2004

# Anger expression as a predictor of outcome in emotion focused therapy with adult survivors of childhood abuse.

Marc F. Carriere
*University of Windsor*

# Analysis of Recurrent Event Data with Environmental Covariates

**By**
**Shahedul Ahsan Khan**

A Thesis

Submitted to the Faculty of Graduate Studies and Research

Through the Department of Mathematics and Statistics

in Partial Fulfillment of the Requirements for

the Degree of Master of Science at the

University of Windsor

Windsor, Ontario, Canada

2005

*Your file Votre référence*
*ISBN: 0-494-04980-4*
*Our file Notre référence*
*ISBN: 0-494-04980-4*

## Abstract

***Study Objectives*** – The objectives of this study were twofold. First, we used simulation to investigate three statistical methodologies for analyzing recurrent event data in the presence of environmental covariates to determine which procedure performs well under what situation. Secondly, we investigated the association between daily air pollution and hospital admissions of respiratory diseases by analyzing data from Vancouver, British Columbia.

***Settings and Study Population*** – Five air pollutants: carbon monoxide (CO), coefficient of haze (CoH), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$) and particulate matter 10 microns or less in diameter ($PM_{10}$) were considered in this study. The event of interest was daily respiratory hospital admissions of residents (65 + years) of Vancouver, B.C. from April 01, 1995 to March 31, 1999. The dates of hospital admission for each individual's repeated visits due to respiratory diseases were recorded.

***Statistical Methodologies*** – Three statistical methods were studied in this study, namely, Dewanji and Moolgavker's model (2000, 2002) based on a Poisson process assumption (*Model I*), Nividi's model (1998, 2002) for bidirectional case-crossover designs (*Model II*), and the usual time series analysis using a generalized linear model with natural splines (ns) to smooth time *(Model III)*. A simulation method was used to evaluate and compare these procedures. The mean square error (MSE) was the criterion used for evaluation.

***Results and Conclusions –***

***Simulation –*** We found that *Model III* performs better than *Model II* in terms of MSE, especially when the baseline intensities are different in different strata with strata being formed by partitioning the observation period. Although *Model III* cannot incorporate between subjects variation in the baseline parameters, it can produce results close to those of *Model I*, especially when the within stratum variation in the baseline intensities is small. An incorrect stratification model under *Model I* can produce serious bias estimate resulting in a large MSE. However, if the right stratification model can be selected, then *Model I* will give the best results. So unless we have a clear perception on the nature of the baseline intensities, we should not use *Model I*.

***Analysis Results –*** The results of *Model I* and *Model III* were similar with *Model II* being a bit different. *Model II* was based on constant baseline intensities with respect to time, which might not be the case for our data. *Model I* showed significant effects of all the pollutants to respiratory admissions. The relative risk (RR) estimates of the pollutants CO, CoH, $NO_2$, $SO_2$ and $PM_{10}$ were 1.04, 1.06, 1.07, 1.021 and 1.037, respectively. *Model III* gave quite similar type of results and conclusions except for $SO_2$ not being significant. We found *Model I* to be better than the other two models because it gave a shorter confidence interval for the RR parameters.

# Acknowledgements

I wish to express my sincere gratitude and appreciation to my supervisor Dr. K. Fung for her guidance, helpful supervision, constant support, and great patience throughout my graduate study and during the preparation of this thesis. Without the computer facilities that she provided me, the simulation part of this thesis could not have been done in such a short time. I would also like to thank Dr. M. Hlynka, Department of Mathematics and Statistics, and Dr. C. Ezeife, Department of Computer Science, for serving on my advisory committee.

I am grateful to Dr. D. Krewski, University of Ottawa, for providing the B.C. data.

I am grateful to the Department of Mathematics and Statistics for providing financial support in terms of a graduate assistantship throughout my studies. The research assistantship from Dr. K. Fung is also appreciated.

I would also like to thank all my friends for their encouragement and support.

I express my gratitude to my entire family. Finally, I express my very special thanks to my wife Rifat Jahan Khan for her encouragement and being exceptionally caring and supportive.

# Contents

## List of Tables

## List of Figures

# Chapter 1

## *Introduction*

Air pollution is a common threat to our health system. Air pollution refers to degradation of air quality resulting from unwanted chemicals or other materials occurring in the air. The so-called pure air is a mixture of gasses containing about 78% nitrogen, 21% oxygen, less than 1% of carbon monoxide, argon, and other gases, and varying amounts of water vapor (Source: http://www.arb.ca.gov/html/gloss.htm). Amounts of foreign and/or natural substances occurring in the atmosphere that may result in adverse effects to humans, animals, vegetation, and/or materials are called air pollutants. Some pollutants affect only small regional areas, while others can travel a long distance. Air pollutants are generally associated with the formation of smog. Smog is a combination of smoke and particulates, ozone, hydrocarbons, nitrogen oxides, and other chemical reactive compounds, which under certain conditions of weather and sunlight may result in a murky haze that causes adverse health effects. The principal greenhouse gases believed to contribute to global climate change are carbon dioxide, methane, and nitrous oxide. (Source: http://www.arb.ca.gov/html/gloss.htm).

Thousands of sources release air pollutants. These sources can be everything from large-scale industrial operations to the service station down the street to our household furnace. Sources of air pollutants can be broadly classified into three categories: mobile sources, area sources and point sources. Mobile sources of air pollutants include light-duty vehicles such as passenger cars and trucks, heavy-duty vehicles, buses, trains, airplanes, ships, agricultural and construction equipment and other sources such as off-

road vehicles. Area sources are generally small emission sources that, when their combined effects are added up, become significant. Burning natural gas in home furnaces, refueling vehicles with gasoline at a service station, farming operations and solvent use are just some activities that are included in this category. Point sources are large industrial and commercial sources that individually emit a significant amount of air pollutants. Oil refineries, sawmills and cement plants are examples of point sources (Source: http://www.gvrd.bc.ca/air/emissions.htm).

The science of air pollution is complex and evolving. Becoming informed is a key to citizens becoming effective players in preventing air pollution. Canada is the second highest per capita emitter of greenhouse gas in the industrialized world (Source: http://www.cleanair.ca/science_facts.html). Over 3 million Canadians must cope with serious respiratory diseases - asthma, chronic obstructive pulmonary disease (COPD), lung cancer, influenza and pneumonia, bronchiolitis, tuberculosis (TB), cystic fibrosis, and respiratory distress syndrome (RDS). These diseases affect people of all ages. Since many of these diseases affect adults over the age of 65, the number of people with respiratory diseases increases as the population ages. The quality of indoor and outdoor air contributes significantly to the exacerbation of symptoms of respiratory diseases. Air quality issues are dependent on geography and solutions vary according to locale. The corresponding increase in demand for services poses a significant challenge for the health care system. Respiratory diseases, including lung cancer, exert a great economic impact on the Canadian health care system. They account for nearly 12.18 billion dollars of expenditures per year (1993 dollars). These costs include the direct or visible costs of health care, such as hospitalization, physician visits and drugs (over 3.79 billion dollars).

They also include the less visible or indirect expenses associated with disability and mortality, which may be even more significant (8.39 billion dollars). (Source: http://www.phac-aspc.gc.ca/publicat/rdc-mrc01/pdf/rdc0901e.pdf). It becomes a key issue to address the extent of adverse effects of air pollutants to our health system. This motivated us to determine the relationship between exposures to air pollutants and hospital admissions due to respiratory diseases.

## 1.1 Pollutants to be Addressed

The pollutants investigated in this study were carbon monoxide (CO), coefficient of haze (CoH), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$) and particles that are less than 10 microns (10$\mu$m) in diameter ($PM_{10}$).

Carbon monoxide (CO) is an odorless, tasteless, colorless gas, which is primarily released by the incomplete combustion (the act of instance of burning some type of fuel such as gasoline to produce energy) of fossil fuels. Over 80% of the CO emitted in urban areas is contributed by motor vehicles. CO interferes with the blood's ability to carry oxygen to the body's tissues and results in numerous adverse health effects. In low doses, CO slows reflexes and reduces perception. In very high concentrations, it can cause unconsciousness and death. (Sources: http://www.arb.ca.gov/html/gloss.htm, http://www.lung.ca/cando/pollutants.html).

Coefficient of haze (CoH) is a measurement of the quantity of dust and smoke in the atmosphere in a theoretical 1000 linear feet of air. Sources of dust and smoke include motor vehicle emissions, industrial emissions, road dust, dust resulting from other agriculture, and smoke from forest fires (Burnett et al. 1997b). A CoH of less than three is considered clean air and more than five is of some concern. CoH readings of 20 or

more can occur in urban areas (Source: http://www.arb.ca.gov/html/gloss.htm). Goldberg et al. (2001b) found in Montreal that coefficient of haze was associated with mortality from respiratory disease and diabetes.

Nitric oxide (NO) is precursor of ozone, nitrogen dioxide ($NO_2$) and nitrate. NO is converted to $NO_2$ in the atmosphere, and then becomes involved in photochemical processes and/or particulate formation. $NO_2$ is a reddish-brown gas and includes a number of gases that are composed of oxygen and nitrogen. It is typically created during combustion processes, and is the major contributor to smog formation and acid deposition. Other natural sources include forest fires, lightning and decaying vegetation. The nitrogen oxides family of gases can be transported long distances in our atmosphere. $NO_2$ is a criteria air pollutant, and may result in numerous adverse health effects. At elevated levels, $NO_2$ can impair lung function, irritate the respiratory system and, at very high levels, make breathing difficult, especially for people who already suffer from asthma or bronchitis. (Sources: http://www.arb.ca.gov/html/gloss.htm, http://www.lung.ca/cando/pollutants.html, http://www.hc-sc.gc.ca/hecs-sesc/air_quality/talk.htm#nitrogen).

Sulfur dioxide ($SO_2$) is a strong smelling, colorless gas that is formed by the combustion of fossil fuels. Primary sources include fossil fuels in petroleum refineries, pulp and paper mills, electricity generating plants, and the smelting process in metal refineries. Natural sources include volcanoes. Power plants, which may use coal or oil high in sulfur content, can be major sources of $SO_2$. $SO_2$ and other sulfur oxides contribute to the problem of acid deposition. It is a criteria air pollutant. $SO_2$ is a naturally occurring substance that becomes problematic at higher concentrations. It can irritate the upper respiratory tract in humans, and it can lead to acid rain. (Sources:

http://www.arb.ca.gov/html/gloss.htm, http://www.lung.ca/cando/pollutants.html, http://www.hc-sc.gc.ca/hecs-sesc/air_quality/talk.htm#sulphur).

Airborne particles are known as particulate matter (PM) or simply particles. These particles are very small solids and/or liquids that are produced by a variety of natural and man-made sources. Particulate matter can also be formed in the air from the chemical transformation of gases such as sulfur dioxide, nitrogen oxides and various hydrocarbons. The size of particles may range from 0.005 microns ($\mu$m) to 100$\mu$m in diameter. $PM_{10}$ is a criteria air pollutant that is 10$\mu$m or less in diameter. The finer particles pose the greatest threat to human health because they can travel deepest into the lungs. Particles are also an important component of smog. Short-term exposure to airborne particles at the levels typically found in urban areas in North America is associated with a variety of adverse effects. Particulates can irritate the eyes, nose and throat and cause coughing, breathing difficulties, reduced lung function and an increased use of asthma medication. Other hazardous pollutants may also adhere to these particles, increasing their toxicity. Exposure to particulates is also associated with an increase in the number of emergency department visits, an increase in hospitalizations of people with cardiac and respiratory disease and in premature mortality. (Sources: http://www.arb.ca.gov/html/gloss.htm, http://www.lung.ca/cando/pollutants.html, http://www.hc-sc.gc.ca/hecs-sesc/air_quality/talk.htm# airborne).

## 1.2 Effects of Air Pollutants on Human Health: Review from Literature

The human health effects of poor air quality are far reaching, but principally affect the body's respiratory system and the cardiovascular system. Individual reactions to air pollutants depend on the type of pollutant a person is exposed to, the degree of exposure,

the individual's health status and genetics. The health effects caused by air pollutants may range from subtle biochemical and physiological changes to difficulty breathing, wheezing, coughing and aggravation of existing respiratory and cardiac conditions. These effects can result in increased medication use, increased doctor or emergency room visits, more hospital admissions and even premature death.

Several studies were conducted to evaluate the effects of air pollutants on mortality. Goldberg et al. (2001a) found evidence of associations between daily non-accidental deaths and most measures of particulate air pollution in Montreal. Neas et al. (1999) found that there was an association between air pollutants and daily mortality in Philadelphia. Kenney et al. (1991) found that mortality was associated with motor-vehicle pollutants in Los Angeles. Zmirou et al. (1996) observed association between $SO_2$ and particulates with daily mortality in Lyon, France. Wojtyniak et al. (1996) analyzed the effects of air pollutants on mortality on four Polish cities and reported that short-term effects of air pollution on mortality in Polish urban population cannot be ignored. Dab et al. (1996) found significant relation between $SO_2$ and particulates with daily mortality in Paris, France. In Milan, Italy, Vigotti et al. (1996) observed significant association between $SO_2$ and particulates with daily mortality.

Numerous studies assessed the association between air pollution and hospital admissions or emergency room visits for respiratory and cardiovascular diseases. Respiratory and cardiovascular diseases are among the leading causes of hospitalization in Canada. In 1996-1997 there were 3.16 million hospital admissions in Canada of which cardiovascular and respiratory diseases accounted for 15% and 9%, respectively. Air pollution exacerbates the condition of people with respiratory and cardiovascular diseases

and causes measurable increases in the rates of hospitalization for these diseases. Figure 1.1 shows the leading causes of hospitalization in Canada during 1996-1997. (Source: http://www.hc-sc.gc.ca/hecs-sesc/air_quality/health_effects.htm#7).

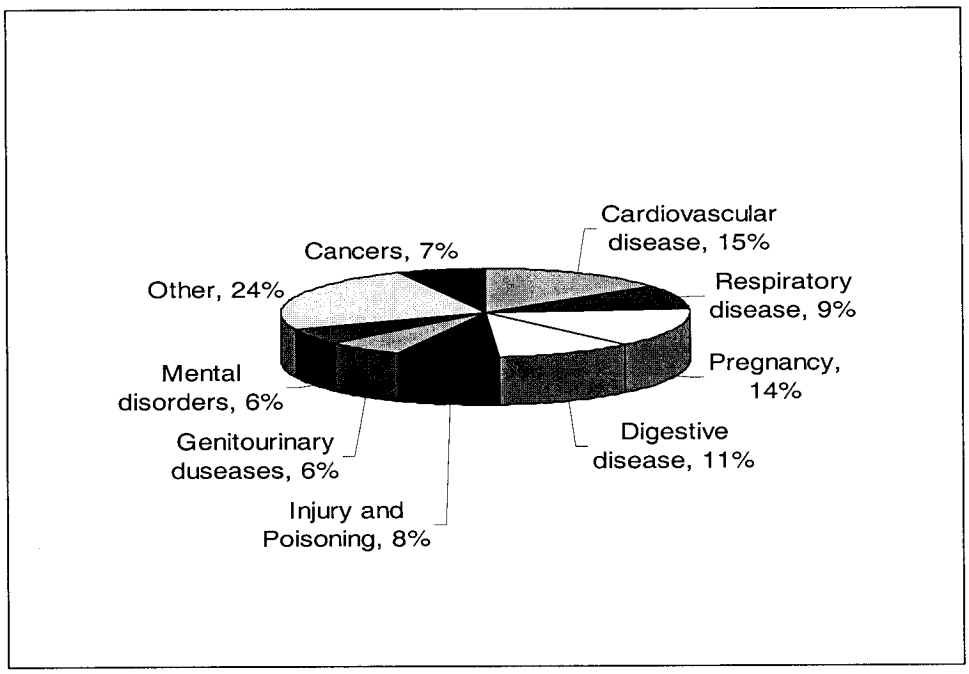



**Figure 1.1:** Leading causes of hospitalization in Canada during 1996-1997 (Total number of hospitalizations: 8,168,799, Source: Statistics Canada, 1999).

Research using large-scale data sets has shown a fairly consistent relationship between air pollutant levels and respiratory diseases in a variety of communities in the industrialized world (Atkinson et al. 1999; Dockery et al. 1993; Pope et al. 1995; Schwartz 1994). Schwartz et al. (1995) found that gaseous pollutants to be harmful to young children, the elderly and those suffering from respiratory or cardiac diseases in Detroit, Michigan. Dewanji and Moolgavkar (2000) found significant association between air pollutants and hospital admissions for chronic respiratory diseases in King County, Seattle, Washington.

In Canada, several reports were published linking air pollution to adverse population health in cities based on data that were collected in the 1980's and early

1990's (e.g. Burnett et al. 1994b, 1999). Lipfert et al. (1992) observed an association between gaseous pollutants and respiratory hospital admissions in Ontario. Burnett et al. (1997a) reported significant correlation between ozone and respiratory hospitalizations in 16 Canadian cities. Burnett et al. (1995) also observed significant association between ambient particulate sulfate and admissions to Ontario hospitals for cardiac and respiratory diseases. Lin et al. (2002) found positive relations between gaseous pollutants (CO, $SO_2$ and $NO_2$) at comparatively low levels and asthma hospitalization in children in Toronto, Ontario. Lin et al. (2004) found that $NO_2$ and $SO_2$ were significantly related to Asthma hospitalization among children with low household income in Vancouver, B.C. Luginaah et al. (2004) found associations between $NO_2$, $SO_2$, CO, CoH, and $PM_{10}$ and daily hospital admission of respiratory diseases especially among females in Windsor, Ontario. Fung at al. (2004) observed that short-term effects of $SO_2$ were associated significantly to daily cardiac hospital admissions for people $\geq$ 65 years of age living in Windsor, Ontario.

## 1.3 The Study Context: Vancouver, B.C.

The present study addresses the issues of air pollution and hospital admission due to respiratory diseases in Vancouver, B.C. The B.C. respiratory admission data consisted of patients, aged 65 or above, who were admitted to the hospital for more than one time (recurrent event data). The study period ranged from April 01, 1995 to March 31, 1999. In 1997, the age-standardized death rates per 100,000 populations due to all respiratory diseases were 62.7 for both sexes, 83.5 for males and 49.4 for females in B.C. On the other hand, these figures for the whole of Canada were 61.0, 87.0 and 45.7. (Source: http://www.statcan.ca/english/freepub/82-221-XIE/00604/tables/html/14153_97.htm). However, Air quality in greater Vancouver is considered to be good and usually meets

federal objectives for main air pollutants. (Source: http://www.gvrd.bc.ca/air/index.htm) For most common air contaminants, the air quality is better than it was 20 years ago, even with the growth in population and economic activity.

## 1.4 Recurrent Event Data

Recurrent events occur frequently in studies in which failures are not necessarily fatal (Kalbfleish and Prentice, 1980, pp. 179-182). Repeated hospital admission is an example. Many regression approaches involved modeling the intensity rate of the point process as a function of covariates or covariate processes (e.g. Prentice et al. 1981). Robust regression analyses of recurrent event data model some other marginal quantities instead of modeling the intensity rates (Wei et al. 1989, Pepe and Chi 1993, Lawless and Nadeau 1995), thus avoiding the strong assumption of recurrent event processes. These approaches can take into account subject specific covariates, but fail for environmental covariates. The usual time series analysis cannot take into account between subject variations in the baseline parameters and cannot incorporate subject specific covariates (Dewanji and Moolgavkar 2000). Dewanji and Moolgavkar (2000) proposed a model for recurrent event data. In their model, the intensity rate was modeled as a function of covariates. This model was also capable of incorporating subject specific covariates and also the previous history of the process.

## 1.5 Objectives of the Study

The main objectives of this study were twofold: (1) evaluation of several statistical methodologies, and (2) analysis of B.C. respiratory hospitalization data (recurrent event data) using these methodologies.

First, three statistical techniques will be examined: Dewanji and Moolgavkar's (2000) model based on the Poisson process, a conditional logistic regression model (Navidi, 1998) for bidirectional case-crossover designs, and the usual time series analysis. Using simulation, we will study how Dewanji and Moolgavkar's method performs in analyzing recurrent event data. Then we will compare Dewanji and Moolgavkar's method with Navidi's method and the usual time series analysis so that we can determine which performs well under what situation.

In the second part of this thesis, we will investigate the trends of respiratory hospital admissions for adults (65 + years) and pollutants in Vancouver, B.C., from 1995 to 1999. Then we will evaluate the extent of the association between air pollutants and respiratory hospital admissions of people $\geq 65$ of age.

This thesis is organized into 4 chapters. Chapter 2 describes the three statistical techniques of interest. Chapter 3 concerns the evaluation of Dewanji and Moolgavkar's model and a comparison among the three models of interest by simulation. Chapter 4 is devoted to the analysis of the Vancouver data set.

# Chapter 2

## *Statistical Methods*

Ordinary time series modeling was used by many authors to find the association between air pollutants and health events (Burnett et al. 1994a, 1995, 1997b, and Cakmak et al. 1998). Dominic et al. (2002) proposed the use of a generalized linear model with a cubic natural spline to smooth time. This technique assumes that all responses are independent. For the case-crossover design of Maclure (1991), Navidi (1998) proposed a multiple failure time model for environmental covariates. This model was based on conditional logistic regression and could be used for bidirectional case-crossover designs. Later, Navidi et al. (2002) discussed a semi-symmetric bidirectional design to restrict the set of eligible control times in case-crossover studies. Dewanji and Moolgavkar (2000) developed a recurrent event model based on the Poisson process assumption. Then, they extended the model by incorporating time dependent baseline intensities. They also proposed (2002) a cross-validation technique to choose the optimal model from different stratification models. In this study, we compared these models using simulation and we also used them in analyzing the Vancouver data. For convenience, we termed these three models as *Model I* (Dewanji and Moolgavkar 2000, 2002), *Model II* (Navidi 1998, 2002) and *Model III* (ordinary time series), respectively. This Chapter is mainly devoted discussing the development and the inference procedures of these three models.

