2009-01-01

A Retrospective Study of Dust Storms and Respiratory Hospitalizations in El Paso, Texas Using a Case-Crossover Study Design

Yanlei Peng
University of Texas at El Paso, pyl_msn@hotmail.com

A RETROSPECTIVE STUDY OF DUST STORMS AND RESPIRATORY HOSPITALIZATIONS IN EL PASO, TEXAS USING A CASE-CROSSOVER STUDY DESIGN

YANLEI PENG

Department of Mathematical Sciences

APPROVED:

Joan Staniswalis, Ph.D., Chair

Ori Rosen, Ph.D.

Sara Grineski, Ph.D.

Patricia D. Witherspoon, Ph.D.
Dean of the Graduate School

To my Parents

A RETROSPECTIVE STUDY OF DUST STORMS AND RESPIRATORY HOSPITALIZATIONS IN EL PASO, TEXAS USING A CASE-CROSSOVER STUDY DESIGN

by

YANLEI PENG

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2009

Acknowledgment

"Everyone has a major weak point. Go through and conquer it! Never give up!" This is what Dr. Joan Staniswalis said to me while I was facing difficulties in my thesis. Her word is a great support and encouragement for me during the writing of the thesis. I took two statistical classes with her, and went to attend a statistics conference under her leadership. Her great enthusiasm and dedication has inspired my determination to pursue a doctor's degree in statistics as the next goal in my life. She stands as a bright example for me.

I also would like to thank other committee members: Dr. Sara E. Grineski and Dr. Ori Rosen. Especially, I want to express my gratitude to Dr. Sara E. Grineski for her financial support, which has been a huge help for me to write this thesis. I wish to thank Dr. Thomas E. Gill, an associate professor in Geological Sciences Department. His comprehensive knowledge in weather and pollutants has given me a guidance to choose correct variables in my thesis analysis. I also extend my appreciation to Maria Barraza, network manager of the Mathematical Sciences Department, for all her help and technical support she gave me at the beginning of writing thesis. I also wish to thank Ms. Poorva Mudgal for assembling part of the weather-pollution data that were utilized in the analysis. My appreciation also extend to Dr. Xiaohui Xu, Department of Epidemiology in the University of Pittsburgh. Thanks for his great help of sending us his SAS codes for fulfilling the reference strategy in my thesis.

At Last, I would like to thank all the people that gave me advising and assistance during the writing of my thesis.

This thesis was supported by Southwest Consortium for Environmental Research and Policy (SCERP) FY2008.

August, 2009.

Abstract

Dust storms frequently occur in El Paso, Texas, but little is known about their respiratory health effects. The association between dust storm and hospitalization for respiratory illness will be studied through a case-crossover study design. The goal is to ascertain the effect of dust storms, adjusted for weather and other pollutants, and to explore the potential harmfulness for different subgroups of residents. A case-crossover study allows patients to serve as their own controls, in order to reduce confounding. The history, relationship, and application of the case-control and case-crossover study designs are reviewed in this thesis. The estimating equations for conditional logistic regression are derived in order to obtain the relative risk for evaluating the impact of dust storms in this case-crossover study design. Primary findings are reported. This thesis contributes to filling the gap in scientific research about the dust storms's effect on respiratory health.

KEYWORDS: Dust storms, case-control, case-crossover, respiratory health.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statement of the Problem

The Chihuahuan Desert region of North America is a significant source of mineral aerosols in the Western Hemisphere, and the Paso del Norte (El Paso, USA/Ciudad Juarez, Mexico) metropolitan area is frequently impacted by the Chihuahuan Desert dust storms. In the early 1990's, the application of econometric time-series studies and prospective cohort studies suggested increased mortality associated with acute (daily) and chronic (decades) exposures to particulate air pollution commonly observed in the developed world [1]. Dust storm, a type of air pollution, emitted from wind-erodible dry land surfaces plays many roles in the climate system and have a significant impact on air quality.

