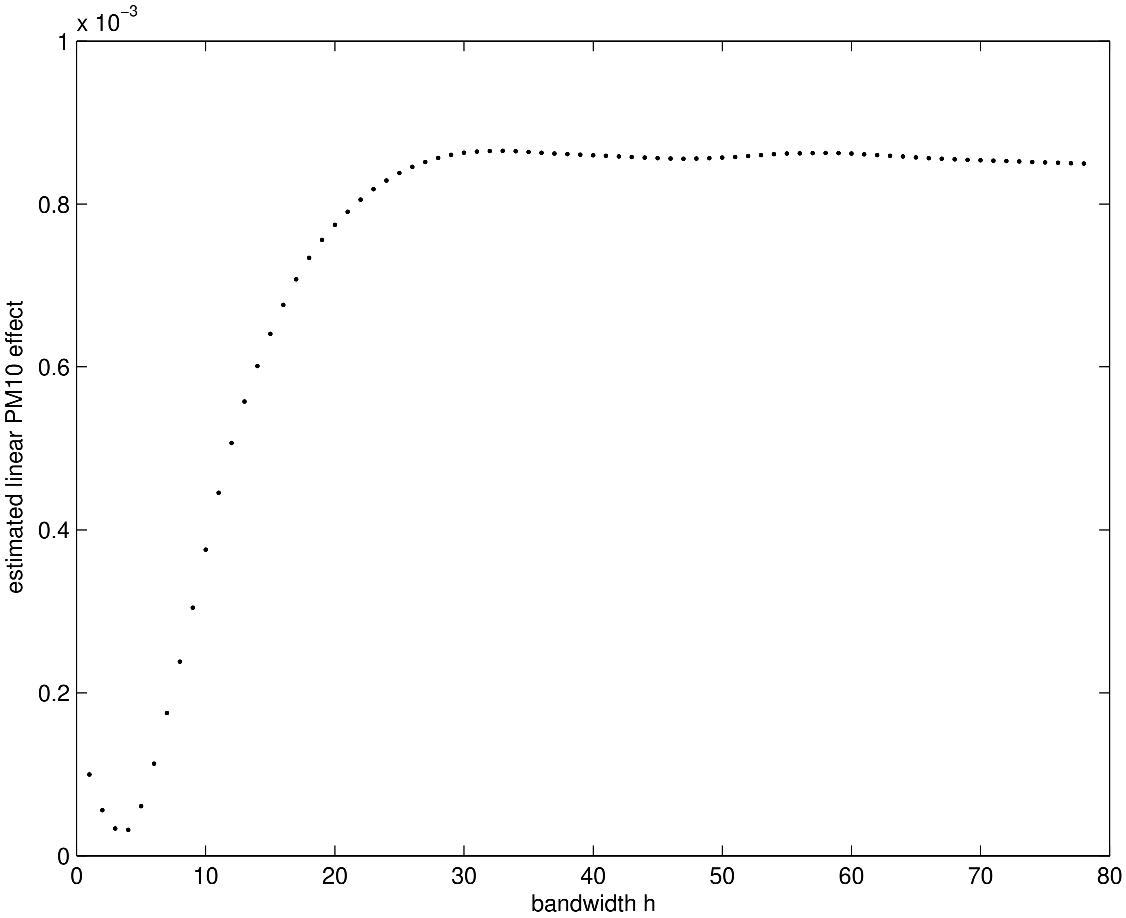
h = 100: PM10 effect vs. confounders effect
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
0
200
400
600
800
1000
1200
Days
estimated PM10 efect
estimated confounders effect