### 2.1 *Model I*

*Model I* is based on the Poisson process assumption. Let $\{N(t), 0 \leq t \leq \tau\}$ denote the point process, which is observed over the period $(0, \tau]$. The history of the process up

to time t is given by $H_t = \{N(s), s \leq t\}$, which is assumed to include information on the covariate process. The intensity of the process $\lambda(t, H_t)$ is then defined as

$$\lambda(t, H_t) = \lim_{dt \to 0} \left[ \frac{\Pr\{dN(t) = 1 \mid H_t\}}{dt} \right] \tag{2.1}$$

where dN(t) is the number of events over the small time interval [t, t + dt). The likelihood contribution from each subject (process) having d events at times $t_1 < t_2 < \ldots < t_d$ over the period $(0, \tau]$ is given by

$$\left[ \prod_{j=1}^{d} \lambda(t_j, H_{t_j}) \right] \times \exp\left[ - \int_0^{\tau} \lambda(t, H_t) dt \right]. \tag{2.2}$$

The idea of the approach is to model the intensity $\lambda(t, H_t)$ in such a way as to make inferences on the effects of the covariates on the occurrences of recurrent events without losing much flexibility in the baseline intensity. First, a non-homogeneous Poisson process model is defined assuming that the baseline intensities $\lambda_i$ vary from subject to subject but are independent of time and the relative risk parameter $\beta$ is constant with respect to all subjects. The intensity of the ith process (subject) is expressed as

$$\lambda(t, H_t) = \lambda(t, x_{it}) = \lambda_i \exp[x_{it}^T \beta] \tag{2.3}$$

where $x_{it}$ is the vector of covariates for the ith subject at time t. The likelihood function is obtained by substituting (2.3) in (2.2):

$$L(\lambda, \beta) = \prod_{i=1}^{n} \left\{ \frac{\lambda_i^{d_i} \prod_{j=1}^{d_i} \exp[x_{it_{ij}}^T \beta]}{\exp\left( \lambda_i \int_0^{\tau_i} \exp[x_{it}^T \beta] dt \right)} \right\} \tag{2.4}$$

where n is the total number of subjects with the ith subject being observed over the period $(0, \tau_i]$ and $d_i$ is the number of events observed for the ith subject at times $t_{i1} < t_{i2} < \ldots < t_{id_i}$, for i = 1, 2, …, n. The maximum likelihood estimates of $\lambda_i$ are

$$\hat{\lambda}_i(\beta) = \frac{d_i}{\int_0^{\tau_i} \exp[x_{it}^T \beta] dt}, \quad i = 1, 2, \ldots, n. \tag{2.5}$$

The profile likelihood of $\beta$ is obtained by substitution (2.5) in (2.4):

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\prod_{j=1}^{d_i} \exp[x_{it_{ij}}^T \beta]}{\left( \int_0^{\tau_i} \exp[x_{it}^T \beta] dt \right)^{d_i}} \right\}. \tag{2.6}$$

The maximum likelihood estimate of $\beta$ is obtained by maximizing equation (2.6). The likelihood basically compares the rate on the days for which events occur to the overall (integrated) rate, so the model looks to pick the $\beta$ which makes that ratio as large as possible.

Next, Dewanji and Moolgavkar extended the model by incorporating time-dependent baseline intensities. Let $\lambda_i$'s, as a function of time t, to be piecewise constant over the interval $(0, \tau_i]$ as follows:

$$\lambda_i(t) = \lambda_{il} \text{ for } t \in I_{il} = (\tau_{i,l-1}, \tau_{il}], \text{ for } l = 1, 2, \ldots, K_i, \tag{2.7}$$

with $0 = \tau_{i0} < \tau_{i1} < \ldots < \tau_{iK_i}$ being prespecified. Intervals $I_{il}$ are regarded as strata. In the formulation (2.7), it is assumed that the baseline intensity for an individual is different in different strata but is the same within a stratum. Because of the Poisson process assumption, the events in disjoint strata $I_{il}$ are independent. Given $d_{il}$, the number of

events for the ith subject in $I_{il}$, a conditional likelihood of the form given inside the braces in (2.6) corresponding to each of the strata $I_{il}$ is formed. The likelihood function is then obtained by taking product over all the (i, l)s to give

$$L(\beta) = \prod_{i=1}^{n} \prod_{l=1}^{K_i} \left\{ \frac{\prod_{j=1}^{d_{il}} \exp(x_{it_{ilj}}^T \beta)}{\left( \int_{I_{il}} \exp[x_{it}^T \beta] dt \right)^{d_{il}}} \right\} \tag{2.8}$$

where $t_{ilj}$ is the time for the jth event in $I_{il}$ for the ith subject. The likelihood function takes a simpler form when $\tau_i$ and $I_{il}$ are the same for all individuals:

$$L(\beta) = \prod_{l=1}^{K} \left\{ \frac{\prod_{i \in D_l} \exp(x_{it_{il}}^T \beta)}{\prod_{i \in D_l} \left( \int_{I_l} \exp[x_{it}^T \beta] dt \right)} \right\}, \tag{2.9}$$

where K is the number of strata and $D_l$ is the set of subjects having events (one subject may have more than one event) in the lth common stratum $I_l$ at times $t_{il}$s. For no subject-specific covariates, $x_{it}$ is replaced by $x_t$. In that case, (2.9) is reduced to

$$L(\beta) = \prod_{l=1}^{K} \left\{ \frac{\prod_{i \in D_l} \exp(x_{t_{il}}^T \beta)}{\left( \int_{I_l} \exp[x_t^T \beta] dt \right)^{d_l}} \right\}, \tag{2.10}$$

where $d_l = ||D_l||$ for $l = 1, 2, \ldots, K$ and $x_{t_{il}}$ is the value of covariates at time $t_{il}$.

One difficulty with the model for the time-dependent baseline intensities is in the choice of the partitions $I_l$s. The decision on the partitioning is crucial as the estimates of the parameters are very sensitive to the choices of the partitions. In the absence of any

prior knowledge, one has to make a difficult judgment on when the baseline intensities may be constant (Dewanji and Moolgavkar, 2002).

Dewanji and Moolgavkar (2002) proposed a cross-validation approach in choosing the optimal model from a few suitably chosen stratification models. Let s (s = 1, 2, ..., S) denote a fixed stratification model. The cross-validated log-likelihood (CVLL(s)) is

$$\mathrm{CVLL}(s) = \sum_{i=1}^{n} \log L_i^{(s_f)}(\hat{\beta}_{-i}^{(s)}) \tag{2.11}$$

where $L_i^{(s_f)}(\cdot)$ is the likelihood contribution from the ith subject corresponding to the finest stratification indexed by $s_f$ and $\hat{\beta}_{-i}^{(s)}$ is the estimate of $\beta$ obtained by removing the ith individual under the model indexed by s. The optimal model $s_0$ is such that $\mathrm{CVLL}(s_0) = \max_s \mathrm{CVLL}(s)$.

If the number of individuals is large, the leave-one-out cross-validation becomes computationally prohibitive. In that case, Dewanji and Moolgavkar suggested m-fold cross-validation with a suitable m, in which the set of individuals was divided into m random parts (referred to as cross-validation samples) with roughly equal sizes and then the above cross-validation method was applied on the m parts.

## 2.2 *Model II*

In case-crossover designs, data are analyzed by considering each subject as one stratum in a case-control study, where the cases are the failure times and the controls are the other times. Conditional logistic regression can be applied for inference purposes in such situations (Navidi, 1998). In standard case-crossover designs, the control times are all taken prior to failure. In bidirectional case-crossover designs, control times are taken both before and after an event.

### 2.2.1 Bidirectional Case-Crossover Design

Navidi (1998) extended the single failure time bidirectional case-crossover designs to the situation of multiple failure times. Let $t_1, t_2, \ldots, t_M$ denote the times at which the failures can occur. Also let $x_{ij}$ be the vector of covariates for the ith subject that is relevant to risk at time $t_j$ and $\beta$ represents a vector of parameters. The log odds of failure for subject i on time $t_j$ is given by

$$\log \frac{p_{ij}}{1-p_{ij}} = \lambda_i + \beta^T x_{ij} \tag{2.12}$$

where $p_{ij}$ is the unconditional probability of failure for subject i on day j and $\lambda_i$ is the baseline intensity for subject i. Rearranging (2.12), we get

$$p_{ij} = \frac{\exp[\lambda_i + \beta^T x_{ij}]}{1+\exp[\lambda_i + \beta^T x_{ij}]}. \tag{2.13}$$

Let $n_i$ be the number of failures for subject i, and $A_i$ be the set of times when those failures occur. Let $P(A_i)$ denotes the probability that the failures for the ith subject occur precisely at the times in the set $A_i$. Then $P(A_i)$ is expressed as

$$P(A_i) = \left(\prod_{t_j \in A_i} p_{ij}\right)\left(\prod_{t_k \notin A_i} (1-p_{ik})\right). \tag{2.14}$$

The probability that the failures occur precisely at those times in $A_i$ conditional on $n_i$ failures is

$$P(A_i \mid n_i) = \frac{\left(\prod_{t_j \in A_i} p_{ij}\right)\left(\prod_{t_k \notin A_i} (1-p_{ik})\right)}{\sum_{S \in D_{n_i}} \left(\prod_{t_j \in S} p_{ij}\right)\left(\prod_{t_k \notin S} (1-p_{ik})\right)}, \tag{2.15}$$

where $D_{n_i}$ is the collection of all sets of $n_i$ times. Substituting (2.13) in (2.15), we get

$$P(A_i \mid n_i) = \frac{\exp\left(\beta^T \sum_{t_j \in A_i} x_{ij}\right)}{\sum_{S \in D_{n_i}} \exp\left(\beta^T \sum_{t_k \in S} x_{ik}\right)}. \qquad (2.16)$$

Let $T_{iS} = \sum_{t_k \in S} x_{ik}$ for any set of times S. This is the sum of the covariates of subject i over the set S. Then (2.16) is simplified to

$$P(A_i \mid n_i) = \frac{\exp\left(\beta^T T_{iA_i}\right)}{\sum_{S \in D_{n_i}} \exp\left(\beta^T T_{iS}\right)}. \qquad (2.17)$$

Summing the logarithm of (2.17) over all subjects gives the likelihood function

$$L(\beta) = \sum_i \left[\beta^T T_{iA_i} - \log \sum_{S \in D_{n_i}} \exp(\beta^T T_{iS})\right]. \qquad (2.18)$$

The number of sets to be summed over in the denominator of (2.17) is $M!/[n_i!(M-n_i)!]$, where M is the total number of time units at risk. If M is sufficiently large compared to the subjects, an unconditional logistic regression can be used (Breslow and Day, 1990, pp. 248-250). Alternatively, the denominator can be approximated with the sum of a few randomly chosen terms (Langholz and Goldstein, 1997).

### 2.2.2 Semi-Symmetric Bidirectional Design

In symmetric bidirectional designs, two control times (before and after the index time) are selected for each case. The potential problem with restricting the set of eligible control times in case-crossover studies is that bias can occur from time trends in the exposure of interest (Navidi et al. 2002). With a slight modification in the sampling procedure, the bias in the symmetric bidirectional design can be eliminated, while retaining control of unmeasured seasonal confounding. According to Navidi et al. (2002),

instead of using both $T_{k-1}$ and $T_{k+1}$, where $T_k$ is the event time for a subject and $T_{k-1}$ and $T_{k+1}$ are the controls to be matched, we may select one of the two at random so that each risk set contains one case and control, and it is no longer possible to determine which of the two members is the case from knowledge of the pair. Navidi et al. (2002) called this method semi-symmetric bidirectional design. Semi-symmetric bidirectional designs are localizable designs in the sense that there exits an unbiased estimating equation restricted to the referent windows. Conditional logistic regression can be applied for inference purposes.

## 2.3 Time Series Analysis

The usual method for removing temporal trends in the time series analysis was to apply a non-parametric smoothed function (LOESS in S-plus) to time with span chosen such that the autocorrelation in the residuals is small. Then a generalized additive model (GAM) is fitted to the data adjusted for air pollutant, day of the week and a LOESS smooth function of time. Dominici et al. (2002) proposed a generalized linear model (GLM) with alternative parametric natural cubic splines (ns) smother instead of the LOESS (lo) function. This method requires the specification of the degrees of freedom defined as df = (number of knots per year) × (total number of years under study ) + 2 so that the autocorrelation in the residuals is largely removed. We first fit the log linear model

$$\ln E(y_t) = \beta_0 + \text{ns}(\text{time}, \text{df}) + (\text{day of the week}) \qquad (2.19)$$

using GLM for different degrees of freedom and pick that degrees of freedom for which time series of the logarithm of events is as close to white noise as possible. This is equivalent to picking the degrees of freedom for which the p-value of Bartlett's test is the

largest. Then each pollutant and the weather variables (temperature and relative humidity) were added into the model:

$$\ln E(y_t) = \beta_0 + \beta_1 x_t + \text{ns}(\text{time}, \text{df}) + (\text{weather variables}) + (\text{day of the week}). \quad (2.20)$$

The parameters can be estimated using GLM. Statistical software like S-plus, R have the GLM function.

# Chapter 3

*Simulation*

In this chapter, we evaluate the three models of analysis discussed in Chapter 2 using simulation. The models include the one based on the Poisson process assumption proposed by Dewanji and Moolgavkar (2000), the model for bidirectional case-crossover designs proposed by Navidi (1998), and the time series method using generalized linear model with natural splines (ns) to smooth the time trend. The performance of the Dewanji and Moolgavkar's models for different stratification schemes with different baseline intensities and with different combinations of pollutant and weather variables is investigated in detail. We also discuss the technique of choosing the optimal model by means of cross-validation suggested by a later paper of Dewanji and Moolgavkar (2002). In addition, an attempt is made to compare the three models mentioned above.

## 3.1 Exposure Data

The B.C. weather and pollutant data were used in our simulation. The data set consisted of four years of information (1461 days) on various pollutants and weather variables such as carbon monoxide (CO), coefficient of Haze (CoH), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$), particulate matter 10μm or less in diameter ($PM_{10}$), temperature and relative humidity. The study period ranged from April 01, 1995 to March, 31 1999. Our objective was to study the association of daily hospital admission with air pollutant. Since the mean daily CO is a highly significant variable to have an effect on hospital admissions for respiratory diseases, we selected the actual CO as the pollutant in our simulations. To ensure that our data were at least somewhat plausible, we

also used the actual data on daily mean temperature and relative humidity as covariates in the simulation.

## 3.2 Two Simulation Scenarios

Two simulation scenarios were considered in this study: *scenario I* and *scenario II*. In *scenario I*, the same set of variables was used to generate as well as to analyze data. In *scenario II*, the set of variables used to generate data differed from the set used to analyze. Figure 3.1 shows a flowchart of different options of data generation and analysis used in this study.

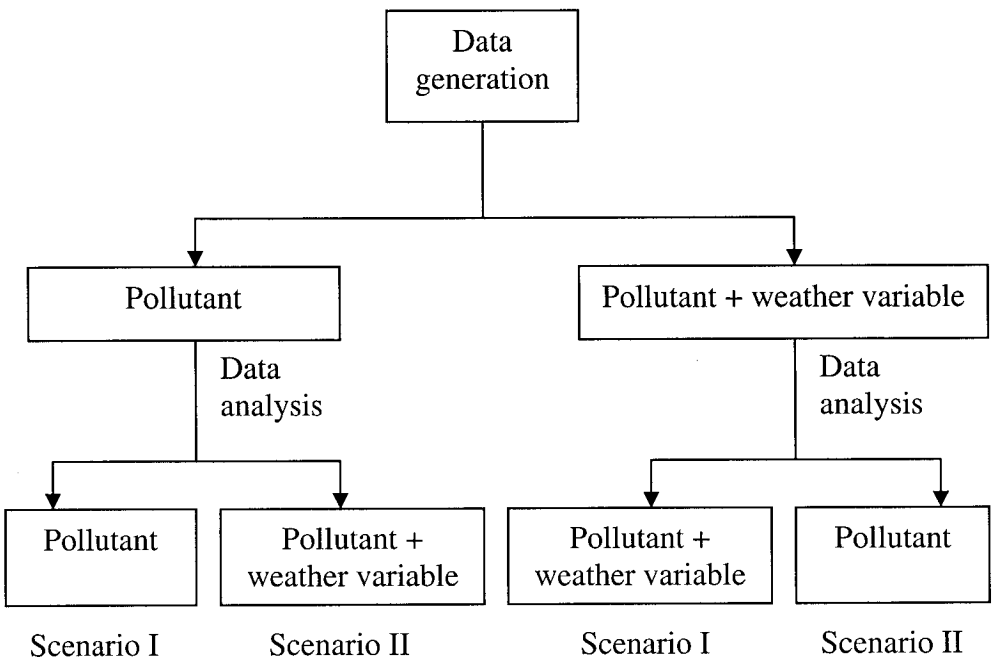



**Figure 3.1:** Flowchart for data generation and analysis options.

The rationale for defining these two scenarios is that in real life, it is not possible to have complete information on all the determinants that are responsible for the response variable. So, it is rare that we analyze the data with the same set of pollutant and

covariates which are responsible for the response. Three different situations are considered here:

(a) **the perfect model** – the data are analyzed using the same set of variables that are responsible for the response;

(b) **the under-specified model** – the data are analyzed using only a subset of variables; and

(c) **the over-specified model** – the data are analyzed using extra covariates which are not responsible for the response.

Situations (b) and (c) help us to evaluate the robustness of a model in terms of the covariates used.

## 3.3 Models

Three models that we considered in this study were the model based on the Poisson process assumption proposed by Dewanji and Moolgavkar, 2000 (*Model I*), the model for bidirectional case-crossover designs proposed by Navidi, 1998 (*Model II*) and the time series method using generalized linear model (*Model IIII*). In this study, we give more emphasis on Dewanji and Moolgavkar's model. The other two models were included for comparison purposes.

## 3.4 Data Generation

Data were generated under the Poisson process assumption. Four types of information were required to generate data:

1. number of strata in the stratification model;
2. baseline intensities;

3. pollutant and weather variable(s); and

4. the regression coefficient $\beta$.

In this simulation study, an eight strata model was selected as the true model and the data were generated from this model using different baseline intensities and exposure variables. The partitioning was made by dividing time into 8 intervals in such a way that each interval corresponds to six months. The strata were (in days): {[1, 183], [184, 366], [367, 549], [550, 731], [732, 914], [915, 1096], [1097, 1279], [1280, 1461]}.

Two types of baseline intensities ($\lambda_i$'s) were considered in this study: increasing $\lambda_i$'s and $\lambda_i$'s with the same distribution within each stratum. Table 3.1 gives the different $\lambda_i$'s from the uniform$[a_i, b_i]$ distribution that were used in generating admission data.

**Table 3.1:** Baseline intensities from the uniform$[a_i, b_i]$ distribution that were used to generate admission data.

| Stratum 1 $[a_1, b_1]$ | Stratum 2 $[a_2, b_2]$ | Stratum 3 $[a_3, b_3]$ | Stratum 4 $[a_4, b_4]$ | Stratum 5 $[a_5, b_5]$ | Stratum 6 $[a_6, b_6]$ | Stratum 7 $[a_7, b_7]$ | Stratum 8 $[a_8, b_8]$ |
|---|---|---|---|---|---|---|---|
| [0.01, 0.015] | [0.015, 0.02] | [0.02, 0.025] | [0.025, 0.03] | [0.03, 0.035] | [0.035, 0.04] | [0.04, 0.045] | [0.045, 0.05] |
| [0.01, 0.03] | [0.03, 0.05] | [0.05, 0.07] | [0.07, 0.09] | [0.09, 0.11] | [0.11, 0.13] | [0.13, 0.15] | [0.15, 0.17] |
| [0.00, 0.04] | [0.04, 0.08] | [0.08, 0.12] | [0.12, 0.16] | [0.16, 0.20] | [0.20, 0.24] | [0.24, 0.28] | [0.28, 0.32] |
| [0.01, 0.03] | [0.01, 0.03] | [0.01, 0.03] | [0.01, 0.03] | [0.01, 0.03] | [0.01, 0.03] | [0.01, 0.03] | [0.01, 0.03] |
| [0.00, 0.16] | [0.00, 0.16] | [0.00, 0.16] | [0.00, 0.16] | [0.00, 0.16] | [0.00, 0.16] | [0.00, 0.16] | [0.00, 0.16] |

We investigated two covariates – temperature and relative humidity (RH). As CO and temperature were negatively correlated ($r = -0.329$) and CO and RH were positively correlated ($r = 0.271$), we wanted to see the differences in performances of the models.

Different values for $\beta$ were chosen under different situations. First of all, when we considered only CO to generate data, we took $\beta = 0.05$. Secondly, when CO and temperature were used to generate data, three situations were considered: no dominance,

temperature dominates CO and CO dominates temperature. The means for CO and temperature were 0.68 and 10.75, respectively. Let $\beta_1$ and $\beta_2$ be the coefficients of CO and temperature, respectively and their chosen values are listed in Table 3.2. We chose $\beta_1 = 0.8$ and $\beta_2 = 0.05$ so that $\beta_1 \times$ CO (mean) = 0.54 and $\beta_2 \times$ temperature (mean) = 0.54, so there was no dominance over one of the other. Then we chose $\beta_1 = 0.1$ and $\beta_2 = 0.05$ so that $\beta_1 \times$ CO (mean) = 0.07 and $\beta_2 \times$ temperature (mean) = 0.54 and hence temperature dominated CO. Finally, we chose $\beta_1 = 0.8$ and $\beta_2 = 0.01$ so that $\beta_1 \times$ CO (mean) = 0.54 and $\beta_2 \times$ temperature (mean) = 0.11 and therefore CO dominated temperature. When CO and RH were used to generate data, the choice of the coefficients was also made so that we could achieve the three situations: no dominance, CO dominates RH and RH dominates CO. Table 3.3 gives different choices of the coefficients for CO and RH.

**Table 3.2:** Choice of the coefficients for CO and temperature to generate data.

| CO (mean = 0.68) | Temperature (mean = 10.75) | $\beta_1 \times$ CO (mean) | $\beta_2 \times$ temp (mean) | Comment |
|---|---|---|---|---|
| $\beta_1 = 0.8$ | $\beta_2 = 0.05$ | 0.54 | 0.54 | No dominance |
| $\beta_1 = 0.8$ | $\beta_2 = 0.01$ | 0.54 | 0.11 | CO dominates temperature |
| $\beta_1 = 0.1$ | $\beta_2 = 0.05$ | 0.07 | 0.54 | Temperature dominates CO |

**Table 3.3:** Choice of the coefficients for CO and relative humidity to generate data.

| CO (mean = 0.68) | RH (mean = 0.78) | $\beta_1 \times$ CO (mean) | $\beta_2 \times$ RH (mean) | comment |
|---|---|---|---|---|
| $\beta_1 = 0.8$ | $\beta_2 = 0.7$ | 0.54 | 0.55 | No dominance |
| $\beta_1 = 0.8$ | $\beta_2 = 0.05$ | 0.54 | 0.04 | CO dominates RH |
| $\beta_1 = 0.05$ | $\beta_2 = 0.7$ | 0.03 | 0.55 | RH dominates CO |

We considered n = 500 subjects in a study and we repeated the simulation 500 times. Let $l_1, l_2, \ldots, l_8$, and $u_1, u_2, \ldots, u_8$ be the lower and the upper limits, respectively, of the strata for the 8 strata model. In each replication, the following algorithm was used to generate data (the statistical software R was used for this simulation).