As the 22nd-largest city of any kind in the United States, El Paso, Texas is the biggest US-Mexico border city, with approximate 700,000 residents in the county (U.S. Census, 2000). Living under a qualified air environment is a concern to researchers in the public health and epidemiology field. Dust storms regularly and frequently occur in this region. However, very little is known about their respiratory health effects. This article addresses this gap by combining hospitalization data and weather-pollution data in a case-crossover study design.

1.2 Outline of the Study

The *Case-Crossover Study design* [2] is a popular tool for estimating the effects of serious outcomes (such as asthma hospitalizations) to short-term exposures (such as dust storms).

This method is equivalent to Poisson regression for time-series modeling [3] except that confounding effects of weather and pollutants are controlled by design (by matching) instead of in the regression model. The case-crossover study is a special case of the *case-control study* in which each individual serves as his/her own control. Chapter 2 will now illustrate the details of these two designs. Chapter 3 will provide the statistical methods of the case-crossover study. The effect of dust storms on respiratory health in El Paso county is our main application. Chapter 4 will describe the data resources prior to focusing our attention on the modeling results in chapter 5. Strengths and limitations of this analysis are discussed in chapter 6.

Chapter 2

Case-Crossover Study Design and Its Applications

2.1 Case-Control Study Design

Before introducing the case-crossover study design, it is first very important to understand the case-control study design, because case-crossover is a special case of the case-control design. *Case-Control Study Design* is a mainstay of epidemiologic research. In recent years, the case-control design has proven to be useful and efficient for evaluation of vaccine effectiveness [4], treatment efficacy [5], screening programs [6], and outbreak investigations [7]. The sophisticated use and understanding of case-control studies is the most outstanding methodological development of modern epidemiology [8].

Breslow and Day[9] have pointed out that studying "rare events" such as death from cancer using randomized clinical trials or other controlled prospective studies requires that large populations be tracked for lengthy periods of time to observe disease development. In the case of lung cancer this could involve 20 to 40 years, potentially longer than the careers of many epidemiologists. In addition, these studies, which generally rely on government funding, are unlikely to be supported because of the low likelihood that the population will develop the disease.

The case-control studies use patients who already have disease or other condition and look back to see if there are characteristics of these patients that differ from those who don't have the disease. It is therefore a *retrospective* study, in contrast to the prospective one. Usually, the case-control studies are used to identify exposure factors that may contribute

to a medical condition by comparing subjects who have that condition with patients who do not have the condition but are otherwise similar.

The purpose of the case-control study is to identify the degree of association between risk for certain disease and the exposure factors under study. Before having a close look at this analysis, we first specify the meaning of "case" and "control". *Case* refers to a subject with an event (e.g., respiratory hospital admission). *Control* refers to a subject with no event (e.g., no respiratory hospital admission). In a case-control study, we match several controls with one case having similar characteristics, such as gender, age, and socio-economic status. That is, we use different individuals as cases and controls.

The basic premise of analytic epidemiology is that disease does not occur randomly but rather in describable patterns that reflect the underlying etiology. This rationale certainly applies to case-control studies. Consider two groups, one in which everyone has the disease of interest (cases) and a comparable one in which everyone is free from the disease (controls). The case-control study seeks to identify possible *cause(s)* of the disease by finding out how the two groups differ. That is, because disease does not occur randomly (it occurs rarely), the case group must have been exposed to some factor, either voluntarily (e.g., through diet, exercise, or smoking) or involuntarily (e.g., through such factors as cosmic radiation, air pollution, occupational hazards, or genetic constitution), that contributed to the *causation* of their disease. Therefore, a comparison of the frequency of exposure among cases and controls may permit inferences as to the basis for the difference in disease status [10].