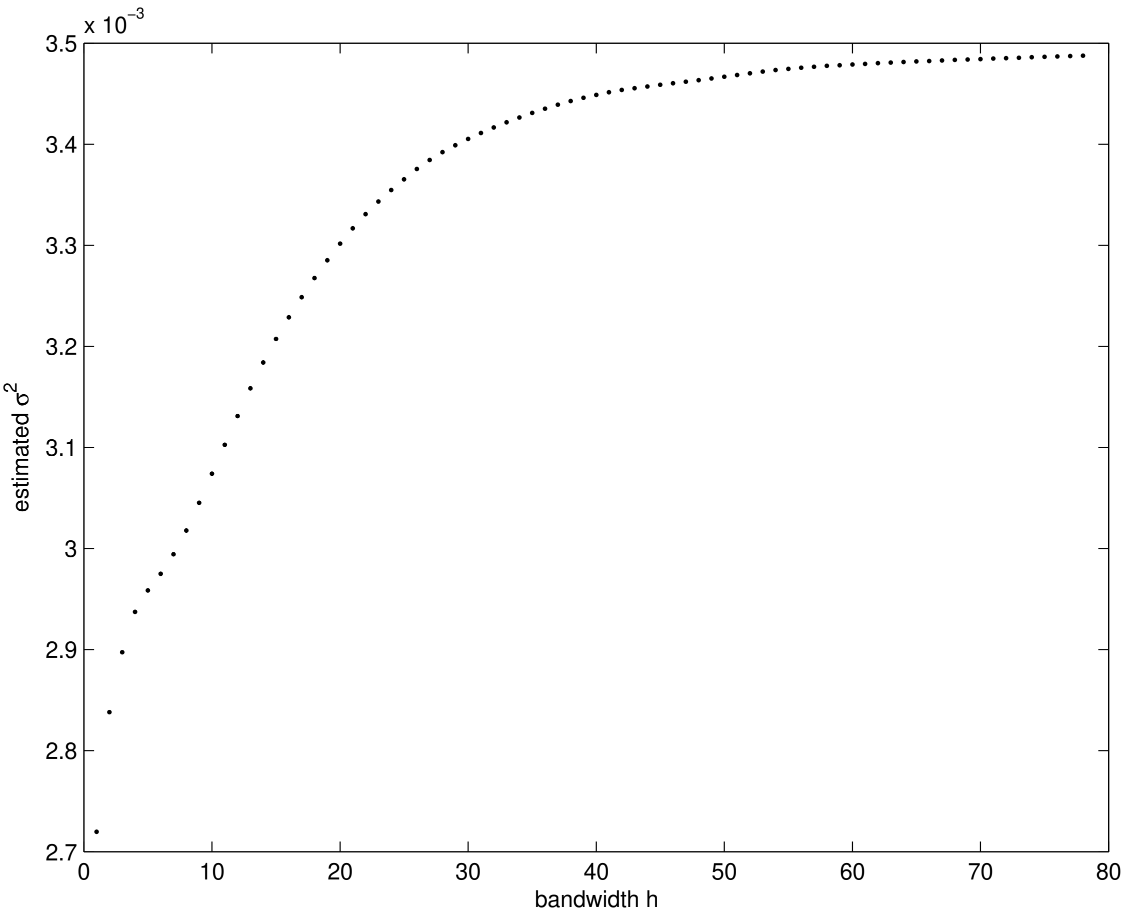

Sample autocorrelation function of the model residuals
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−1500
−1000
−500
0
500
1000
1500