1. Generate 500 random numbers from each of the uniform distributions $U[a_k, b_k]$ in Table 3.2 for k = 1, 2, ..., 8. These numbers are the baseline intensities $\lambda_{ik}$. So for each individual, there are 8 baseline intensities corresponding to the 8 strata.
2. Generate the number of events for the ith individual on the jth day in the kth stratum from a Poisson distribution with mean $\lambda_{ik} \exp[x_{jk}^T \beta]$, where i = 1, 2, ..., 500, k = 1, 2, ..., 8 and $j = l_k, l_k + 1, \ldots, u_k$.
3. The days corresponding to the events for an individual determine the event times for that individual. The values of the pollutants and weather variables for those days are then used in the analysis stage to estimate the parameters.

## 3.5 Analysis

*Scenario I* generated data were analyzed by all three models and *Scenario II* generated data were analyzed only by *Model I* and *Model III*.

### 3.5.1 *Model I*

*Model I* assumes time-dependent baseline intensities. The choice of the partitions $I_l$ is subjective. However, the partition should be made in such a way that the baseline intensity for an individual is the same within a stratum and different in different strata. For a particular study, there may be some natural candidate stratifications like years, seasons, months and so on. In the absence of such logical choices, one has to use

judgment on when the baseline intensities may be constant. The estimates of the parameters $\beta$ are sensitive to the choices of the partitions as noted by Dewanji and Moolgavkar (2002).

For analysis, we selected seven different stratification models including the true one (Table 3.4). The stratifications of the time period were defined as follows: a single stratum over the entire period, 4 strata corresponding to 12 months, 8 strata corresponding to six months, 17 strata corresponding to different seasons (following Dewanji and Moolgavkar, 2000), 48 strata corresponding to one months, 59 strata corresponding to 25 days, and 96 strata corresponding to half of the months. The seasons were defined as follows: Winter was from December to February, Spring was from March to May, Summer was from June to August, and Fall was from September to November. So here the 1 and the 4 strata models are the coarser models and the 17, the 48, the 59, and the 96 strata models are the finer models.

For *Scenario I* simulations, we performed the cross-validation technique (Dewanji and Moolgavkar, 2002) to select an optimal model from different stratification models. Our objective was to verify whether this technique behaves well for the real pollutant values. The *nlm* function in R was used to estimate the parameters for *Model I*.

### 3.5.2 *Model II*

*Model II* was based on a bidirectional case-crossover design. Conditional logistic regression with each subject as one stratum was applied for inference purposes. The function *clogit* in R was used to estimate the parameter.

**Table 3.4:** Models under *Model I* and the corresponding stratification schemes used in simulation.

| Model | Lower bounds of the strata in days (observation period = 1461 days) |
|---|---|
| 1 stratum | 1 |
| 4 strata | 1, 367, 732, 1097 |
| 8 strata* | 1, 184, 367, 550, 732, 915, 1097, 1280 |
| 17 strata | 1, 62, 154, 245, 336, 428, 520, 611, 701, 793, 885, 976, 1066, 1158, 1250, 1341, 1431 |
| 48 strata | 1, 31, 62, 92, 123, 154, 184, 215, 245, 276, 307, 336, 367, 397, 428, 458, 489, 520, 550, 581, 611, 642, 673, 701, 732, 762, 793, 823, 854, 885, 915, 946, 976, 1007, 1038, 1066, 1097, 1127, 1158, 1188, 1219, 1250, 1280, 1311, 1341, 1372, 1403, 1431 |
| 59 strata | 1, 26, 51, 76, 101, 126, 151, 176, 201, 226, 251, 276, 301, 326, 351, 376, 401, 426, 451, 476, 501, 526, 551, 576, 601, 626, 651, 676, 701, 726, 751, 776, 801, 826, 851, 876, 901, 926, 951, 976, 1001, 1026, 1051, 1076, 1101, 1126, 1151, 1176, 1201, 1226, 1251, 1276, 1301, 1326, 1351, 1376, 1401, 1426, 1451 |
| 96 strata | 1, 16, 31, 46, 62, 77, 92, 107, 123, 138, 154, 169, 184, 199, 215, 230, 245, 260, 276, 291, 307, 321, 336, 351, 367, 382, 397, 412, 428, 443, 458, 473, 489, 504, 520, 535, 550, 565, 581, 596, 611, 626, 642, 657, 673, 687, 701, 716, 732, 747, 762, 777, 793, 808, 823, 838, 854, 869, 885, 900, 915, 930, 946, 961, 976, 991, 1007, 1022, 1038, 1052, 1066, 1081, 1097, 1112, 1127, 1142, 1158, 1173, 1188, 1203, 1219, 1234, 1250, 1265, 1280, 1295, 1311, 1326, 1341, 1356, 1372, 1387, 1403, 1417, 1431, 1446 |

* True model

### 3.5.3 *Model III*

*Model III* was a time series model. We used generalized linear model with natural spline to smooth the time trend. Quasi-likelihood method was used for estimation. Table 3.5 gives the knots and the corresponding degrees of freedom that we used in the simulation. The simulations were done in R using the *glm* function.

**Table 3.5** Knots per year and the corresponding degrees of freedom used in simulation.

| Knots per year | 1 | 2 | 3 | 4 | 6 | 12 | 24 |
|---|---|---|---|---|---|---|---|
| Degrees of freedom | 6 | 10 | 14 | 18 | 26 | 50 | 98 |

The following algorithm was used to estimate the parameters in each replication.

1. For each chosen degrees of freedom, run the generalized linear model with ns(time, df) as covariate, and find the p-values of the Bartlett's test;
2. Select the degrees of freedom corresponding to the largest p-value of the Bartlett's test. Label this degrees of freedom as $df^*$;
3. Run the generalized linear model with the pollutant, weather variable and ns(time, $df^*$) to estimate the regression coefficients.

### 3.5.4 Groupings of the Models

We defined four groups on the basis of the similarity of the models. *Group I* consisted of the true model (8 strata), and the finer models (48 and 96 strata) for which each assumed stratum was completely contained in exactly one true stratum. *Group II* consisted of the finer models (17 and 59 strata) for which at least one assumed stratum was not completely contained in exactly one true stratum. *Group III* consisted of the coarser models (the 1 and the 4 strata models) and *Model II*. The 1 stratum model and *Model II* assumed that the baseline intensities are constant with respect to time. Finally, *Model III* was considered as *Group IV* model.

## 3.6 Evaluation Criteria

An estimator is precise if the variance is small, and accurate if the MSE is small. Accuracy (MSE) is how close the estimate is to the true value, whereas precision (variance) measures how close estimates from different replicates are to each other (Lohr, 1998). So the MSE was selected as the main evaluation criteria. However, we also present the biases and the variances of the estimates.

Five hundred replications were executed in this simulation. The means and the variances of the parameter estimates were calculated. The bias was defined as $E(\hat{\theta}) - \theta$ and the MSE as $E(\hat{\theta} - \theta)^2$, where $\hat{\theta}$ is the estimate and $\theta$ is the parameter. The expression for MSE can be simplified as $MSE = E(\hat{\theta} - \theta)^2 = \text{variance} + (\text{bias})^2$.

Under *Model I*, the optimal model was selected in each replication by the cross-validation technique, and the estimate from that optimal model was calculated. The mean of these estimates was referred to as the optimal estimate.

## 3.7 *Scenario I* Simulation Results

In *Scenario I* simulation, the same set of variables was used for generation as well as for analysis of data. We considered three situations that are outlined in Table 3.6.

**Table 3.6:** Variables used in *Scenario I* simulation.

| Data generation | Data analysis |
|---|---|
| CO | CO |
| CO and temperature | CO and temperature |
| CO and relative humidity | CO and relative humidity |

***Case I: Baseline intensities follow different distributions in different strata (CO vs. CO)***

Results are given in rows 1-3 of Table 3.7, when the baseline intensities follow different distributions in different strata. The MSEs were not strictly increasing as the number of strata increased. However, we noticed the increasing pattern under each group of models. That is, for each of *Group I* and *Group II*, MSEs increased as the number of strata increased, and for *Group III*, MSEs increased in the direction *Model II* $\rightarrow$ 1 stratum model $\rightarrow$ 4 strata model. Under any particular group, the means of the estimates were very close (i.e. the biases are similar) but the variances were increasing, thus making the

MSEs increasing. For *Group I* and *Group II* models, MSEs increased at a decreasing rate as the within stratum variation in the baseline intensities increased. For *Group III* models, MSEs increased at an increasing rate as the within stratum variation in the baseline intensities increased. For a particular model within *Groups I, II* and *IV*, the MSEs were decreasing as the within stratum variation in the baseline intensities increased, but for *Group III* models, the reverse holds. These relationships are summarized in Table 3.8.

As expected, the simulation results showed that the 8 strata model always gave a smaller MSE than any of the other models. So a model selection technique should select the true model more often, and the choice between the other models under *Group I* and *Group II* should be similar as the MSEs for those models were close to each other.

The cross-validation technique (Table 3.7) selected the true model most often (37.2%, 53.8% and 36.0%) and the 17 strata model was the second most selected model (22.6%, 20.8% and 21.2%). The selection of the other models (48 strata, 59 strata and 96 strata models) were quite evenly distributed. So the cross-validation technique can be considered as a good method of selecting the optimal model as this method selected the right model most often and the selection percentages for the models under *Group III* were fairly small and close to zero as the within stratum variation in the $\lambda_i$'s increased.

The optimal models obtained by cross-validation (the third last column of Table 3.7) produced means of estimates which were close to the true value 0.05. The variances of the optimal models differed in the three cases and they decreased as the within stratum variation in the $\lambda_i$'s increased. As a consequence, the MSEs were also decreasing as the within stratum variation in the $\lambda_i$'s increased. So, the optimal models became more reliable as the within stratum variations in the $\lambda_i$'s increased.

**Table 3.7:** simulation results. Data generation: CO, data analysis: CO, n = 500, study period = 1461 days, $\beta = 0.05$.

| Baseline Intensities | Statistics | *Model I* | | | | | | | | *Model II* | *Model III* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 Stratum | 4 Strata | 8 Strata | 17 Strata | 48 Strata | 59 strata | 96 Strata | Optimal model** | | |
| ( 0.01, 0.015), ( 0.015, 0.02), | Mean | -0.02070 | 0.17567 | 0.04912 | 0.04312 | 0.04817 | 0.04365 | 0.04817 | 0.04760 | -0.02025 | 0.04024 |
| ( 0.02, 0.025), ( 0.025, 0.03), | Bias | -0.07070 | 0.12567 | -0.00088 | -0.00688 | -0.00183 | -0.00635 | -0.00183 | -0.00240 | -0.07025 | -0.00976 |
| ( 0.03, 0.035), ( 0.035, 0.04), | Variance | 0.00069 | 0.00072 | 0.00084 | 0.00095 | 0.00110 | 0.00122 | 0.00133 | 0.00203 | 0.00068 | 0.00074 |
| ( 0.04, 0.045), ( 0.045, 0.05) | MSE | 0.00569 | 0.01651 | 0.00084 | 0.00100 | 0.00110 | 0.00126 | 0.00133 | 0.00204 | 0.00561 | 0.00084 |
| | Percentage* | 7.8 | 3.8 | 37.2 | 22.6 | 9.4 | 10.6 | 8.6 | - | - | - |
| ( 0.01, 0.03), ( 0.03, 0.05), | Mean | -0.04380 | 0.22346 | 0.04889 | 0.04161 | 0.04878 | 0.04235 | 0.04916 | 0.04814 | -0.04202 | 0.03835 |
| ( 0.05, 0.07), ( 0.07, 0.09), | Bias | -0.09380 | 0.17346 | -0.00111 | -0.00839 | -0.00122 | -0.00765 | -0.00084 | -0.00186 | -0.09202 | -0.01165 |
| ( 0.09, 0.11), ( 0.11, 0.13), | Variance | 0.00024 | 0.00025 | 0.00031 | 0.00033 | 0.00038 | 0.00040 | 0.00044 | 0.00039 | 0.00023 | 0.00040 |
| ( 0.13, 0.15), ( 0.15, 0.17) | MSE | 0.00904 | 0.03034 | 0.00031 | 0.00040 | 0.00038 | 0.00046 | 0.00044 | 0.00039 | 0.00869 | 0.00054 |
| | Percentage* | 0.0 | 0.0 | 53.8 | 20.8 | 9.4 | 6.6 | 9.4 | - | - | - |
| ( 0.00, 0.04), ( 0.04, 0.08), | Mean | -0.05504 | 0.24881 | 0.04965 | 0.04153 | 0.04946 | 0.04236 | 0.05000 | 0.04758 | -0.05132 | 0.04530 |
| ( 0.08, 0.12), ( 0.12, 0.16), | Bias | -0.10504 | 0.19881 | -0.00035 | -0.00847 | -0.00054 | -0.00764 | 0.00000 | -0.00242 | -0.10132 | -0.00470 |
| ( 0.16, 0.20), ( 0.20, 0.24), | Variance | 0.00014 | 0.00014 | 0.00018 | 0.00019 | 0.00021 | 0.00023 | 0.00026 | 0.00027 | 0.00012 | 0.00028 |
| ( 0.24, 0.28), ( 0.28, 0.32) | MSE | 0.01117 | 0.03967 | 0.00018 | 0.00027 | 0.00021 | 0.00028 | 0.00026 | 0.00027 | 0.01039 | 0.00030 |
| | Percentage* | 0.2 | 0.0 | 36.0 | 21.2 | 12.6 | 12.0 | 18.0 | - | - | - |
| ( 0.01, 0.03), ( 0.01, 0.03), | Mean | 0.04923 | 0.04915 | 0.04973 | 0.04966 | 0.04913 | 0.04938 | 0.04974 | 0.04967 | 0.04858 | 0.04970 |
| ( 0.01, 0.03), ( 0.01, 0.03), | Bias | -0.00077 | -0.00085 | -0.00027 | -0.00034 | -0.00087 | -0.00062 | -0.00026 | -0.00033 | -0.00142 | -0.00030 |
| ( 0.01, 0.03), ( 0.01, 0.03), | Variance | 0.00095 | 0.00107 | 0.00117 | 0.00131 | 0.00148 | 0.00165 | 0.00170 | 0.00137 | 0.00105 | 0.00127 |
| ( 0.01, 0.03), ( 0.01, 0.03) | MSE | 0.00096 | 0.00108 | 0.00117 | 0.00131 | 0.00148 | 0.00165 | 0.00170 | 0.00137 | 0.00105 | 0.00127 |
| | Percentage* | 34.8 | 20.2 | 16.6 | 11.0 | 4.4 | 7.6 | 5.4 | - | - | - |
| ( 0.00, 0.16), ( 0.00, 0.16), | Mean | 0.05011 | 0.05000 | 0.04949 | 0.04985 | 0.04906 | 0.04893 | 0.04951 | 0.05005 | 0.04728 | 0.05002 |
| ( 0.00, 0.16), ( 0.00, 0.16), | Bias | 0.00011 | 0.00000 | -0.00051 | -0.00015 | -0.00094 | -0.00107 | -0.00049 | 0.00005 | -0.00272 | 0.00002 |
| ( 0.00, 0.16), ( 0.00, 0.16), | Variance | 0.00042 | 0.00041 | 0.00030 | 0.00032 | 0.00038 | 0.00040 | 0.00044 | 0.00039 | 0.00040 | 0.00038 |
| ( 0.00, 0.16), ( 0.00, 0.16) | MSE | 0.00042 | 0.00041 | 0.00030 | 0.00032 | 0.00038 | 0.00040 | 0.00044 | 0.00039 | 0.00041 | 0.00038 |
| | Percentage* | 23.4 | 19.2 | 27.4 | 10.0 | 5.4 | 8.4 | 6.2 | - | - | - |

* Percentage of the times a model was selected by the cross-validation technique. ** The optimal model selected by the cross-validation technique

**Table 3.8:** Patterns of MSEs for models with $\lambda_i$'s follow different distributions in different strata (data generation: CO, data analysis: CO).

| Models | Pattern |
|---|---|
| **Models under** | |
| *Group I* and *Group II* | 1. MSEs increased as the number of strata increased.<br>2. Rate of increase was decreasing as the within stratum variation in $\lambda_i$'s increased. |
| *Group III* | 1. MSEs increased in the direction *Model II* $\rightarrow$ 1 stratum model $\rightarrow$ 4 strata model.<br>2. Rate of increase was increasing as the within stratum variation in $\lambda_i$'s increased. |
| **Any particular model** | |
| *Group I*, *Group II* and *Group IV* | MSEs decreased as the within stratum variation in $\lambda_i$'s increased. |
| *Group III* | MSEs increased as the within stratum variation in $\lambda_i$'s increased. |

Finally, we noticed from Table 3.7 that MSEs of the true model (8 strata model) and *Model III* were close. For baseline intensities of the form [0.01, 0.015], ...., [0.045, 0.05], they produced almost the same MSE. However, as the within stratum variation in baseline intensities increased, the true model gave smaller a MSE than *Model III*.

***Case I I: Baseline intensities follow same distribution within each stratum (CO vs. CO)***

Results are given in the last two rows of Table 3.7 and in Table 3.9 when the baseline intensities followed the same distribution within each stratum. The MSEs increased in the direction of the true model (8 strata model) to the finest model (96 strata model) when the within stratum variation in the $\lambda_i$'s was large, and the MSEs strictly increased when this variation in the $\lambda_i$'s was small.

The means of the estimates were very close to the true value 0.05 for all models. Looking at the variances for models under *Model I*, we see that although these were increasing when the baseline intensities were of the form [0.01, 0.03], this was not true when the baseline intensities were of the form [0.00, 0.16]. That is, unlike models for different baseline intensities in different strata, the variances here were not strictly increasing. To be more confident about it, we performed some more simulations (Table 3.9) considering baseline intensities of the forms [0.00, 0.02], [0.00, 0.04], [0.00, 0.06], [0.00, 0.08]. We see that the variances were not strictly increasing except for the case where the baseline intensities were of the form [0.00, 0.02]. However, the variances were increasing from the direction of the 8 strata model to the 96 strata model. Thus, we can conclude that the variances were increasing only when the within stratum variation in the $\lambda_i$'s was small, otherwise the variances were not strictly increasing, though they were increasing from the direction of the true model to the finest model. Further, we noticed from Table 3.9 that for any particular model, the MSEs decreased as the within stratum variation in the $\lambda_i$'s increased. These relationships are summarized in Table 3.10.

When the variation in the $\lambda_i$'s was small, the MSE suggested the 1 stratum model was optimal. When the variation was large, the true model produced the smallest MSE. The MSE also suggested that the coarser models and the true model should be selected more than the other models if the variation in the $\lambda_i$'s was not too large (baseline intensities of the form [0.00, 0.02], [0.00, 0.04], [0.00, 0.06] and [0.00, 0.08]). When the variation in the $\lambda_i$'s was large (baseline intensities of the form [0.00, 0.16]), it was clear that the true model and the finer models should be selected more frequently.

**Table 3.9:** Simulation results for *Model I* considering the baseline intensities to follow the same distribution in each stratum. Data generation: CO, data analysis: CO, n = 500, study period = 1461 days, $\beta = 0.05$.

| Baseline Intensities | Statistics | *Model I* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 Stratum | 4 Strata | 8 Strata | 17 Strata | 48 Strata | 59 strata | 96 Strata | Optimal model** |
| ( 0.00, 0.02), ( 0.00, 0.02), | Mean | 0.04969 | 0.04958 | 0.04866 | 0.04895 | 0.04721 | 0.04706 | 0.04790 | 0.05023 |
| ( 0.00, 0.02), ( 0.00, 0.02), | Bias | -0.00031 | -0.00042 | -0.00134 | -0.00105 | -0.00279 | -0.00294 | -0.00210 | 0.00023 |
| ( 0.00, 0.02), ( 0.00, 0.02), | Variance | 0.00238 | 0.00246 | 0.00260 | 0.00295 | 0.00346 | 0.00368 | 0.00411 | 0.00390 |
| ( 0.00, 0.02), ( 0.00, 0.02) | MSE | 0.00238 | 0.00246 | 0.00260 | 0.00295 | 0.00347 | 0.00369 | 0.00412 | 0.00390 |
| | Percentage* | 21.8 | 21.0 | 18.6 | 11.4 | 6.6 | 10.4 | 10.2 | |
| ( 0.00, 0.04), ( 0.00, 0.04), | Mean | 0.05009 | 0.04995 | 0.05023 | 0.05043 | 0.05001 | 0.04965 | 0.05060 | 0.049991 |
| ( 0.00, 0.04), ( 0.00, 0.04), | Bias | 0.00009 | -0.00005 | 0.00023 | 0.00043 | 0.00001 | -0.00035 | 0.00060 | -0.00009 |
| ( 0.00, 0.04), ( 0.00, 0.04), | Variance | 0.00117 | 0.00121 | 0.00117 | 0.00130 | 0.00146 | 0.00166 | 0.00178 | 0.00144 |
| ( 0.00, 0.04), ( 0.00, 0.04) | MSE | 0.00117 | 0.00121 | 0.00117 | 0.00130 | 0.00146 | 0.00166 | 0.00178 | 0.00144 |
| | Percentage* | 23.4 | 17.2 | 21.2 | 12.2 | 6.2 | 9.0 | 10.8 | |
| ( 0.00, 0.06), ( 0.00, 0.06), | Mean | 0.05023 | 0.05018 | 0.05032 | 0.05038 | 0.05072 | 0.05006 | 0.05131 | 0.05055 |
| ( 0.00, 0.06), ( 0.00, 0.06), | Bias | 0.00023 | 0.00018 | 0.00032 | 0.00038 | 0.00072 | 0.00006 | 0.00131 | 0.00055 |
| ( 0.00, 0.06), ( 0.00, 0.06), | Variance | 0.00085 | 0.00085 | 0.00077 | 0.00090 | 0.00100 | 0.00111 | 0.00117 | 0.00098 |
| ( 0.00, 0.06), ( 0.00, 0.06) | MSE | 0.00085 | 0.00085 | 0.00077 | 0.00090 | 0.00100 | 0.00111 | 0.00117 | 0.00098 |
| | Percentage* | 24.6 | 23.6 | 16.8 | 11.0 | 8.2 | 7.6 | 8.2 | |
| ( 0.00, 0.08), ( 0.00, 0.08), | Mean | 0.05102 | 0.05106 | 0.05120 | 0.05151 | 0.05150 | 0.05120 | 0.05236 | 0.0519 |
| ( 0.00, 0.08), ( 0.00, 0.08), | Bias | 0.00102 | 0.00106 | 0.00120 | 0.00151 | 0.00150 | 0.00120 | 0.00236 | 0.00195 |
| ( 0.00, 0.08), ( 0.00, 0.08), | Variance | 0.00067 | 0.00066 | 0.00060 | 0.00070 | 0.00077 | 0.00085 | 0.00093 | 0.00077 |
| ( 0.00, 0.08), ( 0.00, 0.08) | MSE | 0.00067 | 0.00066 | 0.00060 | 0.00070 | 0.00078 | 0.00085 | 0.00094 | 0.00078 |
| | Percentage* | 27.6 | 18.0 | 23.2 | 10.0 | 5.2 | 10.0 | 6.0 | |
| ( 0.00, 0.16), ( 0.00, 0.16), | Mean | 0.05011 | 0.05000 | 0.04949 | 0.04985 | 0.04906 | 0.04893 | 0.04951 | 0.04728 |
| ( 0.00, 0.16), ( 0.00, 0.16), | Bias | 0.00011 | 0.00000 | -0.00051 | -0.00015 | -0.00094 | -0.00107 | -0.00049 | -0.00272 |
| ( 0.00, 0.16), ( 0.00, 0.16), | Variance | 0.00042 | 0.00041 | 0.00030 | 0.00032 | 0.00038 | 0.00040 | 0.00044 | 0.00040 |
| ( 0.00, 0.16), ( 0.00, 0.16) | MSE | 0.00042 | 0.00041 | 0.00030 | 0.00032 | 0.00038 | 0.00040 | 0.00044 | 0.00041 |
| | Percentage* | 23.4 | 19.2 | 27.4 | 10.0 | 5.4 | 8.4 | 6.2 | |

* Percentage of the times a model was selected by the cross-validation technique. ** The optimal model selected by the cross-validation technique.