As just mentioned, the objective of case-control studies is to identify differences in exposure frequency that might have effects on one group displaying the outcome of interest (certain disease) and the other group not having it. At this point, we introduce the most frequently calculated measure of effect. The principle is to determine how much more (or less) likely the cases are to be exposed than the controls. In Table (2.1), first consider only the controls. The proportion of the controls exposed is B/(B+D). The proportion of controls not exposed is D/(B+D). The odds of exposure, given that an individual is a

member of the control group, are simply the ratio of these two proportions:

$$B/(B+D) \div D/(B+D) = B/D.$$

This term B/D represents the odds of exposure among the control group. Repeating the same calculations, one determines that the odds of exposure among the case group are A/C. To evaluate whether the odds of exposure for the case group are different from the odds of exposure for the control group, we compute the ratio of these two odds to obtain an *odds ratio* (OR): (A/C)÷(B/D). But this can be more conveniently expressed as (AD)/(BC), which is the cross-product of the cells from our 2 by 2 table.

Table 2.1: Distribution of Exposures in A Case-Control Study Design

		Disease Status	
		Yes	No
Exposure Status	Yes	A	B
	No	C	D
		A+C	B+D

Literally, the odds ratio (OR) measures the odds of exposure for a given disease. An OR of 1.0 implies that a particular exposure is not a risk factor for the disease, because the odds of exposure are equal among the cases and controls. An OR of 2.0 suggests that the cases were twice as likely as the controls to be exposed, and thus this particular exposure is a risk factor for the disease. However, not all risk factors increase risk; a factor with an OR of less than 1.0 would be associated with lower risk of disease (i.e., a protective factor). Note that "hazard ratio" mentioned in chapter 5 (modeling results) is equivalent to odds ratio for the conditional logistic regression model. They both are estimates for the "relative risk", which will be discussed in section 2.2 next.

Here is an example from Lopez-Carrillo (1994) [11] for those of us who have a predilection for spicy foods and have wondered about the health hazards associated with consumption of chili peppers. Lopez-Carrillo (1994) used a population-based case-control design in Mexico City to study the association between chili pepper consumption and gastric cancer risk. Subjects for the study included 220 incident cases and 752 controls randomly selected from the general population. They reported that consumption of chili peppers was significantly related with a high risk for gastric cancer (age and sex adjusted OR = 5.49). The data from this study are abstracted to illustrate how to calculate the OR. From Table (2.2), we can obtain the OR (unadjusted for age and sex):

$$\frac{AD}{BC} = \frac{(204)(145)}{(552)(9)} = 5.95$$

So that Chili pepper consumption puts people at almost 6 times higher risk for gastric cancer as compared to those who refrain from eating Chili peppers.

Table 2.2: The 2 by 2 Table of Lopez-Carrillo's Example (1994)

Chili Pepper Consumption	Case of Gastric Cancer	Controls
Yes	A = 204	B = 552
No	C = 9	D = 145

Confounding is the term used to describe distortion of the exposure effect of interest, because it is associated with the effect of an extraneous factor. It is intimately connected to the concept of causality. The concepts of confounding and bias are most easily understood in the context of cohort studies. A cohort is a subset of the population identified by common characteristics. For example, a birth cohort would consist of all people born in 1990, another cohort would be all people born in 1991, etc.. In a cohort study, information is obtained on exposures before disease status is determined, and all cases of disease arising in a given time period are ascertained. In a cohort study, if some exposure E is associated with disease

status, then the incidence of the disease varies among the strata defined by different levels of E. If these differences in incidence are partially caused by some other factor C, then we say that C has partially confounded the association between E and disease. If C is not causally related to disease, then the differences in incidence cannot be caused by C, thus C does not confound the exposure/disease association. In case-control studies, however, information on exposure is normally obtained after disease status is established, and the cases and controls represent samples from the total. Confounding in a case-control study has the same basis as in a cohort study. Breslow and Day [9] point out that it arises from the association in the causal network in the study population and cannot always be removed by appropriate study design alone. An essential part of the design is an examination of possible confounding effects and how they may be controlled, for example, by matching or by randomization when feasible.

Bias in a case-control study, however, arises from the differences in design between case-control and cohort studies. And it is thus a consequence of the study design. Eliminating bias is one of the goals of a good study design. The effects of bias are often difficult to control in the analysis, although they can be treated by matching or adjusting for covariate effect in the regression model.