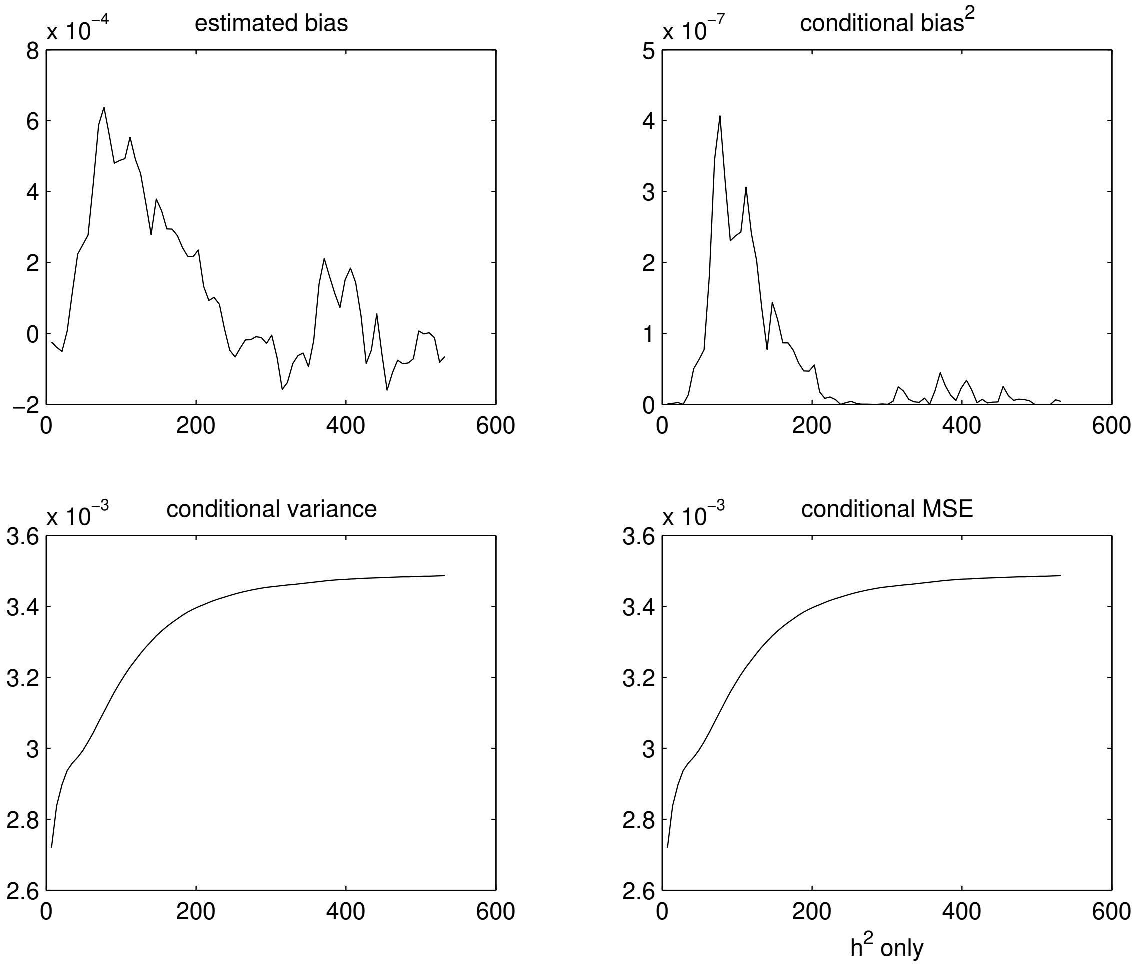

h = 100: Histogram of Residuals
0.4
0.2
0
−0.2
−0.4
Residuals
0
100
200
300
400
500
600
700
800
900
1000
Days