**Table 3.10:** Patterns of MSEs for models with baseline intensities follow the same distribution within each stratum (data generation: CO, data analysis: CO).

| Models | Pattern |
|---|---|
| **Models under *Model I*** | |
| Small within stratum variation in the $\lambda_i$'s | MSEs increased as number of strata increased. |
| Large within stratum variation in the $\lambda_i$'s | MSEs increased in the direction of the true model to the finest model. |
| **Any particular model** | MSEs decreased as the within stratum variation in $\lambda_i$'s increased. |

Looking at the cross-validation results (Table 3.7 and Table 3.9), we see that the coarser models (the 1 stratum model and the 4 strata model) and the true model were chosen most of the times. Except for the case [0.00, 0.16], the 1 stratum model was chosen more often than the other models. On the other hand, the MSE suggested the true model to be the optimal one for the cases [0.00, 0.06], [0.00, 0.08] and [0.00, 0.16]. Further, the true model and the finer models should be selected more often when the variation in the $\lambda_i$'s was quite large (the case [0.00, 0.16]). Therefore, we conclude that the cross-validation technique may not perform well in terms of MSE when the baseline intensities followed the same distribution within each stratum. However, the optimal models from the cross-validation technique gave estimates close to the true value 0.05, and the MSE's of the optimal models decreased as the within stratum variation in the $\lambda_i$'s increased.

Finally, from Table 3.7, we see that the estimates under *Model II and Model III* were also close to the true parameter value. MSEs were also small but not as small as the MSE of the optimal model under *Model I.*

Results for *Scenario I* simulations with CO and the weather variables are given in Table 3.11 and Table 3.12. We considered *Model I* and *Model III* and investigated two situations: CO and temperature vs. CO and temperature, and CO and relative humidity (RH) vs. CO and RH. We only investigated the case where the baseline intensities followed the same distribution in each stratum. The 8 strata model was the true model and the baseline intensities were [0.01, 0.03], ..., [0.01, 0.03] for the 8 strata.

***CO and temperature vs. CO and temperature***

When CO and temperature were used to generate data, we considered three situations as described in Section 3.4, namely, no dominance, temperature dominates CO and CO dominates temperature. Results for the coefficients of CO are given in rows 4-6 of Table 3.11. Since our main interest was on the pollutant CO, we did not present the results for the coefficients of temperature.

We see that the biases of $\hat{\beta}_1$ were fairly small in each case. In all the cases, the variances and the MSEs were increasing as the number of strata increased. However, if the within stratum variation in the baseline intensities were large, we might not get this trend for the variances and MSEs. We noticed before that the variances and MSEs were not strictly increasing when the within stratum variation in the baseline intensities was large rather they were increasing in the direction from the true model to the finest model. Hence, the same conclusion can be made as the CO vs. CO case with small variation in the baseline intensities, namely, the coarser models and the true model are the best in these situations. The results for *Model III* are given in the last column of rows 4-6 in Table 3.11. The estimates were quite close to the true parameter values. The MSEs were also small, but not as small as the MSE of the optimal model under *Model I*.

*CO and RH vs. CO and RH*

Here again we considered three cases: no dominance, RH dominates CO and CO dominates RH. The results for the coefficients of CO are given in rows 4-6 of Table 3.12. We noticed the same type of results we found for the case CO and temperature, that is, the biases were small and the variances and MSEs were increasing. So the coarser models and the true model are the best for *Scenario I* simulation when the variations in the baseline intensities are small.

## 3.8 *Scenario II* Simulation Results

In *scenario II* simulation, different sets of variables were used in data generation and in data analysis. We only investigated the case where the baseline intensities followed the same distribution in each stratum. The 8 strata model was the true model and the baseline intensities were [0.01, 0.03], ..., [0.01, 0.03] for the 8 strata.

First we investigated the situation when we generated data using CO and temperature/RH, but analyzed them without the covariate. This case is common in real life, since we do not have information on all the determinants of the response variable and often omit some covariates that are actually responsible for the response variable. These results are given in Table 3.11 and Table 3.12 (rows 1-3).

*CO and temperature/RH vs. CO: No dominance*

When CO and temperature were used to generate the data (Table 3.11, rows 1), we found no obvious pattern in the MSEs. This was due to the fact that though the variances were increasing in the direction from the true model to the finest model, the biases did not follow any strict pattern and we noticed large variations among them. The coarser

models produced large biases compared to the other models. The finer models (59 and 96 strata models) seemed to be better in terms of MSEs. The MSE of *Model III* was not so far away from the MSE of the optimal model under *Model I*. Further, the means of estimates underestimated the true value all the time. Since CO and temperature were negatively correlated, the number of admissions observed was less than what would have been if we generated the data using CO only. This was due to the negative effects of temperature on CO. So when we analyzed the data using CO only, we obtained estimates which underestimated the true regression coefficient.

When CO and RH were used to generate the data (Table 3.12, row 1), we found that MSEs decreased in the direction from the true model to the finest model. In this case, clearly the finer models were better. The variances were increasing and the biases were decreasing in the same direction too. The coarser models gave large bias but not as large as that for the case of negative correlation. The MSE of *Model III* was also quite close to the optimal model. We also noticed that the means of the estimates always overestimated the true value. Since CO and RH were positively correlated, the number of admissions observed was more than what we would anticipate if the data were generated using CO only. This was due to the positive effects of RH on CO. So when we analyzed the data using CO only, we obtained estimates which overestimated the true regression coefficient.

***CO and temperature/RH vs. CO: Temperature/RH dominates CO***

Tables 3.11 and 3.12 (row 2) show the results for the case of CO and temperature vs. CO and CO and RH vs. CO, respectively. We noticed the same conclusions for these two cases with the corresponding two cases for no dominance.

**Table 3.11:** Simulation results using CO and temperature. $\lambda$ : (0.01, 0.03), ..., (0.01, 0.03), n = 500, study period = 1461 days.

| Data Generation | Data Analysis | Statistics | *Model I* | | | | | | | *Model III* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 Stratum | 4 Strata | 8 Strata | 17 Strata | 48 Strata | 59 strata | 96 Strata | |
| **No Dominance** | **Model:** Different | Mean | 0.46036 | 0.46056 | 0.75064 | 0.72700 | 0.74394 | 0.78163 | 0.76424 | 0.75010 |
| **Model:** 8 Strata with CO and temp | stratification models | Bias | -0.33964 | -0.33944 | -0.04936 | -0.07300 | -0.05606 | -0.01837 | -0.03576 | -0.04990 |
| $\beta_1 = 0.8$; $\beta_1 \times$ mean(CO) = 0.54 | and *Model III* | Variance | 0.00031 | 0.00031 | 0.00030 | 0.00035 | 0.00042 | 0.00045 | 0.00050 | 0.00053 |
| $\beta_2 = 0.05$; $\beta_2 \times$ mean(temp) = 0.54 | with CO only | MSE | 0.11567 | 0.11553 | 0.00274 | 0.00568 | 0.00356 | 0.00078 | 0.00178 | 0.00302 |
| **Temp dominates CO** | **Model:** Different | Mean | -0.26971 | -0.26843 | 0.06149 | 0.03924 | 0.05950 | 0.09802 | 0.08094 | 0.06972 |
| **Model:** 8 Strata with CO and temp | stratification models | Bias | -0.36971 | -0.36843 | -0.03851 | -0.06076 | -0.04050 | -0.00198 | -0.01906 | -0.03028 |
| $\beta_1 = 0.1$; $\beta_1 \times$ CO(mean) = 0.07 | and *Model III* | Variance | 0.00071 | 0.00071 | 0.00070 | 0.00079 | 0.00090 | 0.00097 | 0.00102 | 0.00107 |
| $\beta_2 = 0.05$; $\beta_2 \times$ mean(temp) = 0.54 | with CO only | MSE | 0.13740 | 0.13646 | 0.00219 | 0.00448 | 0.00254 | 0.00098 | 0.00139 | 0.00199 |
| **CO dominates temp** | **Model:** Different | Mean | 0.73177 | 0.73161 | 0.78578 | 0.78130 | 0.78415 | 0.79175 | 0.78769 | 0.78448 |
| **Model:** 8 Strata with CO and temp | stratification models | Bias | -0.06823 | -0.06839 | -0.01422 | -0.01870 | -0.01585 | -0.00825 | -0.01231 | -0.01552 |
| $\beta_1 = 0.8$; $\beta_1 \times$ mean(CO) = 0.54 | and *Model III* | Variance | 0.00042 | 0.00042 | 0.00046 | 0.00053 | 0.00063 | 0.00068 | 0.00077 | 0.00080 |
| $\beta_2 = 0.01$; $\beta_2 \times$ mean(temp) = 0.11 | with CO only | MSE | 0.00507 | 0.00510 | 0.00066 | 0.00088 | 0.00089 | 0.00075 | 0.00092 | 0.00104 |
| **No dominance** | **Model:** Different | Mean | 0.80010 | 0.80006 | 0.80017 | 0.80025 | 0.80017 | 0.80002 | 0.80057 | 0.79954 |
| **Model:** 8 Strata with CO and temp | stratification models | Bias | 0.00010 | 0.00006 | 0.00017 | 0.00025 | 0.00017 | 0.00002 | 0.00057 | -0.00046 |
| $\beta_1 = 0.8$; $\beta_1 \times$ mean(CO) = 0.54 | and *Model III* | Variance | 0.00030 | 0.00030 | 0.00030 | 0.00036 | 0.00042 | 0.00045 | 0.00050 | 0.00127 |
| $\beta_2 = 0.05$; $\beta_2 \times$ mean(temp) = 0.54 | with CO and temp | MSE | 0.00030 | 0.00030 | 0.00030 | 0.00036 | 0.00042 | 0.00045 | 0.00050 | 0.00127 |
| **Temp dominates CO** | **Model:** Different | Mean | 0.10028 | 0.10020 | 0.10053 | 0.10054 | 0.10029 | 0.10015 | 0.10043 | 0.09976 |
| **Model:** 8 Strata with CO and temp | stratification Models | Bias | 0.00028 | 0.00020 | 0.00053 | 0.00054 | 0.00029 | 0.00015 | 0.00043 | -0.00024 |
| $\beta_1 = 0.1$; $\beta_1 \times$ mean(CO) = 0.07 | and *Model III* | Variance | 0.00068 | 0.00069 | 0.00070 | 0.00079 | 0.00091 | 0.00097 | 0.00105 | 0.00110 |
| $\beta_2 = 0.05$; $\beta_2 \times$ mean(temp) = 0.54 | with CO and temp | MSE | 0.00068 | 0.00069 | 0.00070 | 0.00079 | 0.00091 | 0.00097 | 0.00105 | 0.00110 |
| **CO dominates temp** | **Model:** Different | Mean | 0.80016 | 0.80006 | 0.80014 | 0.80034 | 0.80008 | 0.79987 | 0.80022 | 0.79972 |
| **Model:** 8 Strata with CO and temp | stratification models | Bias | 0.00016 | 0.00006 | 0.00014 | 0.00034 | 0.00008 | -0.00013 | 0.00022 | -0.00028 |
| $\beta_1 = 0.8$; $\beta_1 \times$ mean(CO) = 0.54 | and *Model III* | Variance | 0.00044 | 0.00045 | 0.00046 | 0.00054 | 0.00064 | 0.00069 | 0.00078 | 0.00082 |
| $\beta_2 = 0.01$; $\beta_2 \times$ mean(temp) = 0.11 | with CO and temp | MSE | 0.00044 | 0.00045 | 0.00046 | 0.00054 | 0.00064 | 0.00069 | 0.00078 | 0.00082 |
| **Model:** 8 Strata with CO | **Model:** Different | Mean | 0.04946 | 0.04951 | 0.04969 | 0.04979 | 0.04927 | 0.04943 | 0.04980 | 0.04987 |
| $\beta = 0.05$ | stratification models | Bias | -0.00054 | -0.00049 | -0.00031 | -0.00021 | -0.00073 | -0.00057 | -0.00020 | -0.00013 |
| | and *Model III* | Variance | 0.00117 | 0.00119 | 0.00120 | 0.00134 | 0.00152 | 0.00167 | 0.00174 | 0.00136 |
| | with CO and temp | MSE | 0.00117 | 0.00119 | 0.00120 | 0.00134 | 0.00152 | 0.00167 | 0.00174 | 0.00136 |

Note: Analysis results are given only for the coefficient of CO.

**Table 3.12:** Simulation results using CO and relative humidity (RH). $\lambda$ : (0.01, 0.03), ..., (0.01, 0.03), n = 500, study period = 1461 days.

| Data Generation | Data Analysis | Statistics | *Model I* | | | | | | | *Model III* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 Stratum | 4 Strata | 8 Strata | 17 Strata | 48 Strata | 59 strata | 96 Strata | |
| **No dominance** | **Model:** Different | Mean | 0.86265 | 0.86916 | 0.83915 | 0.82676 | 0.82116 | 0.81688 | 0.81234 | 0.81468 |
| **Model:** 8 Strata with CO and RH | stratification models | Bias | 0.06265 | 0.06916 | 0.03915 | 0.02676 | 0.02116 | 0.01688 | 0.01234 | 0.01468 |
| $\beta_1 = 0.8$; $\beta_1 \times$ co(mean) = 0.54 | and *Model III* | Variance | 0.00024 | 0.00024 | 0.00025 | 0.00029 | 0.00035 | 0.00037 | 0.00043 | 0.00044 |
| $\beta_2 = 0.7$; $\beta_2 \times$ RH(mean) = 0.55 | with CO only | MSE | 0.00416 | 0.00502 | 0.00178 | 0.00100 | 0.00080 | 0.00066 | 0.00059 | 0.00065 |
| **RH dominates CO** | **Model:** Different | Mean | 0.12304 | 0.13058 | 0.09581 | 0.08147 | 0.07706 | 0.07144 | 0.06717 | 0.07655 |
| **Model:** 8 Strata with CO and RH | stratification models | Bias | 0.07304 | 0.08058 | 0.04581 | 0.03147 | 0.02706 | 0.02144 | 0.01717 | 0.02655 |
| $\beta_1 = 0.05$; $\beta_1 \times$ co(mean) = 0.03 | and *Model III* | Variance | 0.00056 | 0.00057 | 0.00061 | 0.00072 | 0.00082 | 0.00090 | 0.00098 | 0.00084 |
| $\beta_2 = 0.7$; $\beta_2 \times$ RH(mean) = 0.55 | with CO only | MSE | 0.00589 | 0.00706 | 0.00270 | 0.00171 | 0.00155 | 0.00136 | 0.00128 | 0.00154 |
| **CO dominates RH** | **Model:** Different | Mean | 0.80448 | 0.80483 | 0.80296 | 0.80207 | 0.80173 | 0.80121 | 0.80136 | 0.80109 |
| **Model:** 8 Strata with CO and RH | stratification models | Bias | 0.00448 | 0.00483 | 0.00296 | 0.00207 | 0.00173 | 0.00121 | 0.00136 | 0.00109 |
| $\beta_1 = 0.8$; $\beta_1 \times$ CO(mean) = 0.54 | and *Model III* | Variance | 0.00043 | 0.00044 | 0.00049 | 0.00057 | 0.00069 | 0.00073 | 0.00083 | 0.00085 |
| $\beta_2 = 0.05$; $\beta_2 \times$ RH(mean) = 0.04 | with CO only | MSE | 0.00045 | 0.00046 | 0.00050 | 0.00058 | 0.00069 | 0.00073 | 0.00083 | 0.00086 |
| **No dominance** | **Model:** Different | Mean | 0.80015 | 0.80028 | 0.80032 | 0.80037 | 0.80021 | 0.79983 | 0.80064 | 0.79999 |
| **Model:** 8 Strata with CO and RH | stratification models | Bias | 0.00015 | 0.00028 | 0.00032 | 0.00037 | 0.00021 | -0.00017 | 0.00064 | -0.00001 |
| $\beta_1 = 0.8$; $\beta_1 \times$ CO(mean) = 0.54 | and *Model III* | Variance | 0.00027 | 0.00027 | 0.00027 | 0.00030 | 0.00037 | 0.00038 | 0.00044 | 0.00045 |
| $\beta_2 = 0.7$; $\beta_2 \times$ rh(mean) = 0.55 | with CO and RH | MSE | 0.00027 | 0.00027 | 0.00027 | 0.00030 | 0.00037 | 0.00038 | 0.00044 | 0.00045 |
| **RH dominates CO** | **Model:** Different | Mean | 0.05014 | 0.05008 | 0.05049 | 0.05031 | 0.05033 | 0.05023 | 0.05075 | 0.05080 |
| **Model:** 8 Strata with CO and RH | stratification models | Bias | 0.00014 | 0.00008 | 0.00049 | 0.00031 | 0.00033 | 0.00023 | 0.00075 | 0.00080 |
| $\beta_1 = 0.05$; $\beta_1 \times$ CO(mean) = 0.03 | and *Model III* | Variance | 0.00062 | 0.00063 | 0.00065 | 0.00074 | 0.00085 | 0.00093 | 0.00100 | 0.00082 |
| $\beta_2 = 0.7$; $\beta_2 \times$ RH(mean) = 0.55 | with CO and RH | MSE | 0.00062 | 0.00063 | 0.00065 | 0.00074 | 0.00085 | 0.00093 | 0.00100 | 0.00082 |
| **CO dominates RH** | **Model:** Different | Mean | 0.80018 | 0.79998 | 0.80016 | 0.80018 | 0.80014 | 0.80000 | 0.80043 | 0.79993 |
| **Model:** 8 Strata with CO and RH | stratification models | Bias | 0.00018 | -0.00002 | 0.00016 | 0.00018 | 0.00014 | 0.00000 | 0.00043 | -0.00007 |
| $\beta_1 = 0.8$; $\beta_1 \times$ CO(mean) = 0.54 | and *Model III* | Variance | 0.00048 | 0.00049 | 0.00052 | 0.00059 | 0.00070 | 0.00074 | 0.00084 | 0.00087 |
| $\beta_2 = 0.05$; $\beta_2 \times$ RH(mean) = 0.04 | with CO and RH | MSE | 0.00048 | 0.00049 | 0.00052 | 0.00059 | 0.00070 | 0.00074 | 0.00084 | 0.00087 |
| **Model:** 8 Strata with CO | **Model:** Different | Mean | 0.04948 | 0.04929 | 0.04960 | 0.04960 | 0.04880 | 0.04943 | 0.04980 | 0.04991 |
| $\beta = 0.05$ | stratification models | Bias | -0.00052 | -0.00071 | -0.00040 | -0.00040 | -0.00120 | -0.00057 | -0.00020 | -0.00009 |
| | and *Model III* | Variance | 0.00106 | 0.00110 | 0.00115 | 0.00125 | 0.00142 | 0.00159 | 0.00163 | 0.00132 |
| | with CO and RH | MSE | 0.00106 | 0.00110 | 0.00115 | 0.00125 | 0.00142 | 0.00159 | 0.00163 | 0.00132 |

Note: Analysis results are given only for the coefficient of CO.

***CO and temperature/RH vs. CO: CO dominates Temperature/RH***

Tables 3.11 and 3.12 (row 3) give the results for CO and temperature vs. CO and CO and RH vs. CO, respectively.

For the case of CO and temperature vs. CO (the case of negative correlation), we found that the true model had the smallest MSE. Except for the 59 strata model, MSEs were increasing in the direction from the true model to the finest model. We also found that the variances were strictly increasing in this direction. Coarser models gave large bias but not as large as the two cases mentioned before for negative correlation. Biases seemed to be smaller when the number of strata was large. Means of the estimates always underestimated the true value because of the same reason discussed earlier. *Model III* produced larger MSE than that of the optimal model (true model).

For CO and RH vs. CO (the case of positive correlation), we found that the MSEs were increasing. So the coarser models seemed to be better in this case. The means of the estimates overestimated the true value in all cases. *Model III* produced larger MSE than that of the optimal model (1 stratum model).