Summarizing Breslow and Day [9], confounding will manifest itself similarly in both cohort and case-control studies, and will reflect the causal association between variables in the population under study. In contrast to a cohort study, the essence of case-control studies is *sampling*. Hence the bias, by contrast, is not a property of the study population and should not arise in a cohort study. Bias results from inadequacies in the design of case-control studies, either from the manner in which the data are acquired or in the selection of cases or assignment of controls.

Case-control studies are conducted for efficiency reasons. Under certain circumstances, it may be cumbersome or even impossible to study an entire dynamic population or cohort during a certain time period. Examples include: when the measurement of the causally related factor(s) and other relevant variables (e.g, confounders) is time-consuming, patient-

burdensome, and/or expensive (e.g., when imaging techniques or DNA-material are involved); or when the outcome of interest (e.g., certain disease) is very rare (e.g., anaphylactic shock following the use of a certain drug); or when the time between exposure to the occurrence of the outcome is very long or unknown (for example, the use of Diethylstilbestrol by pregnant women and the occurrence of vaginal carcinoma in their daughters). Instead of studying the cohort in detail, which is all members of the population at risk during the entire followup period, it is more efficient to study only those who develop the outcome of interest/certain disease during the study period (the cases) and a sample of the population at risk without the disease from where the cases emerge (the controls). The exposure of interest and other relevant factors are then measured in cases and controls only. This is the rationale and essence of the case-control study [8].

Before finishing this section, I would like to state a brief history of case-control studies in clinical research. Actually, the case-control method was developed in the field of sociology. According to Grobbee and Hoes (2007) [8], the first case-control study in medicine was published in 1920 by Broders [12]. That paper is studying the role of smoking in the development of epithelioma of the lip. Smoking habits of 537 patients with epithelioma of the lip were compared to 500 patients without epithelioma. The proportion of pipe smokers was much higher in the cases (78%) than in the controls (38%), although tobacco use was similar in both groups (79% and 80%, respectively). In this first case-control study, the author not only did not explain how controls were sampled, but did not give additional characteristics of the control group. In addition, no discussion on possible confounding was included, and no formal measure of association between pipe smoking and lip carcinoma was calculated. However, it is understandable for this analysis, because it took an additional 30 years for the odds ratio to be introduced and eight more years before a method to adjust for confounding was first described [8]. Nevertheless, in other studies, a causal association between pipe smoking and epithelioma of the lip was later confirmed.

2.2 Case-Crossover is A Special Case of the Case-Control Study Design

Earlier in this chapter, we stated that the ideal control is a person free of the disease of interest and similar in every respect to the case except for the exposure of interest. This objective is not only hard to achieve but also difficult to evaluate. As part of the study, one approach is that we choose relatives, associates, or neighbors of cases as the control group. This is a good method to control for possible differences in socioeconomic status, education, or other characteristics assumed to be determinants of friendship or neighborhood.

However, what if I tell you that there is another approach using the cases serving as their own controls? This kind of analysis was first developed by Maclure (1991) [2], as a cross between the case-control design in which referents are sampled from the case's own history, and the cross-over design in which exposure is switched according to the interests of the study hypothesis. The original motivation for posing case-crossover methods appropriate for the analysis of environmental data was to control for temporal trends, while estimating association between a short-term exposure, as occurs with air pollution, and the risk of a rare event, such as death or a respiratory event. The exposure of an individual to a pollutant immediately prior to some catastrophic event (e.g. death, asthma attack) is compared with the exposure of *the same individual to the same pollutant* at other, control or "referent" times. Hence, when we are selecting the control group, we do not need to spend so much effort to match the cases' age, gender or socio-econometric conditions, and we will just simply consider these cases to be their own controls. Therefore, *the case-crossover is a special case of the case-control study design.*

The *Case-Crossover Study Design* [2] is a popular tool for estimating the effects of serious outcomes (such as asthma attack) to short-term exposures (such as dust storms). This method is equivalent to Poisson regression for time-series modeling [3]. In a typical air pollution time series study, for instance, daily event counts (e.g. hospital admission counts) are regressed on the shared weather-pollution exposure series, using a Poisson regression

model. The case-crossover study is equivalent to a Poisson regression analysis except that confounding effects of weather and pollutants are controlled by design (by matching) instead of in the regression model.