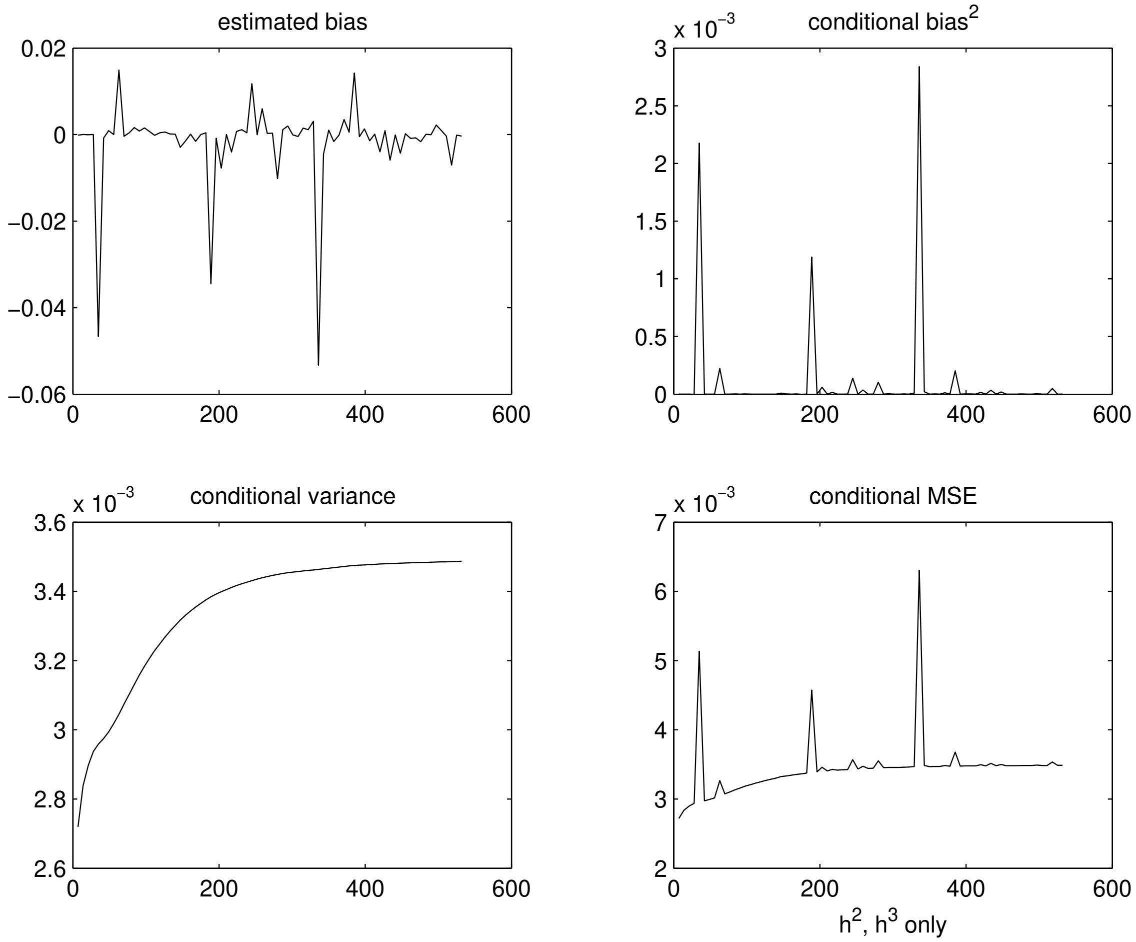

h = 100: Histogram of Residuals
300
250
200
150
100
50
0
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3