***CO vs. CO and temperature/RH***

Finally we investigated the results for CO vs. CO and temperature/RH. In this case, we were analyzing the data with an extra variable that was not responsible for the response in addition to the variable which was truly responsible for the response. The results are given in Tables 3.11 and 3.12 (last row). For both cases, the MSEs increased as the number of strata increased. The coarser models seemed to be better. *Model III* produced MSEs which were not so far away from the optimal model under *Model I.*

## 3.9 Summary of Results and Discussion

The simulation results of our study demonstrated some important features of the three models under consideration. We selected the 8 strata model as the data generator model for our simulation. Seven models for *Model I* were purposively selected for analysis. *Model II* and *Model III* were chosen for comparison purposes.

For *Scenario I* simulation with data generated using a single pollutant (CO), we found that the true model was the optimal model according to the MSE when the baseline intensities were increasing. The cross-validation technique also selected the true model most of the times. So the cross-validation technique can be regarded as a good technique for this situation. Dewanji and Moolgavlar (2002) stated that the asymptotic variance depends on the stratification. If there is more between-subject variation in the baseline parameters in a stratum, more information in the regression coefficient is expected which leads to a smaller variance. We also found that the optimal model became more reliable (smaller variance and MSE) as the within stratum variation in the baseline intensities increased. For the case when the baseline intensities followed the same distribution, the coarser models seemed to be better. According to the MSE, the coarsest model (1 stratum model) was the optimal model when the within stratum variation in the baseline intensities was small, and the true model was optimal when the within stratum variation in the baseline intensities was not small. The cross-validation technique might not perform well when the within stratum variation in the baseline intensities is quite large.

Dewanji and Moolgavkar (2002) stated for *Model I* that the variance increases as the number of strata increases. In our simulation, we found this statement valid when the baseline intensities followed different distributions in different strata. But, for baseline

intensities with the same distribution within each stratum, this statement was no longer true. We found that the variance increased in the direction from the true model to the finest model in this case.

We found for *Scenario I* simulation with data generated by a single pollutant (CO) that the estimates of *Model II* were close to the estimates of the 1 stratum model. For different baseline intensities in different strata, *Model II* produced serious biased estimates resulting in a large MSE. However, *Model III* gave MSEs which were close to those of the optimal model for all cases. We found that when the within stratum variation in the baseline intensities was small (e.g. [0.010, 0.015], [0.015, 0.020], ..., [0.045, 0.050] ), *Model III* was as good as the optimal model. However, when the within stratum variation in the baseline intensities was not too small, the optimal model produced smaller MSE than *Model III.*

*Scenario I* simulation with data generated by a pollutant (CO) and a weather variable (temperature or relative humidity) was performed only for one type of baseline intensities of the form (0.01, 0.03), ..., (0.01, 0.03). We found, on the basis of MSE, that the coarsest model (1 stratum model) was optimal in that case. For the cases of no dominance and covariate dominated pollutant, MSEs were smaller when the pollutant and the covariate were positively correlated compared to negatively correlated pollutant and covariate. The reverse was true when the pollutant dominated the covariate. However, the differences in the MSEs were fairly small. *Model III* also produced estimates close to the true parameter values, but the MSEs were larger than that of the optimal model.

*Scenario II* simulation was done only for one type of baseline intensities of the form (0.01, 0.03), ..., (0.01, 0.03). When we generated the data with pollutant alone (CO) and

analyzed with the pollutant and an extra weather variable (CO and temperature/RH), we found the 1 stratum model to be optimal in most cases. We also found that the MSE was smaller in the case of positive correlation between the pollutant and the weather variable (CO and RH) compared to the case of negative correlation between the pollutant (CO) and the weather variable (temperature). MSEs were larger with an extra weather variable than those without when the pollutant and the weather variable were negatively correlated. When the pollutant and the weather variable were positively correlated, MSEs were larger with an extra weather variable than those without for the coarser stratification models, but the reverse holds for the true model and the finer stratification models. *Model III* also produced a reasonably good estimate in this situation, but not as good as that of the optimal model (1 stratum model).

If we omit the weather variable from the model in the analysis stage (CO and temperature/RH vs. CO), we found that the conclusions differed on the type of relationship between the pollutant and the weather variable and the type of dominance. The conclusions for the case of no dominance and the weather variable dominating the pollutant, were similar. In both of these cases, the finer models seemed to be better. However, the MSE was much smaller for the case when the pollutant and the weather variable were positively correlated than the case when the pollutant and the weather variable were negatively correlated. When pollutant dominates the weather variable, we found that the true model was better for the case of negative correlation (CO and temperature), and the coarser models were better for the case of positive correlation (CO and RH). Here again *Model III* performed reasonable well.

Fung et al. (2003) found by simulation that *Model III* gives a more precise estimate of risk than *Model II* for single failure times. In this study, we also found that *Model III* performs better than *Model II* in terms of MSE, especially when the baseline intensities are different in different strata. Although *Model III* cannot incorporate between subjects variation in the baseline parameters (Dewanji and Moolgavkar, 2000), it can produce results close to those of *Model I*, especially when the within stratum variation in the baseline intensities is small. For real data, it is difficult to find the right stratification model as we might not be certain about the nature of the baseline intensities. We found that an incorrect stratification model could produce a seriously biased estimate resulting in a large MSE. However, if the right stratification model could be selected, then *Model I* would give the best results. So we recommend using *Model III* unless we have a clear perception of the nature of the baseline intensities.

Since *Scenario II* simulation was done only for one type of baseline intensities, the conclusions drawn were not general. Also the cross-validation technique was not applied. So there is much scope for further analysis of *Scenario II* simulation.

We also did not consider the case of multiple pollutant models as well as the fact that there may be multicolinearity among the pollutants. For these cases, the simulation technique can be useful in finding the behavior of the three models considered in this study.

# Chapter 4

*Data Analysis*

The analysis of data was done for the three models described in Chapter 2, namely, *Model I* (Dewanji and Moolgavkar 2000, 2002), *Model II* (Navidi 1998, 2002) and *Model III* (time series analysis using generalized linear model). In this Chapter, we discuss the results obtained from analyzing the Vancouver data using the three methods. The main objective is to see which pollutants are significantly associated with respiratory hospital admissions and to find the extent of the association.

## 4.1 Data Source

The analysis was based on British Columbia air pollution and hospital admission data for respiratory diseases (ICD-9 code 460-519). Hospital admission data for the period of April 01, 1995 to March 31, 1999 were obtained from the B.C. Linked Health Dataset. The database covered residents of greater Vancouver, B.C. The dates of hospital admission for each individual's repeated visits due to respiratory diseases were recorded. The information on each individual's gender was also available.

Daily measurements on air pollutants were available from five to 31 monitoring stations of B.C. between April 01, 1995 to March 31, 1999. Specifically, daily mean pollutant levels on carbon monoxide (CO), coefficient of haze (CoH), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$), and particulate matter 10μm or less in diameter ($PM_{10}$) were available. Daily data on weather conditions, including temperature and relative humidity were obtained from the Vancouver International Airport.

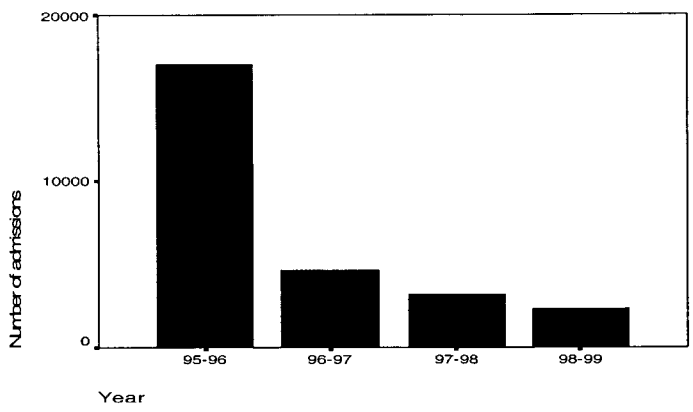
## 4.2 Description of Admission Data

The original admission data consisted of 39,329 individuals and 67,447 admissions. The study population consisted of adults 65 years of age or older in greater Vancouver, B.C. For this analysis, we considered only a cohort of individuals who were admitted during April 01, 1995 to March 31 1996. The cohort consisted of 12,458 (6,499 males and 5,959 females) individuals and was followed up till March 31, 1999 for subsequent detection of admissions.

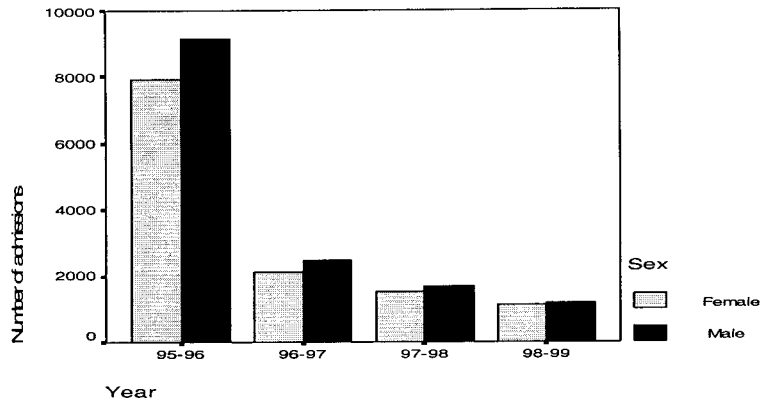
Table 4.1 (a) gives the distribution of admissions per year for the cohort data. A total of 27,162 admissions were recorded during the study period. The number of male admissions is more than that of the female admissions in each year. The first year accounted for most of the admissions (63.0% for males, 62.4% for females and 62.7% for males/females combined), and the number of admissions decreases as time increases. This is due to the fact that there are a lot of people who were only admitted once. Figure 4.1 gives the corresponding bar diagrams.

**Table 4.1 (a):** Distribution of respiratory admissions by year for adults (65 + years) in greater Vancouver, B.C., during 1995-1999.

| Period | Total number of admissions | | |
|---|---|---|---|
| | Male | Female | Male/Female combined |
| April 01, 1995 – March 31, 1996 | 9126 (63.0%) | 7898 (62.4%) | 17024 (62.7%) |
| April 01, 1996 – March 31, 1997 | 2482 (17.1%) | 2135 (16.9%) | 4617 (17.0%) |
| April 01, 1997 – March 31, 1998 | 1689 (11.7%) | 1521 (12.0%) | 3210 (11.8%) |
| April 01, 1998 – March 31, 1999 | 1198 (8.3%) | 1113 (8.8%) | 2311 (8.5%) |
| April 01, 1995 – March 31, 1999 | 14495 (100%) | 12667 (100%) | 27162 (100%) |




Males/Females combined




Males/Females separate

**Figure 4.1:** Bar diagram of respiratory admissions for adults (65 + years) in greater Vancouver, B.C., during 1995-1999.

**Table 4.1 (b):** Descriptive statistics of respiratory admissions by year for adults (65 + years) in greater Vancouver, B.C., during 1995-1999.

| Period | Daily maximum | | | Daily minimum | | | Mean number of admissions per day | | | Standard deviation of the number of admissions | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Total | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| April 01, 1995 – March 31, 1996 | 47 | 44 | 83 | 7 | 7 | 17 | 24.9 | 21.6 | 46.5 | 7.4 | 7.0 | 12.6 |
| April 01, 1996 – March 31, 1997 | 16 | 14 | 29 | 0 | 0 | 2 | 6.8 | 5.8 | 12.6 | 3.1 | 2.7 | 4.7 |
| April 01, 1997 – March 31, 1998 | 13 | 12 | 23 | 0 | 0 | 1 | 4.6 | 4.2 | 8.8 | 2.3 | 2.3 | 3.6 |
| April 01, 1998 – March 31, 1999 | 11 | 11 | 18 | 0 | 0 | 0 | 3.3 | 3.0 | 6.3 | 1.8 | 1.8 | 2.8 |
| April 01, 1995 – March 31, 1999 | 47 | 44 | 83 | 0 | 0 | 0 | 9.9 | 8.7 | 18.6 | 9.8 | 8.5 | 17.8 |

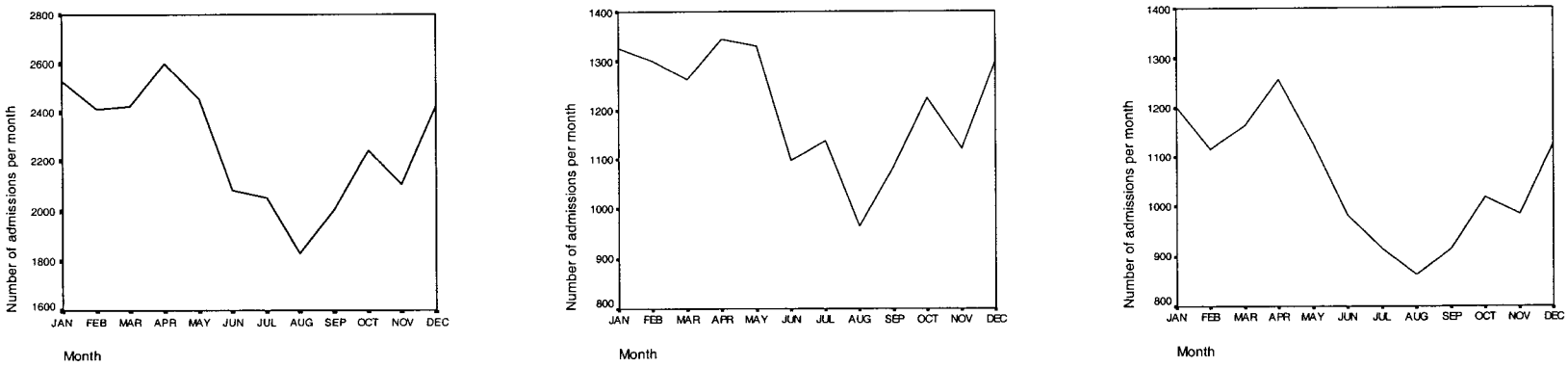
Table 4.1 (b) presents the descriptive statistics for each year as well as for the whole study period. For the entire period, male, female and male/female combined admissions reached a daily maximum of 47, 44 and 83, respectively, and a daily minimum of zero for each case. The means and the standard deviations of the number of admissions per day are 9.9 and 9.8, respectively, for male, 8.7 and 8.5, respectively, for female, and 18.6 and 17.8, respectively, for male/female combined.

Table 4.2 gives the descriptive statistics for the number of daily admissions by month and Figure 4.2 (a) shows the line diagrams. We see that the percent of the number of admissions reached a maximum in April (9.3% for male, 9.9% for female and 9.6% for male/female combined), and then it came down and reached a minimum in August (6.7% for male, 6.8% for female and 6.7% for male/female combined). After August, it again started to increase, but not strictly. Male, female and male/female combined admissions reached a daily maximum in October (47), January and March (44), and January (83), respectively. The mean number of admissions per day reached the highest point in April (11.2, 10.5 and 21.7 for male, female and male/female combined, respectively).

Figure 4.2 (b) presents the line diagram for the number of admissions per day. We see that the first year (day 1 – day 366) accounted for a large number of admissions. After that it showed some cyclic pattern.

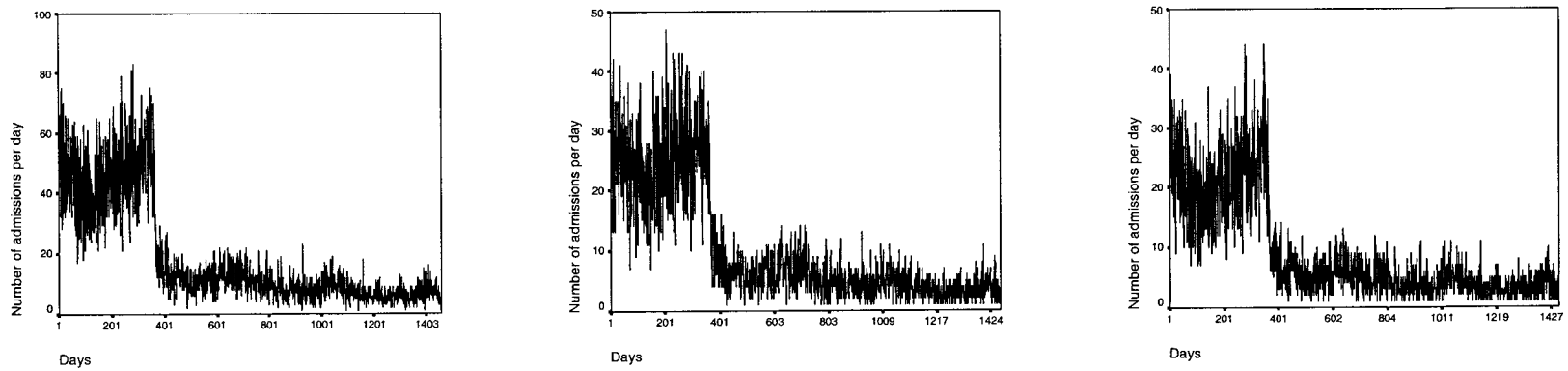
**Table 4.2:** Descriptive statistics of respiratory admissions by month for adults (65 + years) in greater Vancouver, B.C., during 1995-1999.

| Period | Number of admissions | | | Daily maximum | | | Daily minimum | | | Mean number of admissions per day | | | Standard deviation of the number of admissions | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Male | Female | Total | Male | Female | Total | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| January | 1326 (9.1%) | 1201 (9.5%) | 2527 (9.3%) | 41 | 44 | 83 | 1 | 0 | 1 | 10.7 | 9.7 | 20.4 | 10.4 | 9.7 | 19.7 |
| February | 1301 (9.0%) | 1115 (8.8%) | 2416 (8.9%) | 40 | 38 | 73 | 1 | 0 | 4 | 11.5 | 9.9 | 21.4 | 11.3 | 9.0 | 19.8 |
| March | 1263 (8.7%) | 1164 (9.2%) | 2427 (8.9%) | 40 | 44 | 75 | 0 | 0 | 0 | 10.2 | 9.4 | 19.6 | 10.3 | 10.3 | 20.1 |
| April | 1344 (9.3%) | 1256 (9.9%) | 2600 (9.6%) | 42 | 39 | 75 | 0 | 0 | 2 | 11.2 | 10.5 | 21.7 | 9.8 | 9.4 | 18.8 |
| May | 1330 (9.2%) | 1126 (8.9%) | 2456 (9.0%) | 41 | 35 | 65 | 0 | 0 | 1 | 10.7 | 9.1 | 19.8 | 10.1 | 8.5 | 18.0 |
| June | 1098 (7.6%) | 983 (7.8%) | 2081 (7.7%) | 38 | 30 | 59 | 0 | 0 | 2 | 9.2 | 8.2 | 17.3 | 8.4 | 7.5 | 15.4 |
| July | 1137 (7.8%) | 915 (7.2%) | 2052 (7.6%) | 38 | 31 | 63 | 0 | 1 | 2 | 9.2 | 7.4 | 16.5 | 8.9 | 6.7 | 15.2 |
| August | 964 (6.7%) | 863 (6.8%) | 1827 (6.7%) | 28 | 37 | 65 | 0 | 0 | 2 | 7.8 | 7.0 | 14.7 | 7.8 | 7.3 | 14.7 |
| September | 1086 (7.5%) | 916 (7.2%) | 2002 (7.4%) | 40 | 28 | 64 | 0 | 0 | 2 | 9.1 | 7.6 | 16.7 | 9.4 | 7.2 | 16.2 |
| October | 1225 (8.5%) | 1018 (8.0%) | 2243 (8.3%) | 47 | 33 | 69 | 0 | 0 | 0 | 9.9 | 8.2 | 18.1 | 9.8 | 8.0 | 17.2 |
| November | 1122 (7.7%) | 984 (7.8%) | 2106 (7.8%) | 43 | 37 | 79 | 0 | 0 | 1 | 9.4 | 8.2 | 17.6 | 10.2 | 8.6 | 18.4 |
| December | 1299 (9.0%) | 1126 (8.9%) | 2425 (8.9%) | 43 | 32 | 70 | 0 | 0 | 0 | 10.5 | 9.1 | 19.6 | 10.3 | 9.1 | 18.6 |
| Apr 01, 1995 – Mar 31, 1999 | 14495 (100%) | 12667 (100%) | 27162 (100%) | 47 | 44 | 83 | 0 | 0 | 0 | 9.9 | 8.7 | 18.6 | 9.8 | 8.5 | 17.8 |




Males/Females combined Males Females

**Figure 4.2 (a):** Number of respiratory admissions by month for adults (65 + years) in greater Vancouver, B.C., during 1995-1999.




Males/Females combined Males Females

**Figure 4.2 (b):** Daily respiratory admissions for adults (65 + years) in greater Vancouver, B.C., during 1995-1999.

## 4.3 Description of Pollutant and Weather Data

Since 1979, the Index of the Quality of Air (IQUA) has been published for Canadian cities as a single value derived from the measurement of several air pollutants that are averaged from 1 to 24 hours. The IQUA is based on the National Ambient Air Quality Objectives (NAAQOs) (Source: http://www.hc-sc.gc.ca/hecs-sesc/air_quality/ regulations.htm#3), which have been defined under the *Canadian Environmental Protection Act* (CEPA) and which provide a scale for assessing the quality of the air in all parts of Canada. The maximum acceptable levels are intended to provide adequate protection against adverse effects on humans, animals, vegetation, soil, water, materials, and visibility. These levels are 13 ppm (parts per million) per 8 hours for CO, 106 ppb (parts per billion) per 24 hours for $NO_2$ and 115 ppb per 24 hours for $SO_2$. These standards are provided in the last row of Table 4.3 (a). CoH and $PM_{10}$ were not included in the NAAQOs. However, a CoH of less than three is considered clean air and more than five is of some concern (Source: http://www.arb.ca.gov/html/gloss.htm). The recommended reference Level for $PM_{10}$ (24 hour averages), statistically derived on the basis of several key epidemiological studies is 25 $\mu g/m^3$ (micrograms per cubic meter) (Source: http://www.hc-sc.gc.ca/hecs-sesc/air_quality/publications/particulate_ matter_ exec_summary/reference_levels.htm).

Table 4.3 (a) gives the descriptive statistics of the daily mean pollutants and the weather conditions of B.C. for the period from April 01, 1995 to March 31, 1999. The daily average levels for CO, CoH, $NO_2$ and $SO_2$ never exceeded the standards. The daily average level of $PM_{10}$ exceeded the statistically derived reference level in only 72 days out of the 1461 days (entire study period).