In the case-crossover study, each individual who is a "case" serves as his/her own control. That is, we do not need to match the "similar characteristics" in the case-crossover study because we are using the same individual. A case happens during an *index time*, i.e., a hazard or at-risk period of time before the hospitalization. A control happens during a *reference time*, i.e., a defined period of time before and after the hospitalization, e.g. 7 days.

Next, I'll draw an example of case-crossover study. Suppose a patient whose name is Harry was admitted as an asthma hospitalization at an index time. We define two reference days, say 7 days, before and after his hospitalization. This kind of strategy of choosing referents is called *symmetric bi-directional* referent selection. We assume that Harry does not have any asthma admissions in any of the reference days. Then Harry can serve as his own controls on these two reference days. Dust storm and other weather-related exposures at the "index time" are compared to exposures at "referent times". Generally, the case-crossover design is a scientific way of asking and answering the question clinicians so often ask patients: "Were you doing anything unusual just before the episode?" [2]. For Harry, the "episode" is the asthma hospitalization.

Of course, there are other strategies of choosing referents. Besides the symmetric bi-directional referent selection, there is another popular strategy called *time stratified* referent selection. For example, we can use the dates $\{3^{rd}, 6^{th}, 9^{th}, \ldots, 30^{th}\}$ of each month to be the referent days. However, the advantage to the symmetric bi-directional referent is that we can avoid the "confounding" effects much better [13]. Hence, we are using the method of symmetric bi-directional referent selection *restricted to be within the same month and year* in our analysis. More specifically, we are using not only 7 days before and after the hospitalization as our reference days, but also 14, 21, and 28 days before and after the hospitalization, so that the referent days occur on the same day of the week.

This specific symmetric bi-directional referent selection is matching on the day of week within strata defined by month and year as in many epidemiological papers ([14] and [15]). This makes sense for El Paso, because the weather varies greatly from month to month. May is comfortable and June is exceedingly hot and dry; July is rainy; August is the monsoon season. We restrict the referent window so that the reference days are within the same month as the hospitalization date. For example, Harry was admitted on 05/10/2003 in the hospital because of asthma attack. Then the reference days for Harry will be 05/03/2003, 05/17/2003, 05/24/2003 and 05/31/2003. By doing this, we can then compare the exposures between "index time" and "referent times" with Harry serving as his own controls on those reference days.

Denote

$$
\begin{aligned}
Y &= \begin{cases} 1 & \text{hospital admission (case)} \\ 0 & \text{no hospital admission (control)} \end{cases} \\
X &= \begin{cases} 1 & \text{exposed (dust storm)} \\ 0 & \text{not exposed (no dust storm)} \end{cases}
\end{aligned} \tag{2.1}
$$

Particularly, we want to estimate the relative risk (approximated by odds ratio) of asthma hospitalization ($Y = 1$) given that a dust storm day occurred:

$$
\text{relative risk} = \frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)} \tag{2.2}
$$

We will utilize the conditional logistic regression model to estimate the relative risk by using the odds ratio. In our analysis, the event of interest rarely happens, that is,

$$
\frac{P(Y = 0|x = 0)}{P(Y = 0|x = 1)} \approx 1
$$

Thus, the relative risk (RR) can be estimated by odds ratio (OR):

$$\begin{aligned}
\text{RR} &= \frac{P(Y=1|x=1)}{P(Y=1|x=0)} \\
&\approx \frac{P(Y=1|x=1)}{P(Y=1|x=0)} \times \frac{P(Y=0|x=0)}{P(Y=0|x=1)} \\
&= \left[\frac{P(Y=1|x=1)}{P(Y=0|x=1)}\right] / \left[\frac{P(Y=1|x=0)}{P(Y=0|x=0)}\right] \\
&= OR
\end{aligned}$$

Details on the method will be introduced next.