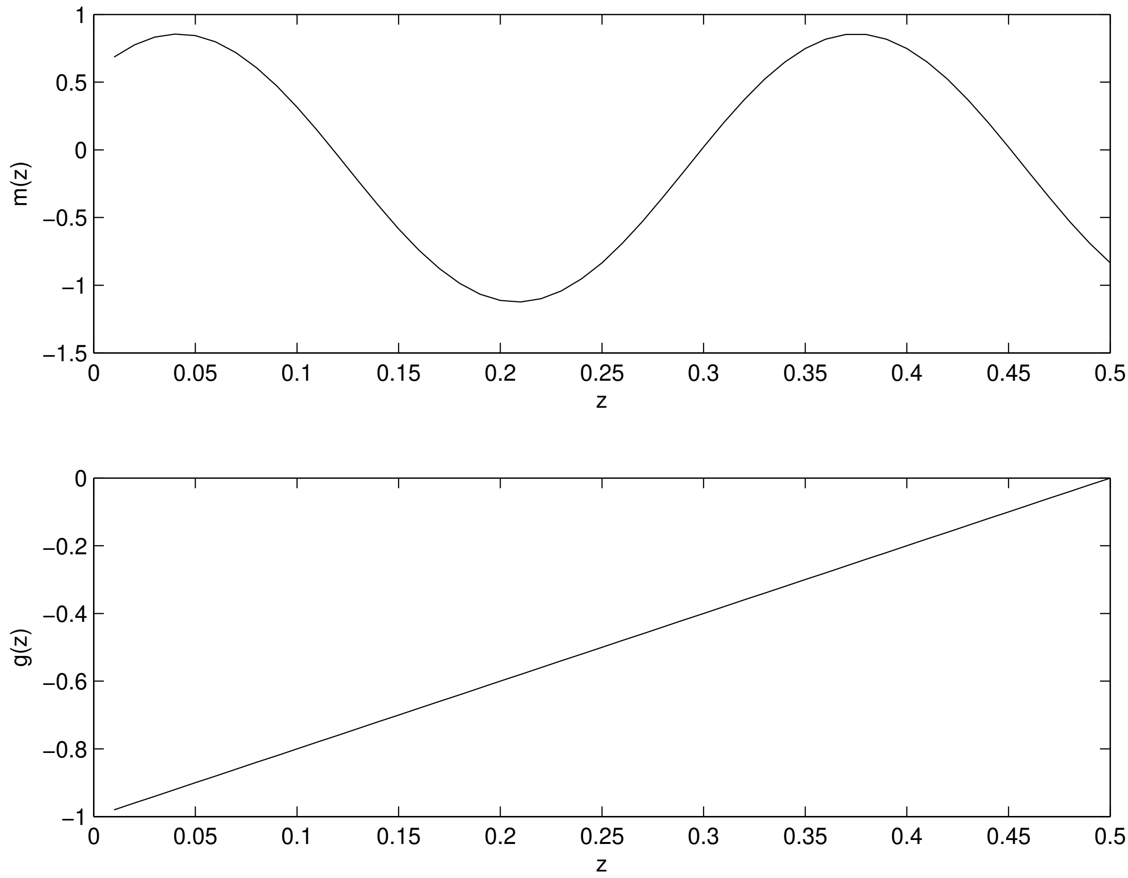

bandwidth = 100
0.1
0.05
0
−0.05
m3_hat(Days)
100
200
300
400
500
600
700
800
900
1000
Days


x 10^−3
1
0.8
0.6
0.4
0.2
0
estimated linear PM10 effect
0
10
20
30
40
50
60
70
80
bandwidth h


x 10−3
3.5
3.4
3.3
3.2
3.1
3
2.9
2.8
2.7
estimated σ2
0
10
20
30
40
50
60
70
80
bandwidth h


estimated bias
x 10^−4
conditional bias^2
x 10^−7
conditional variance
x 10^−3
conditional MSE
x 10^−3
h^2 only


estimated bias
0.02
0
−0.02
−0.04
−0.06
0
200
400
600
x 10−3
conditional bias2
3
2.5
2
1.5
1
0.5
0
0
200
400
600
x 10−3
conditional variance
3.6
3.4
3.2
3
2.8
2.6
0
200
400
600
x 10−3
conditional MSE
7
6
5
4
3
2
0
200
400
600
h2, h3 only


m(z)
z
g(z)
z

Asymptotic normality of $\widehat{\boldsymbol{\beta}}$:

Under certain regularity assumptions:

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \rightarrow \mathcal{N}\left(\mathbf{0}, \sigma^2 C^{-1}\right)$$

To test whether β_i significantly different from zero, use test statistic:

$$\frac{\widehat{\beta}_i(h)}{\widehat{SE}(\widehat{\beta}_i(h))}$$

with h chosen to guarantee the $\sqrt{n}-$ consistency of $\widehat{\beta}_i$.

Compare test statistic with appropriate critical value corresponding to standard normal distribution.