**Table 4.3 (a):** Descriptive statistics for the daily mean concentrations of pollutants and the weather conditions in greater Vancouver, B.C., during 1995-1999.

| Descriptive statistics | Pollutants | | | | | Weather conditions | |
|---|---|---|---|---|---|---|---|
| | CO (ppm) | CoH ($\mu g/m^3$) | $NO_2$ (ppb) | $SO_2$ (ppb) | $PM_{10}$ ($\mu g/m^3$) | Relative humidity (%) | Mean temperature ($^0C$) |
| Mean | 0.682 | 0.261 | 16.846 | 3.478 | 13.421 | 78.376 | 10.753 |
| Standard deviation | 0.251 | 0.133 | 4.308 | 1.812 | 6.143 | 9.881 | 5.836 |
| $25^{th}$ percentile | 0.520 | 0.167 | 13.941 | 2.000 | 8.800 | 71.000 | 6.500 |
| $75^{th}$ percentile | 0.760 | 0.300 | 19.313 | 4.500 | 16.800 | 86.000 | 15.700 |
| Inter-quartile range | 0.240 | 0.133 | 5.372 | 2.500 | 8.00 | 15.000 | 9.200 |
| Maximum value | 2.025 | 1.000 | 33.889 | 12.500 | 52.167 | 99.000 | 25.900 |
| Minimum value | 0.275 | 0.000 | 7.222 | 0.000 | 3.769 | 31.000 | -9.200 |
| Maximum acceptable level* | 13 per 8 hrs | - | 106 per 24 hrs | 115 per 24 hrs | - | - | - |

* National Ambient Air Quality Objectives

**Table 4.3 (b):** Monthly averages of pollutants in greater Vancouver, B.C., during 1995-1999.

| Month | Pollutant | | | | |
|---|---|---|---|---|---|
| | CO | CoH | $NO_2$ | $SO_2$ | $PM_{10}$ |
| January | 0.79 | 0.29 | 19.26 | 3.36 | 10.52 |
| February | 0.79 | 0.27 | 19.12 | 3.39 | 11.34 |
| Mach | 0.62 | 0.20 | 17.02 | 2.97 | 10.42 |
| April | 0.59 | 0.21 | 16.57 | 3.25 | 12.91 |
| May | 0.55 | 0.22 | 15.46 | 3.43 | 15.00 |
| June | 0.51 | 0.20 | 14.05 | 3.16 | 13.99 |
| July | 0.56 | 0.24 | 14.58 | 3.75 | 16.65 |
| August | 0.59 | 0.26 | 15.92 | 4.01 | 16.99 |
| September | 0.76 | 0.32 | 17.94 | 4.28 | 17.63 |
| October | 0.76 | 0.31 | 16.04 | 3.45 | 13.64 |
| November | 0.83 | 0.31 | 17.52 | 3.32 | 11.80 |
| December | 0.83 | 0.30 | 18.83 | 3.36 | 10.08 |
| Total | 0.68 | 0.26 | 16.85 | 3.48 | 13.42 |

Figure 4.3 (a) shows the time series plots of the daily mean pollutants. Because of daily fluctuations, it is hard to see a trend. Table 4.3 (b) presents the monthly averages of the daily mean pollutants and Figure 4.3 (b) shows the corresponding plots. From Figure 4.3 (b), we see that CO, CoH and $NO_2$ have similar pattern. These decreased rapidly from February to June, and reached at their minimum in June. Then they began to increase. CO reached at its maximum in November, and after December, it began to decrease again. CoH reached at its maximum in September and then began to decrease. $NO_2$ reached its maximum in January. $SO_2$ and $PM_{10}$ seemed to have similar types of trends. Both of these reached their maxima in September and minima in April.

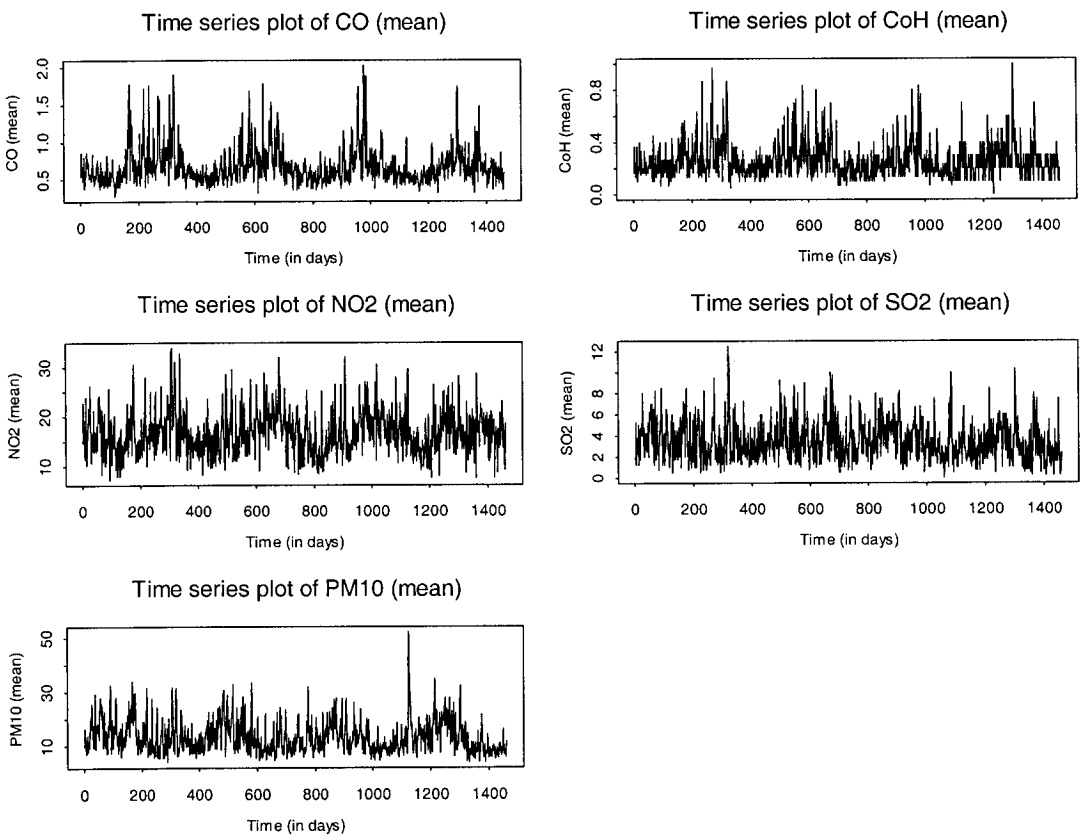



**Figure 4.3 (a):** Time series plot of daily mean pollutants in greater Vancouver, B.C., during 1995-1999.

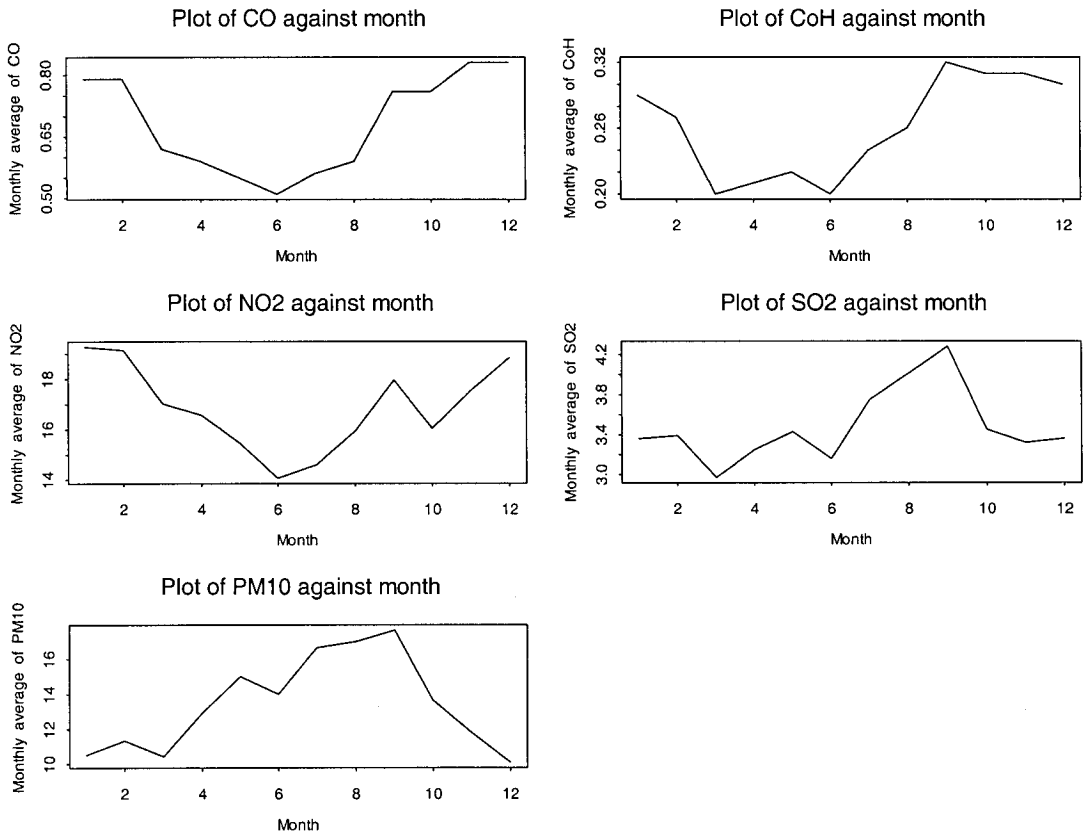



**Figure 4.3 (b):** Monthly averages of pollutants in greater Vancouver, B.C., during 1995-1999.

Table 4.3 (c) presents the correlation matrix of the daily mean pollutants and weather variables. Daily mean temperature and relative humidity are negatively correlated (r = -0.267). Temperature has negative correlations with all the pollutants except $SO_2$ and $PM_{10}$, but relative humidity has positive correlations with all the pollutants except $SO_2$ and $PM_{10}$. Pollutants are positively correlated with each other with CO and CoH most highly correlated (r = 0.853). Correlations between CO and $NO_2$ (r = 0.735) and CoH and $NO_2$ (r = 0.721) are also high.

## 4.4 Analysis under *Model I*

Analysis was done taking one pollutant at a time. In addition to the pollutant, we also included temperature and relative humidity in our analysis. We did not include day

of the week in the model, because a reasonable stratification scheme should control the effects of this variable. In Section 4.4.1, we propose a graphical approach to selecting a reasonable stratification model.

**Table 4.3 (c):** Correlation matrix of daily mean concentrations of pollutants and weather conditions in greater Vancouver, B.C., during 1995-1999.

| | Temperature | Relative humidity | CO | CoH | $NO_2$ | $SO_2$ | $PM_{10}$ |
|---|---|---|---|---|---|---|---|
| Temperature | 1.000 | -0.267 | -0.329 | -0.115 | -0.305 | 0.115 | 0.368 |
| Relative humidity | | 1.000 | 0.271 | 0.147 | 0.073 | -0.005 | -0.311 |
| CO | | | 1.000 | 0.853 | 0.735 | 0.598 | 0.442 |
| CoH | | | | 1.000 | 0.721 | 0.645 | 0.600 |
| $NO_2$ | | | | | 1.000 | 0.579 | 0.545 |
| $SO_2$ | | | | | | 1.000 | 0.610 |
| $PM_{10}$ | | | | | | | 1.000 |

### 4.4.1 A Graphical Approach of Selecting Stratification Models

We found in Chapter 3 that the regression parameter estimates largely depend on the choice of strata. Incorrect stratification can lead to negative estimates of the regression coefficients that are difficult to explain. On the other hand, choice of strata is a very difficult task and also a subjective decision. We also found that if a reasonably correct stratification model could be selected, then the finer models, with each stratum completely contained in exactly one stratum of that reasonably correct model, produced estimates which were fairly close to each other.

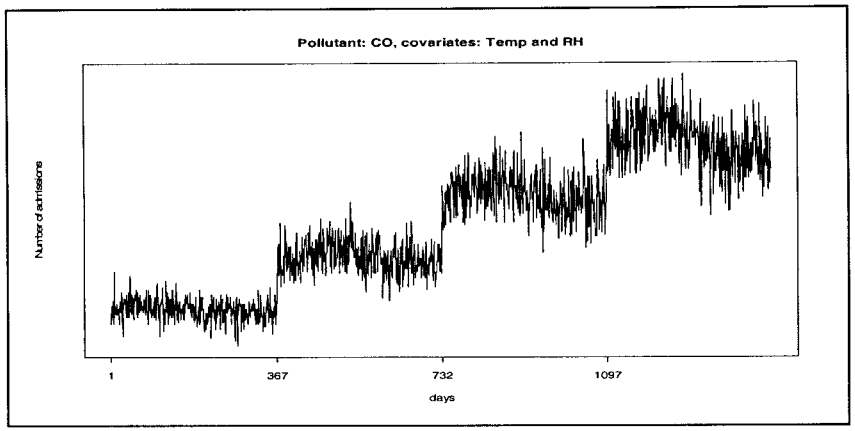
We first formulated different stratification models using some natural candidates like years, seasons, months etc. We found negative estimates of the regression coefficients for every pollutant for at least one stratification model. Negative estimates indicate that an increase in the corresponding air pollutant was associated with decreased hospital admission or that air pollutant had a protective effect. The general consensus is that increases in an air pollutant even at low levels were associated with an increase in hospital admissions. For this reason, the negative estimates ware not biologically plausible and these negative estimates raised questions of interpretation. Also we found that the estimates differ widely for different stratification models. In this Section, we propose a graphical approach of selecting stratification models so that we can avoid difficulties in interpretation of the estimates.

By simulation, we found that the number of admissions per day variable depends on three things: (1) partition of the time period, (2) baseline intensities, and (3) regression parameters. In the plot of admissions per day, the overall shape of the curve varies with the partition and the baseline intensities, and the number of admissions and the shape of the curve within a stratum depend on the regression coefficients.

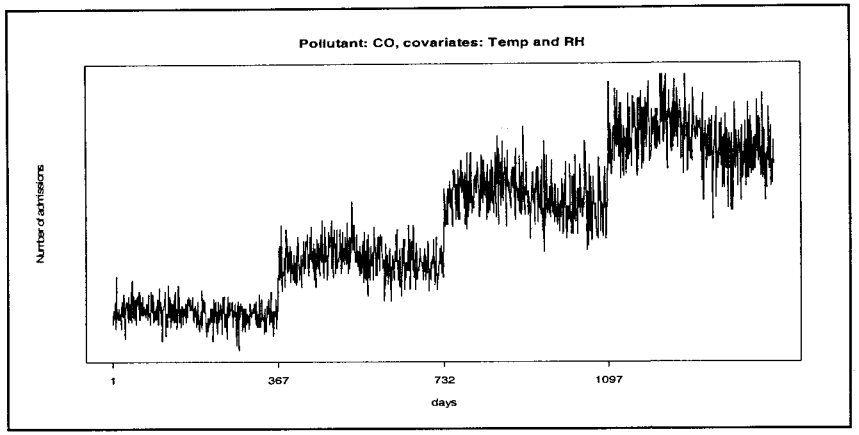
Figure 4.4 (a) and Figure 4.4 (b) present some plots of admissions per day created by the R software using the same simulation technique used in Chapter 3. A four strata model was selected as the base model with partitions [1, 366], [367, 731], [732, 1096], [1097, 1461]. In Figure 4.4 (a), we used the baseline intensities [0.00, 0.04], [0.04, 0.08], [0.08, 0.16], [0.16, 0.20] and in Figure 4.4 (b), [0.00, 0.08], [0.00, 0.02], [0.00, 0.06], [0.00, 0.03]. We see a similar type of pattern within each stratum. As the baseline intensities change from one stratum to another, we notice a jump in the plot. So if we

could select, by simulation, a stratification scheme and the corresponding baseline intensities so that the plot of the number of admissions takes a similar type of shape as the plot of the real admission data, then that stratification model would be a reasonable choice for analysis. By trial and error, such a stratification scheme can be obtained. Then we can formulate some finer stratification models so that each stratum is completely contained in exactly one stratum of the selected model. This should enable the finer models to give estimates which are fairly close to each other.

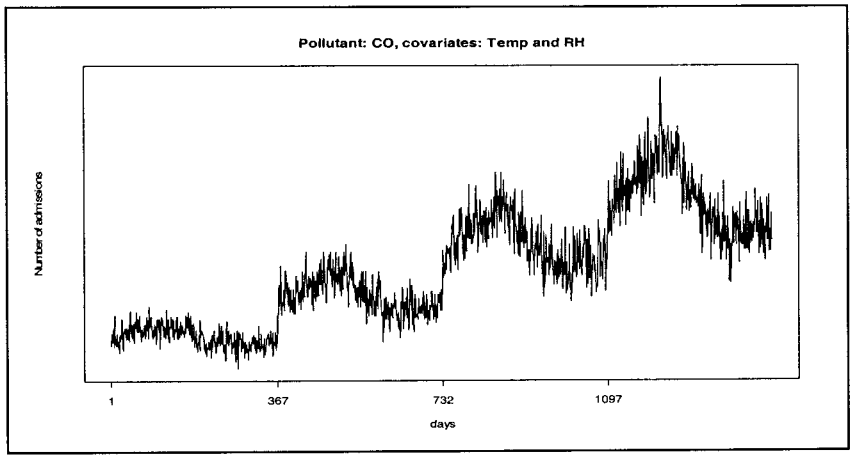
In Figure 4.2 (b), we plotted the number of admissions per day for the real data. To find a reasonable stratification model, we used the simulation technique to generate the number of admissions using information of the pollutant, temperature and relative humidity in such a way that we could get a similar type of plot as in Figure 4.2 (b). Figure 4.4 (c) presents the plots of daily admission predicted by each of the pollutants. It is clear that the plots in Figure 4.2 (b) and Figure 4.4 (c) have similar shapes. The strata used to create Figure 4.4 (c) were [1, 26], [27, 53], [54, 79], [80, 105], [106, 115], [116, 180], [181, 230], [231, 280], [281, 367], [367, 520], [521, 732], [733, 915], [916, 1097], [1098, 1280], [1281, 1461] and the baseline intensities for the corresponding strata were [0.00, 0.12], [0.00, 0.08], [0.00, 0.07], [0.00, 0.06], [0.00, 0.05], [0.00, 0.04], [0.00, 0.05], [0.00, 0.06], [0.00, 0.09], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02]. We selected this stratification model as the base and some finer models were formed on the basis of this base model. In addition to these models, we also used the one stratum model in analysis. Table 4.4 (a) shows the models and the lower bound of the partitions of the study period. After fitting these models using the real data, we applied the cross-validation technique to find the optimal model.



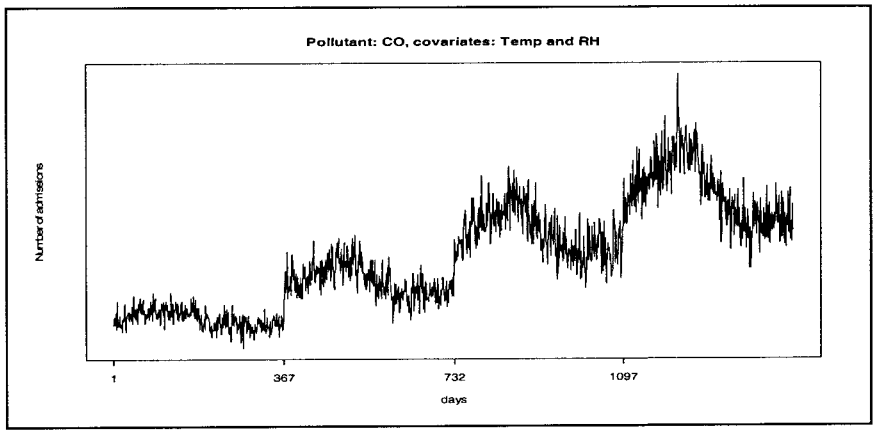

(i) b1 = 0.05, b2 = 0.01 and b3 = 0.03




(ii) b1 = 0.08, b2 = 0.01 and b3 = 0.03

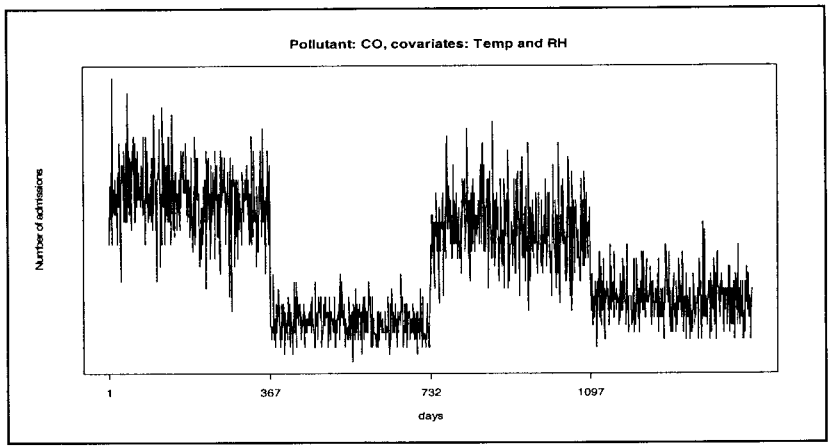



(iii) b1 = 0.05, b2 = 0.03 and b3 = 0.01

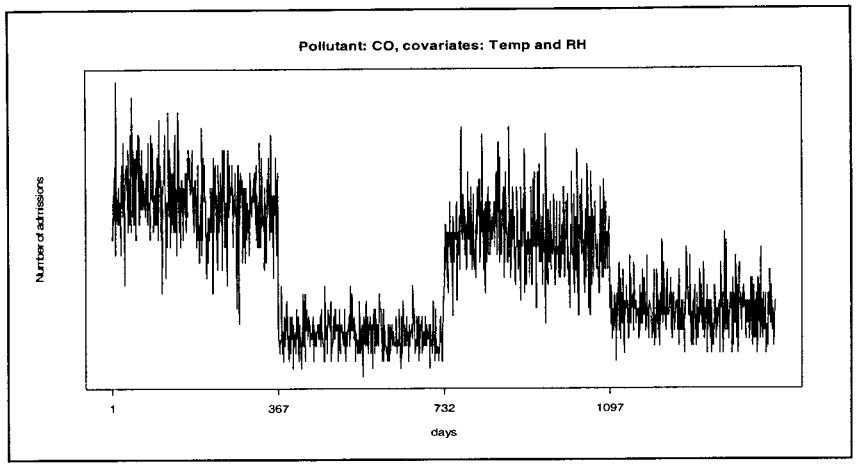



(iv) b1 = 0.08, b2 = 0.03 and b3 = 0.01

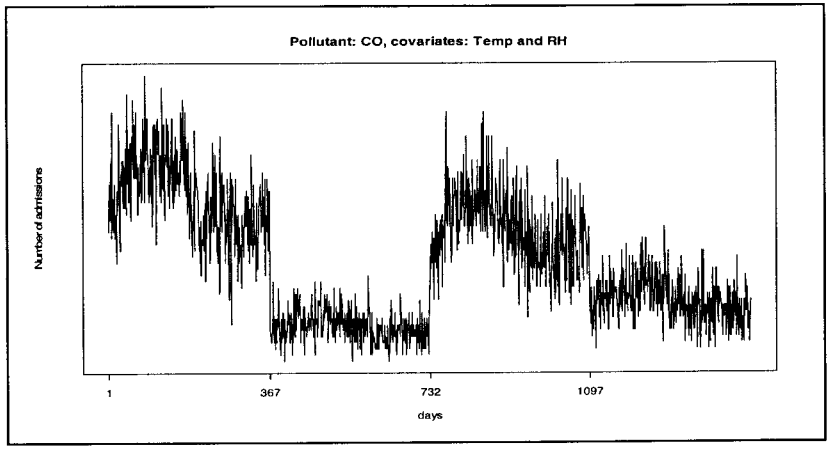
**Figure 4.4 (a):** Plot of hospital admissions generated by simulation (strata: [1, 366], [367, 731], [732, 1096], [1097, 1461], $\lambda$ = [0.00, 0.04], [0.04, 0.08], [0.08, 0.16], [0.16, 0.20]).



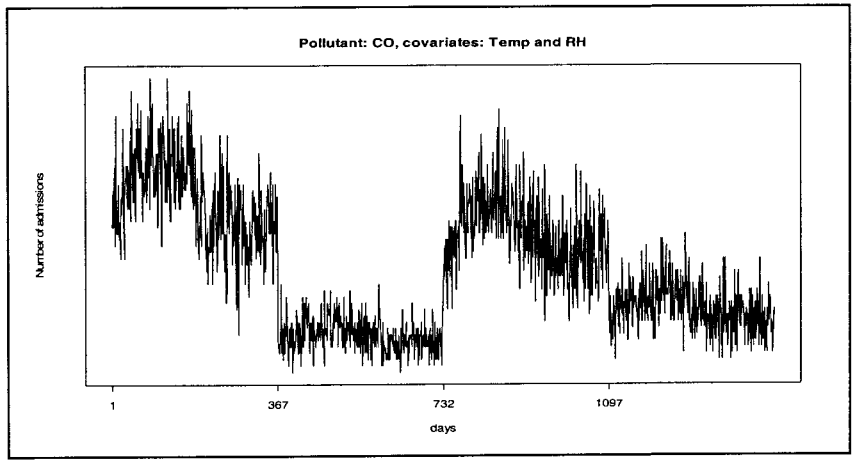

(i) b1 = 0.05, b2 = 0.01 and b3 = 0.03




(ii) b1 = 0.08, b2 = 0.01 and b3 = 0.03

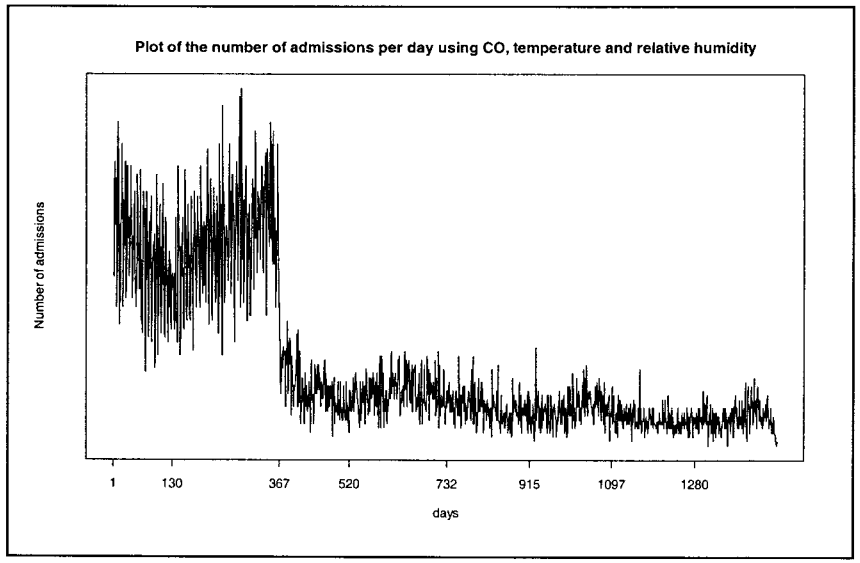


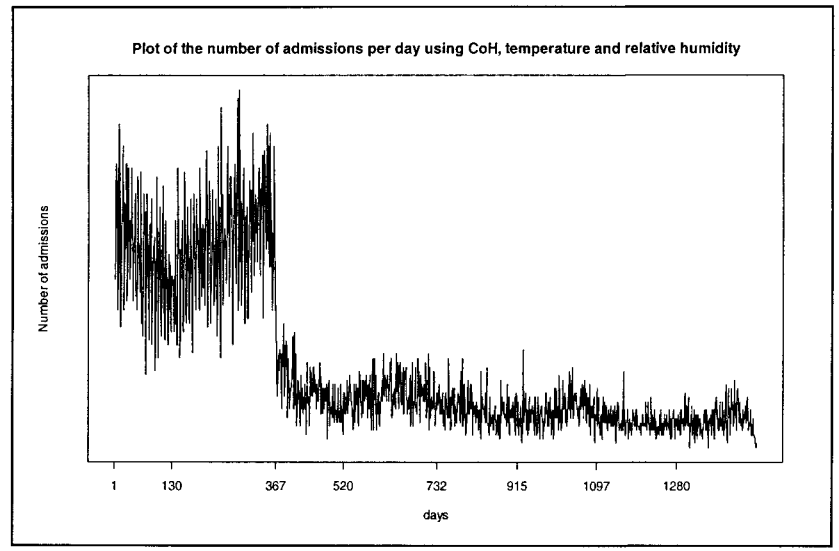

(iii) b1 = 0.05, b2 = 0.03 and b3 = 0.01

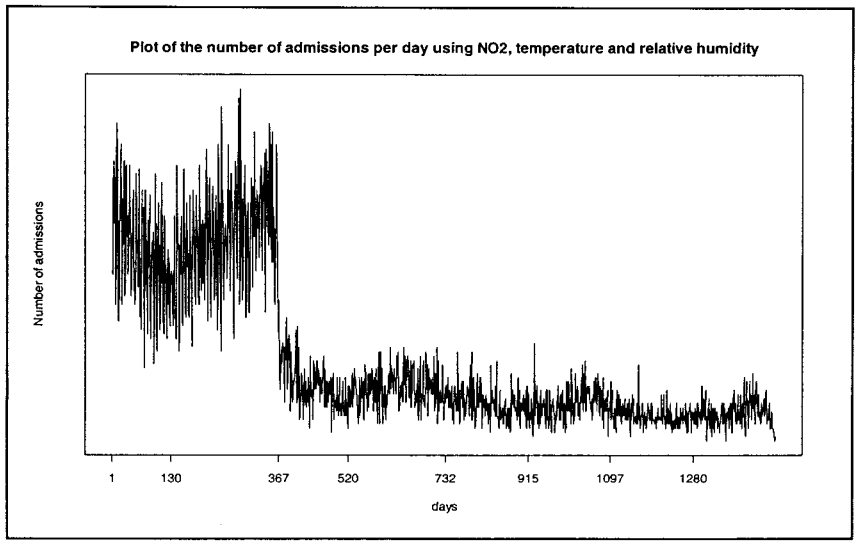


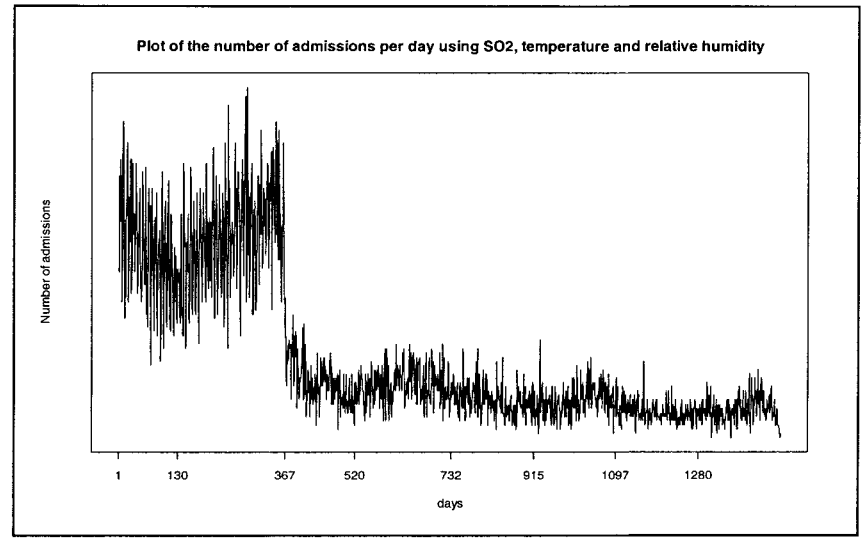

(iv) b1 = 0.05, b2 = 0.03 and b3 = 0.01

**Figure 4.4 (b):** Plot of hospital admissions generated by simulation (strata: [1, 366], [367, 731], [732, 1096], [1097, 1461], $\lambda$ = [0.00, 0.08], [0.00, 0.02], [0.00, 0.06], [0.00, 0.03]).



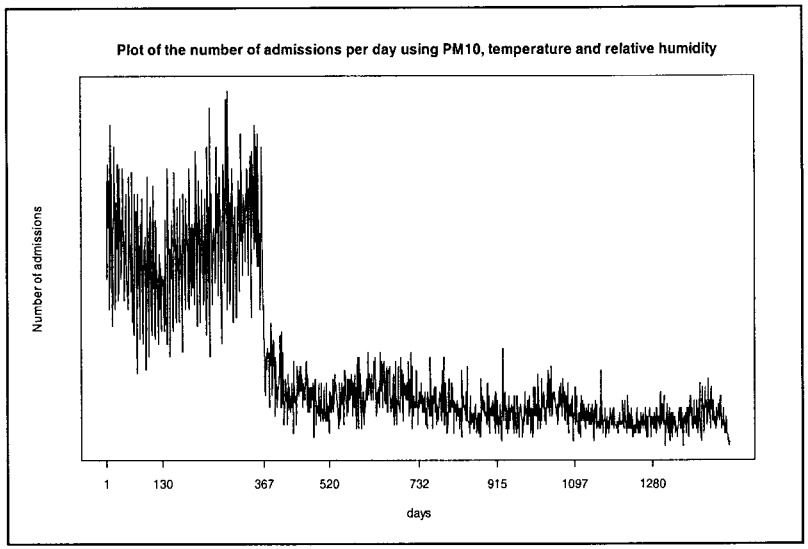



**Figure 4.4 (c):** Plot of hospital admissions generated by simulation using the pollutants and weather variables (strata: [1, 25], [26, 52], [53, 78], [79, 104], [105, 114], [115, 179], [180, 229], [230, 279], [280, 366], [367, 519], [520, 731], [732, 914], [915, 1096], [1097, 1279], [1280, 1461], $\lambda$ = [0.00, 0.12], [0.00, 0.08], [0.00, 0.07], [0.00, 0.06], [0.00, 0.05], [0.00, 0.04], [0.00, 0.05], [0.00, 0.06], [0.00, 0.09], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02], [0.00, 0.02], b1 = 0.05, b2 = 0.03, b3 = 0.01 ).

**Table 4.4 (a):** Selected models and corresponding stratification schemes.

| Model | Lower bounds of the strata in days (observation period = 1461 days) |
|---|---|
| 1 stratum | 1 |
| 15 strata* | 1, 26, 53, 79, 105, 115, 180, 230, 280, 367, 520, 732, 915, 1097, 1280 |
| 22 strata | 1, 26, 53, 79, 105, 115, 180, 230, 280, 367, 400, 460, 520, 626, 732, 823, 915, 1006, 1097, 1188, 1280, 1370 |
| 34 strata | 1, 26, 53, 79, 105, 115, 180, 230, 280, 367, 400, 430, 460, 490, 520, 573, 626, 679, 732, 777, 823, 869, 915, 960, 1006, 1051, 1097, 1142, 1188, 1234, 1280, 1325, 1370, 1415 |
| 39 strata | 1, 26, 53, 79, 105, 115, 150, 180, 210, 230, 250, 280, 320, 367, 380, 400, 430, 460, 490, 520, 573, 626, 679, 732, 777, 823, 869, 915, 960, 1006, 1051, 1097, 1142, 1188, 1234, 1280, 1325, 1370, 1415 |
| 61 strata | 1, 26, 53, 79, 105, 115, 150, 180, 210, 230, 250, 280, 300, 320, 340, 367, 380, 400, 430, 460, 490, 520, 550, 573, 590, 626, 650, 679, 710, 732, 750, 777, 800, 823, 845, 869, 890, 915, 940, 960, 980, 1006, 1025, 1051, 1070, 1097,1120, 1142, 1160, 1188, 1210, 1234, 1260, 1280, 1300, 1325, 1350, 1370, 1390, 1415, 1440 |

* Base model

### 4.4.2 Results

Table 4.4 (b) gives the analysis results of respiratory admission data for each of the pollutants using *Model I*. The results were presented for all six stratification models selected and only for the daily mean pollutants, though we controlled daily mean temperature and relative humidity when fitting the models. We provided the inter-quartile range (IQR) of the pollutants, estimates of the regression parameters, standard errors of

the estimates, p-values for the large sample test of the regression parameters based on a normal approximation, estimates of relative risks (RR) and an approximate large sample 95% confidence interval of the relative risk parameter. The analyses were done using S-Plus software. Relative risks were calculated by

$$RR = \exp(\hat{\beta} \times IQR) \tag{4.1}$$

where $\hat{\beta}$ is the estimate of the regression coefficient and IQR is the inter-quartile range of the pollutant. Relative risks measure how much more likely it is to be hospitalized with an IQR unit increase in the pollutant. The asymptotic 95% confidence interval of the relative risk parameter was calculated using a normal approximation as follows:

$$\left(\exp[(\hat{\beta} - 1.96\,SE) \times IQR], \exp[(\hat{\beta} + 1.96\,SE) \times IQR]\right). \tag{4.2}$$

In Table 4.4 (b), we see that the estimates of the regression coefficients for each of the pollutants were close for the finer models. Also the standard errors were increasing in the direction of the base model (15 strata model) to the finest model (61 strata model). By the cross-validation technique, we found that the 61 strata model was the optimal model for all the pollutants. The estimates, and consequently the relative risks, were significant for the all the pollutants at 5% level of significance. However, $SO_2$ was not significant at the 1% level of significance.

Relative risk of CO for the optimal model was 1.039 with 95% confidence interval (1.02573, 1.05170). Thus, the percentage increase in hospital admissions among the elderly (65 + years) in Vancouver was 3.9% for every increase of 0.24 units (the IQR) in CO. Similarly, relative risks and 95% confidence intervals for CoH, $NO_2$, $SO_2$ and $PM_{10}$ were 1.05569 and (1.04233, 1.06922), 1.06981 and (1.05201, 1.08790), 1.02080 and (1.0039, 1.03795), and 1.03738 and (1.01779, 1.05735), respectively.

**Table 4.4(b):** Results under *Model I* for respiratory admissions of adults (65 + years) in greater Vancouver, B.C., during 1995-1999.

| Pollutant (mean) | Statistics | *Model I* | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 Stratum | 15 strata | 22 strata | 34 strata | 39 strata | 61 strata* |
| CO | Estimate | 0.39395 | 0.10410 | 0.15875 | 0.15926 | 0.14996 | 0.15794 |
| IQR = 0.24 | S.E. | 0.02403 | 0.02386 | 0.02423 | 0.02437 | 0.02573 | 0.02657 |
| | P-value | 0.00000 | 0.00001 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| | Relative risk | 1.09916 | 1.02530 | 1.03884 | 1.03896 | 1.036646 | 1.03863 |
| | 95% Confidence interval | (1.08681, 1.11166) | (1.01385, 1.03687) | (1.02706, 1.05075) | (1.02712, 1.05094) | (1.02418, 1.04927) | (1.02573, 1.05170) |
| CoH | Estimate | 0.54213 | 0.25228 | 0.38788 | 0.39689 | 0.38373 | 0.40646 |
| IQR = 0.13 | S.E. | 0.04526 | 0.04484 | 0.04588 | 0.04644 | 0.04772 | 0.04873 |
| | P-value | 000000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| | Relative risk | 1.07496 | 1.03421 | 1.05308 | 1.05434 | 1.05250 | 1.05569 |
| | 95% Confidence interval | (1.06232, 1.08775) | (1.02216, 1.04640) | (1.04053, 1.06578) | (1.04163, 1.06722) | (1.03945, 1.06570) | (1.04233, 1.06922) |
| $NO_2$ | Estimate | 0.01258 | 0.00938 | 0.01235 | 0.01231 | 0.012168 | 0.01256 |
| IQR = 5.37 | S.E. | 0.00151 | 0.00144 | 0.00147 | 0.00149 | 0.00156 | 0.00159 |
| | P-value | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| | Relative risk | 1.06990 | 1.05165 | 1.06857 | 1.06834 | 1.06754 | 1.06981 |
| | 95% Confidence interval | (1.05300, 1.08707) | (1.03579, 1.06776) | (1.05212, 1.08528) | (1.05176, 1.08519) | (1.05015, 1.08522) | (1.05201, 1.08790) |
| $SO_2$ | Estimate | 0.01135 | 0.00398 | 0.00933 | 0.00930 | 0.00728 | 0.00823 |
| IQR = 2.50 | S.E. | 0.00343 | 0.00320 | 0.00322 | 0.00324 | 0.00330 | 0.00340 |
| | P-value | 0.00092 | 0.21392 | 0.00380 | 0.00412 | 0.02737 | 0.01550 |
| | Relative risk | 1.02879 | 1.00999 | 1.02360 | 1.02353 | 1.01836 | 1.02080 |
| | 95% Confidence interval | (1.01166, 1.04620) | (0.99428, 1.02596) | (1.00756 1.03990) | (1.00739, 1.03992) | (1.00203, 1.03496) | (1.0039, 1.03795) |
| $PM_{10}$ | Estimate | 0.01876 | 0.00156 | 0.00473 | 0.00481 | 0.00405 | 0.00459 |
| IQR = 8.00 | S.E. | 0.00111 | 0.00109 | 0.00111 | 0.00112 | 0.00118 | 0.00122 |
| | P-value | 0.00000 | 0.14234 | 0.00002 | 0.00002 | 0.00062 | 0.00016 |
| | Relative risk | 1.16191 | 1.01287 | 1.03853 | 1.03920 | 1.03294 | 1.03738 |
| | 95% Confidence interval | (1.14181, 1.18236) | (0.99571, 1.03032) | (1.02055, 1.05683) | (1.02105, 1.05766) | (1.01396, 1.05229) | (1.01779, 1.05735) |

* Sixty one strata model was found optimal for all the pollutants by the cross-validation technique.

## 4.5 Analysis under *Model II*

*Model II* was based on the semi-symmetric case-crossover design (Navidi et. al, 2002) with conditional logistic regression model with each subject as one stratum. Hospital admission times were the cases and a single set of control days for each subject was selected. Every control date was two weeks (14 days) away from the case date. We generated a random sample from a uniform[0, 1] distribution to choose the control dates: either before two weeks or after two weeks from the case date. If there were $n_i$ hospital admissions for the ith subject, then we chose $n_i$ control dates for that subject in such a way that a random number less than 0.5 would indicate a control date two weeks before a case date, and a random number greater than 0.5 would indicate a control date two weeks after a case date.

Results are given in Table 4.5. All the pollutants, except $NO_2$, were significant at 5% level of significance. We also see that the results were not similar to those obtained by *Model I*. This was not surprising because *Model II* assumed that the baseline intensities were constant with respect to time and, from the plot of the admission data, we noticed that the baseline intensities were not constant with respect to time.

**Table 4.5:** Results under *Model II* (semi-symmetric case-crossover design) for respiratory admissions of adults (65 + years) in greater Vancouver, B.C., during 1995-1999.

| Pollutant(mean) | Statistics | | | | |
|---|---|---|---|---|---|
| | Estimate | S.E. | P-value | Relative risk | 95% Confidence interval |
| CO | 0.39410 | 0.03378 | 0.00000 | 1.09921 | (1.08188, 1.11681) |
| CoH | 0.44940 | 0.06322 | 0.00000 | 1.06176 | (1.04436, 1.07945) |
| $NO_2$ | 0.00381 | 0.00207 | 0.06500 | 1.02066 | (0.99870, 1.04309) |
| $SO_2$ | 0.03108 | 0.00436 | 0.00000 | 1.08080 | (1.05798, 1.10412) |
| $PM_{10}$ | 0.01353 | 0.00145 | 0.00000 | 1.11429 | (1.08924, 1.13991) |

## 4.6 Analysis under *Model III* (Time Series Analysis)

We fit the generalized log-linear model

$$\ln E(y_t) = \beta_0 + ns(time, df) + D_t \qquad (4.3)$$

where $E(y_t)$ was the mean number of admissions on day t, ns was the natural spline function and $D_t$ was a covariate that controls the effects of day t of a week. We removed the seasonal cycles and the effects of the day of the week so that a time series of the logarithm of hospital admissions could be as close to white noise as possible. We defined $D_t$ as the ratio of the average number of admissions of day t of a week to the average number of daily admissions over the entire study period. Based on the Vancouver data, these seven values were 1.12, 1.16, 1.12, 1.10, 1.01, 0.73 and 0.76 starting with Monday. $D_t$ reflected that there were more admissions on Tuesdays and less on weekends.

We had to choose the degrees of freedom to smooth time by ns. The four choices were 4 knots per year (intervals of 3 months), 6 knots per year (intervals of 2 months), 12 knots per year (intervals of 1 month) and 24 knots per year (intervals of half a month). Table 4.6 (a) gives the number of knots and the corresponding degrees of freedom. These four knots were used in model (4.3). We found that 50 degrees of freedom produced residuals that were closest to white noise and also produced the largest p-value. Then we incorporated the pollutants and weather variables (daily mean temperature and relative humidity (RH)) into the following model:

$$\ln E(y_t) = \beta_0 + \beta_1 \times \text{pollutant} + ns(time,50) + \text{daily mean temp} + RH + D_t \qquad (4.4)$$

**Table 4.6 (a):** Primarily selected degrees of freedom.

| Knots per year | 4 | 6 | 12 | 24 |
|---|---|---|---|---|
| Degrees of freedom | 18 | 26 | 50 | 98 |

Table 4.6 (b) shows the results obtained using GLM in S-plus. We see that except for $SO_2$, all the pollutants were significant. Though the estimates here were pretty close to those obtained under *Model I*, their standard errors were larger than those for *Model I*.

**Table 4.6 (b):** Results under *Model III* (Time series analysis) for respiratory admissions of adults (65 + years) in greater Vancouver, B.C., during 1995-1999.

| Pollutant(mean) | Statistics | | | | |
|---|---|---|---|---|---|
| | Estimate | S.E. | P-value | Relative risk | 95% Confidence interval |
| CO | 0.15100 | 0.03388 | 0.00000 | 1.03691 | (1.02050, 1.05360) |
| CoH | 0.41974 | 0.06193 | 0.00000 | 1.05756 | (1.04060, 1.07480) |
| $NO_2$ | 0.01158 | 0.00200 | 0.00000 | 1.06417 | (1.04200, 1.08680) |
| $SO_2$ | 0.00692 | 0.00437 | 0.11313 | 1.01746 | (0.99590, 1.03950) |
| $PM_{10}$ | 0.00413 | 0.00151 | 0.00629 | 1.03363 | (1.00940, 1.05840) |

## 4.7 Discussion

We evaluated the effects of environmental pollutants on repeated hospital admissions of respiratory diseases among the elderly (65 + years) in Vancouver, Canada, for the period of April 01, 1995 to March 31, 1999. Three statistical methods were employed: *Model I* (Dewanji and Moolgavkar, 2000), *Model II* (Navidi, 1998) and *Model III* (time series analysis). The results of *Model I* and *Model III* were similar while *Model II* produced slightly different results. This was because *Model II* was based on a constant baseline intensity with respect to time, which might not be the case for our data. *Model I*

showed significant effects of all the pollutants under study (CO, CoH, $NO_2$, $SO_2$ and $PM_{10}$) with hospital admissions of respiratory diseases. *Model III* gave the same conclusion except for $SO_2$. We found that *Model I* gave shorter confidence intervals of the relative risk parameters for all the pollutants compared to *Model III*. In this respect, *Model I* can be regarded superior to the other two models.

Since CO has strong correlation with other pollutants regularly used in air pollution studies, the effects of CO are often obscured and difficult to examine independently of other pollutants (Burnett et al., 1999). Previous findings regarding the association between CO and respiratory admissions were inconsistent. Atkinson et al. (1999) found no significant association between CO and respiratory admissions in London. Luginaah et al. (2004) found one and two day delayed effects of CO on respiratory admissions for females 0-14 years of age in Windsor, Ontario. Controlling for seasonal and temperature effects, Cho et al. (2000) in Korea found significant association between CO and hospital admissions for respiratory diseases. Dewanji and Moolgavkar (2000) also observed significant association between CO and hospital admissions for chronic respiratory diseases in King County, Seattle, Washington, over the period 1990-1995. In this study, we found significant association between CO and respiratory admissions. Both *Model I* and *Model III* gave relative risk estimate 1.04 with 95% confidence interval (1.02, 1.05).

The effects of CoH on respiratory admissions were rarely examined before, as compared to the other pollutants (Goldberg et al. 2001a). However, Burnett et al. (1997) found that CoH was the strongest predictor of hospitalizations for respiratory diseases among particle-related pollutants. Luginaah et al. (2004) observed significant association between CoH and female respiratory hospitalization in Windsor, Ontario. In this study,

we also found CoH to be highly significant. The relative risk estimate was 1.06 with 95% confidence interval (1.04, 10.7) under both *Model I* and *Model III.*

Although $NO_2$ is known to increase susceptibility to respiratory infections (Speizer et al. 1980), conclusions from previous studies were inconsistent. Atkinson et al. (1999) found no significant effects of $NO_2$ on respiratory admissions for three age groups (0-14, 15-64 and 65 +) in London. Spix et al. (1998) reported in Air Pollution and Health: A European Approach (APHEA) that there was no significant association between $NO_2$ and hospital admissions due to respiratory diseases. In Paris, France (Dab et al. 1996) and in Birmingham, England (Wordley et al. 1997), a lack of association between $NO_2$ and respiratory admissions was observed. On the other hand, Wong et al. (1999) found significant association between $NO_2$ and respiratory admissions for age groups 0-4, 5-64 and 65 + in Hong Kong. Ponce de Leon et al. (1996) found significant association between $NO_2$ lag 2 and respiratory admissions for children 0-14 years of age in London, England. Luginaah et al. (2004) observed significant association between $NO_2$ lag 2 and respiratory admissions for females 0-14 years of age, but not for any of the other female or male groups in Windsor, Ontario. In this study, we found significant association between $NO_2$ and respiratory hospital admissions. The relative risk estimates were 1.07 and 1.06 with 95% confidence intervals (1.05, 1.08) and (1.04, 1.09) under *Model I* and *Model III*, respectively.

Next we consider $SO_2$. Spix et al. (1998), Sunyer et al. (2003) and Wordley et al. (1997) reported no significant effects of $SO_2$ on hospital admissions due to respiratory diseases. Ponce de Leon et al. (1996) in London, England also found no significant association for the 65 + age group. On the other hand, studies in Milan , Italy (Vigotti et

al. 1996), in Paris, France (Dab et al. 1996) and in London, England (Walters et al. 1994) found significant association between $SO_2$ and respiratory hospital admissions. Atkinson et al. (1999) reported strong association between $SO_2$ and respiratory admission among 0-14 years olds in London. Wong et al. (1999) found significant association among the elderly 65 +, but not among younger age groups in Hong Kong. Bates and Sizto (1987) observed significant association between $SO_2$ lag 2 and respiratory admissions in Southern Ontario. Luginaah et al. (2004) found significant association between $SO_2$ (lag 1) and respiratory admissions for females 0-14 years of age in Windsor, Ontario. We also found inconsistent results for $SO_2$ under *Model I* and *Model III*. *Model I* showed significant association between $SO_2$ and respiratory admissions (relative risk = 1.021 with 95% confidence interval (1.004, 1.038)), but *Model III* did not (relative risk = 1.017 with 95% confidence interval (0.996, 1.040)).

Several previous studies found significant association between $PM_{10}$ and hospitalizations for respiratory diseases. Burnett et al. (1999) in Toronto and Wong et al. (1999) in Hong Kong found $PM_{10}$ to be significantly associated with respiratory admissions. Schwartz (1996) observed significant association between $PM_{10}$ and respiratory admissions among women 65 years of age and older in Spokane, Washington. Atkinson et al. (1999) found significant effects of $PM_{10}$ on respiratory admissions for all age groups in London. Dewanji and Moolgavkar (2000) observed significant association between $PM_{10}$ and hospital admissions for chronic respiratory diseases in King County, Seattle, Washington. Luginaah et al. (2004) found significant association between $PM_{10}$ (lag 2) and respiratory admissions of males 15-64 years of age in Windsor, Ontario. In the present study, we also found significant association between $PM_{10}$ and respiratory

hospital admissions. The relative risks and the 95% confidence intervals under *Model I* and *Model III* were 1.037 and 1.033, and (1.018, 1.057) and (1.009, 1.058), respectively.

Although we declared that Model I performs the best in analyzing recurrent event data with environmental pollutants, we also found difficulties in the formulation of the stratification models. The estimates are sensitive to the stratification schemes used. Misleading and uninterpretable results may result from the use of improper stratification models. The graphical approach we proposed requires subjective decision. Also it is not an easy task to form the stratification models by trial and error method. Further studies can be conducted to find some methods for the formulation of stratification models. Nevertheless, we found the graphical approach effective in deciding the stratification models in our analysis.

We did not analyze the lag effects of different pollutants in our study. Also, a single pollutant model was used for inference purposes without controlling for sex. So there is scope for further analysis. Finally, these results pertain only to the effects of environmental pollutants on repeated hospital admissions of respiratory diseases among the elderly (65 + years) in Vancouver, B.C, during 1995-1999. Hence, these results must be interpreted with caution. Nevertheless, the findings still have implications for public health policy.

# References


Atkinson RW, Bremner SA, Anderson HR, Strachan DP, Bland JM, Ponce de Leon A. (1999). Short-term associations between emergency hospital admissions for respiratory and cardiovascular disease and outdoor air pollution in London. *Arch Environ Health* **54**: 398-411.

Bates CV, Sizto R. (1987). Air pollution and hospital admissions in Southern Ontario: the acid summer haze effect. *Environ Res* **43**: 317-331.

Breslow NE, Day NE. (1980). *Statistical Methods in Cancer Research*, Volume I, *The Analysis of Case-Control Studies. IARC Scientific Publications*, 32. Lyon: International Agency for Research on Cancer.

Burnett RT, Bartlett S and Krewski D, Raad-Young M. and Roberts G. (1994a). Air pollution effects on hospital admissions: a statistical analysis of parallel time series. *Environ and Ecological Stat* **1**: 325-332.

Burnett RT, Dales RE, Raizenne ME, Krewski D, Summers PW, Roberts GR, Raad-Young M, Dann T, Brook JR. (1994b). Effect of low ambient levels of ozone and sulfates on the frequency of respiratory admissions to Ontario hospitals. *Environ Res* **65**: 172-194.

Burnett RT, Dales RE, , Krewski D, Vincent Renaud, Dann Tom, Brook JR. (1995). Association between ambient particulate sulfate and admissions to Ontario hospitals for cardiac and respiratory disease. *Am J of Epidemiol* **142**:15-22.

Burnett RT, Brook JR, Yung WT, Dales RE, Krewski D. (1997a). Association between ozone and hospitalization for respiratory diseases in 16 Canadian cities. *Environ Res* **72**: 24-31.

Burnett RT, Cakmak S, Brook JR, Krewski D. (1997b). The role of particulate size and chemistry in the association between summertime ambient air pollution and hospitalization for cardiorespiratory disease. *Environ Health Prespect* **105**: 614-20.

Burnett RT, Smith-Doiron M, Stieb D, Cakmak S, Brook JR. (1999). Effects of particulate and gaseous air pollution on cardiorespiratory hospitalizations. *Arch Environ Health* **54**: 130-139.

Cakmak S, Burnett RT, Krewski D. (1998). Adjusting for temporal variation in the analysis of parallel time series of health and environmental variables. *J Exposure Anal Environ Epidemiol* **8**: 129-144.

California Air Resource Board. Glossary of Air Pollution Terms. May 2003. (http://www.arb.ca.gov/html/gloss.htm).

Canadian Lung Association. Outdoor Air Quality. Pollutants. 2004 (http://www.lung.ca/cando/pollutants.html).

Cho B, Choi J, Yum YT. (2000). Air pollution and hospital admissions for respiratory disease in certain areas of Korea. *J Occup Health* **42(4)**: 185-191.

Cleanair.ca. Clean Air Handbook. Pollution Science. Science Facts. (http://www.cleanair.ca/science_facts.html).

Dab W, Medina S, Quenel P, Le Moullec Y, Le Tertre A, Thelot B, Monteil C, Lameloise P, Pirard P, Momas I, Ferry R, Festy B. (1996). Short term respiratory health effects of ambient air pollution: results of the APHEA project in Paris. *Journal of Epidemiology and Community Health* **50 (Suppl 1)**: S42-S46.

Dewanji A, Moolgavkar SH. (2000). A Poisson process approach for recurrent event data with environmental covariates. Environmetrics **11**: 665-673.

Dewanji A, Moolgavkar SH. (2002). Choice of stratification in Poisson process analysis of recurrent event data with environmental covariates. *Statistics in Medicine* **21**: 3383-3393.

Dockery DW, Schawartz J, Spengler JD. (1993). Air pollution and daily mortality: Association with particulates and acid aerosols. *Environ Res* **9**:362-73.

Dominici F, McDermott A, Zeger SL, Samet JM. (2002). On the use of generalized additive models in time series of air pollution and health. *Am J Epidemiol* **156**: 193-203.

Fung KY, Krewski D, Chen Y, Burnett RT, Cakmak S. (2003). Comparison of time series and case-crossover analyses of air pollution and hospital admission data. *Int J of Epidemiol* **32**: 1064-1070.

Fung KY, Luginaah I, Gorey KM, Webster G. (2004). Air pollution and daily hospital admissions for Cardiovascular Diseases in Windsor, Ontario. *Canadian Journal of Public Health*. (In Press).

Greater Vancouver Regional District. Air. 2003. (http://www.gvrd.bc.ca/air/index.htm)

Greater Vancouver Regional District. Air. Emissions. 2003. (http://www.gvrd.bc.ca/air/emissions.htm).

Goldberg MS, Burnett R, Bailar JC III, Brook J, Bonvalot Y, Tamblyn R, Singh R, Valois M-F. (2001a). The association between daily mortality and short-term effects of

ambient air particle pollution in Montreal, Quebec: 1. Nonaccidental mortality. *Environ Res* **A86**: 12-25.

Goldberg MS, Burnett RT, Bailar JC, Brook J, Bonvalot Y, Tamblyn R. Singh R, Valois M-F, Vincent R. (2001b). The association between daily mortality and ambient air particle pollution in Montreal, Quebec: 2. Cause-specific mortality. *Environ Res* **A86**: 26-36.

Health Canada. Health Environment and Consumer Safety. Environmental Contaminants. Health and Air Quality. Health Effects. Leading Causes of Hospitalization. December 2003. (http://www.hc-sc.gc.ca/hecs-sesc/air_quality/health_effects.htm#7).

Health Canada. Health Environment and Consumer Safety. Environmental Contaminants. Health and Air Quality. Let's Talk About Air. Airborne Particles. December 2003. (http://www.hc-sc.gc.ca/hecs-sesc/air_quality/talk.htm#airborne).

Health Canada. Health Environment and Consumer Safety. Environmental Contaminants. Health and Air Quality. Let's Talk About Air. Nitrogen. December 2003. (http://www.hc-sc.gc.ca/hecs-sesc/air_quality/talk.htm#nitrogen).

Health Canada. Health Environment and Consumer Safety. Environmental Contaminants. Health and Air Quality. Let's Talk About Air. Nitrogen. December 2003. (http://www.hc-sc.gc.ca/hecs-sesc/air_quality/talk.htm#sulphur).

Health Canada. Health Environment and Consumer Safety. Environmental Contaminants. Health and Air Quality. Publications. General Publications. National Ambient Air Quality Objectives for Particulate Matter, Executive Summary [HTML]. Identification of Reference Levels. June 2003. (http://www.hc-sc.gc.ca/hecs-sesc/air_quality/publications/particulate_matter_exec_summary/reference_levels.htm).

Health Canada. Health Environment and Consumer Safety. Environmental Contaminants. Health and Air Quality. Regulations. National Ambient Air Quality Objectives (NAAQOS). January 2003. (http://www.hc-sc.gc.ca/hecs-sesc/air_quality/regulations.htm#3).

Kalbfleish JD, Prentice RL. (1980). *The Statistical Analysis of Failure Time Data.* Wiley: New York.

Kinney PL, Ozkaynak H. (1997). Association of daily mortality and air pollution in Los Angeles County. *Environ Res* **54**: 99-120.

Langholz BM, Goldstein L. (1997). *Fitting logistic models using the conditional logistic likelihood when there are large strata.* Technical Report 113, Department of Preventive Medicine, University of Southern California, Los Angeles, California.

Lawless JF, Nadeau JC. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**: 158-168.

Lin M, Chen Y, Burnett RT, Villeneuve P and Krewski D. (2002). Effect of short-term exposure to gaseous pollution on asthma hospitalization in children: a bi-directional case-crossover analysis. *J Epidemiol Comm Health* **57**: 50-55.

Lin M, Chen Y, Villeneuve PJ, Burnett RT, Lemyre L, Hertzman C, McGrail KM, Krewski D. (2004). Gaseous Air Pollutants and Asthma Hospitalization of Children with Low Household Income in Vancouver, British Columbia, Canada. *Am J of Epidemiol* **159(3)**: 294-303.

Lipfert FW, Hammerstrom T. (1992). Temporal patterns in air pollution and hospital admission. *Environ Res* **59**: 374-399.

Lohr S. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Luginaah IN, Fung KY, Gorey KM, Webster G, Wills C. (2004). Association of ambient air pollution with respiratory hospitalization in a Government designated 'Area of Concern': The case of Windsor, Ontario. *Environmental Health Perspectives*. (Accepted).

Maclure M. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *Am J Epidemiology* **133**: 144-153.

McCullagh P, Nelder JA. (1989). *Generalized Linear Model*. Chapman and Hall.

Navidi W. (1998). Bidirectional case-crossover designs for exposures with time trends. *Biometrics* **54**: 596-605.

Navidi W, Weinhandl E. (2002). Risk set sampling for case-crossover designs. *Epidemiology* **13(1)**: 100-105.

Neas LM, Schwartz J, Dockery D. (1999). A Case-Crossover Analysis of Air Pollution and Mortality in Philadelphia. *Environ Health Perspect* **107**:629-631.

Pepe MS, Chi J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**: 811-820.

Ponce de Leon A, Anderson HR, Bland JM, Strachan DP. (1996). Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92. *J Epidemiol Community Health* **33(Suppl)**: S63-S70.

Pope CA III, Bates D, Raizenne M. (1995). Environmental Health Issues. *Environmental Health Perspect* **103**: 472-480.

Prentice RL, Williams BJ, Peterson AV. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**: 373-379.

Respiratory Disease in Canada (2001). Canadian Institute for Health Information, Canadian Lung Association, Health Canada, Statistics Canada. (http://www.phac-aspc.gc.ca/publicat/rdc-mrc01/pdf/rdc0901e.pdf).

Schwartz J. (1994). Air pollution and hospital admissions for the elderly in Detroit, Michigan. *Am J Respir Crit Care Med* **150**: 648-655.

Schwartz J. (1996). Air pollution and hospital admissions for respiratory disease. *Epidemiology* **7(1)**: 20-28.

Schwartz J, Morris R. (1995). Air pollution and hospital admissions for cardiovascular disease in Detroit, Michigan. *Am J Epidmiol* **142(1)**: 23-35.

Speizer FE, Ferris B JR, Bishop YM, Spengler J. (1980). Respiratory disease rates and pulmonary function in children associated with $NO_2$ exposure. *Am Rev Respir Dis* **121**: 3-10.

Spix C, Anderson HR, Schwartz J, Vigotti MA, Le Tertre A, Vonk JM, et al. (1998). Short term effects of air pollution on hospital admissions of respiratory disease in Europe. A quantitative summary of APHEA study results. *Arch Environ Health* **53**: 54-64.

*S-Plus 2000: Modern Statistics and Advanced Graphics. Guide to Statistics (Volume I and II)*. (2003). Data Analysis Products Division. MathSoft, Inc. Seattle, Washington.

*S-Plus 2000: Modern Statistics and Advanced Graphics. Programmer's Guide*. (2003). Data Analysis Products Division. MathSoft, Inc. Seattle, Washington.

Statistics Canada. Search the Website: Respiratory Diseases. All respiratory disease deaths (ICD-9 460-519), age-standardized rate per 100,000 population and confidence interval, by sex, Canada, provinces, territories, health regions and peer groups, 1997. June 2004. (http://www.statcan.ca/english/freepub/82-221-XIE/00604/tables/html/14153_97.htm).

Sunyer J, Atkinson R, Ballester F, Le Tertre A, Ayres JG, Forastiere F, Forsberg B, Vonk JM, Bisanti L, Anderson RH, Schwartz J, Katsouyanni K. (2003). Respiratory effects of sulphur dioxide: a hierarchical multicity analysis in the APHEA 2 study. *Occup. Environ. Med* **60**: 2.

The R Development Core Team (2004). *R: A Language and Environment for Statistical Computing. Reference Index. Version 2.0.1.*

The R Development Core Team (2004). *R Data Import/Export. Version 2.0.1.*

Venables WN, Smith DM, The R Development Core Team (2004). *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics. Version 2.0.1.*

Vigotti MA, Rossi G, Bisanti L, Zanobetti A, Schwartz J. (1996). Short term effects of urban air pollution on respiratory health in Milan, Italy. *Journal of Epidemiology and Community Health* **50 (Suppl 1):** S71-S75.

Walters S, Griffiths R, Ayres J. (1994). Temporal association between hospital admissions for asthma in Birmingham and ambient levels of sulphur dioxide and smoke. *Thorax* **49**: 133-140.

Wei LJ, Lin DY, Weissfeld L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**: 1065-1073.

Wojtniak B, Piekarski T. (1996). Short term effect of air pollution on mortality in Polish urban populations – what is different. *Journal of Epidemiology and Community Health* **50 (Suppl 1)**: S36-S41.

Wong TW, Lau TS, Yu TS, Neller A, Wong SL, Tam W, et al. (1999). Air pollution and hospital and cardiovascular diseases in Hong Kong. *Occup Environ Med* **56**: 679-683.

Wordley J, Walters S, Ayres JG. (1997). Short term variations in hospital admissions and mortality and particulate air pollution. *Occup Environ Med* **54**: 108-116.

Zmirou D, Barumandzadeh T, Balducci F, Ritter P, Laham G, Ghilardi JP. (1996). Short term effect of air pollution on mortality in the city of Lyon, France, 1985-90. *Journal of Epidemiology and Community Health* **50 (Suppl 1)**: S30-S35.

# Vita Auctoris

| | |
|---|---|
| Name: | Shahedul Ahsan Khan |
| Place of Birth: | Dhaka, Bangladesh |
| Year of Birth: | 1973 |
| Education: | University of Dhaka, Bangladesh<br>1991-1994 B.Sc. in Statistics |
| | University of Dhaka, Bangladesh<br>1994-1995 M.Sc. in Statistics |
| | University of Windsor, Windsor, Ontario<br>2003-2004 M.Sc. in Statistics |