Chapter 3

Statistical Methods

The derivation of the conditional likelihood function is broken up into three steps. I will introduce them in the following subsections.

3.1 Logistic Regression Applied to the Matched Prospective Study

The case-crossover analysis uses a retrospective design to look into the past. But first, let's review how the logistic regression works for the prospective study. Suppose we have an exposure series (including dust storm status), $\boldsymbol{x}$. In a *prospective* design, we know the distribution of y (e.g. asthma hospitalizations) given $\boldsymbol{x}$. Let x_{ik} denote the dust storm status for the i^{th} observation, where $k = 1, 2$.

In a *matched prospective design*, suppose we have a sample of n i.i.d. observations of the pairs $(Y_{11}, Y_{12}), (Y_{21}, Y_{22}), \ldots, (Y_{m1}, Y_{m2})$, where

$$Y_{11}, Y_{21}, \ldots, Y_{m1} \text{ are under exposure } (x_{i1} = 1)$$

and

$$Y_{12}, Y_{22}, \ldots, Y_{m2} \text{ are not under exposure } (x_{i2} = 0)$$

$$i = 1, 2, \ldots, m.$$

Let the quantity $\pi(x) = E(Y|x)$ represent the conditional mean of Y given x when the logistic distribution is used. The specific form of the logistic regression model we will use is as follows:

$$\text{exposed:} \quad \pi(x_{i1}) = P(Y_{i1} = 1) = \frac{e^{\alpha+\beta}}{1+e^{\alpha+\beta}}$$

$$\text{not exposed:} \quad \pi(x_{i2}) = P(Y_{i2} = 1) = \frac{e^{\alpha}}{1+e^{\alpha}}$$

or

$$\text{logit}[P(Y_{i1} = 1)] = \alpha + \beta, \qquad \text{logit}[P(Y_{i2} = 1)] = \alpha.$$

Equivalently,

$$\text{logit}[P(Y_{ik} = 1)] = \alpha + \beta x_{ik} \qquad i = 1, 2, \ldots, m \quad k = 1, 2. \tag{3.1}$$

The general method of estimation is *maximum likelihood*. The method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method, we must first construct a function, called the likelihood function. In our case, for those subjects where $Y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$, and for those where $Y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$. Let $\boldsymbol{\beta} = (\alpha, \beta)$. The likelihood function can be expressed as:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{m} \prod_{k=1}^{2} \pi(x_{ik})^{y_{ik}} [1 - \pi(x_{ik})]^{1-y_{ik}}.$$

However, it is easier mathematically to work with the log likelihood function:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{m} \sum_{k=1}^{2} \{y_{ik} \ln[\pi(x_{ik})] + (1 - y_{ik}) \ln[1 - \pi(x_{ik})]\}. \tag{3.2}$$

Then we maximize the log-likelihood to get the estimates by solving

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0.$$

For logistic regression the expressions in the above equation are nonlinear in $\boldsymbol{\beta}$, and thus require special method for their solutions.

3.2 Logistic Regression Applied to a Retrospective Study

The case-crossover study seeks to identify possible causes of the disease by finding out how the situations differ before and after the event happened. That is, because disease does not occur randomly (it occurs rarely), the subject must have been exposed to some factor, either voluntarily (e.g., through diet, exercise, or smoking) or involuntarily (through such factors as cosmic radiation, air pollution, or genetic constitution), that contributed to the causation of their disease [8].

Let Y denote the event that a subject has a certain disease (case). Suppose we have a sample of size n_1 cases ($Y = 1$) and n_0 controls ($Y = 0$). Let S denote the event that a subject is sampled, that is,

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

Suppose the prospective logistic regression model holds, with $\pi(\boldsymbol{x}_j) = P(Y_j = 1|\boldsymbol{x}_j)$ denoting the probability of disease. And let $\pi^*(\boldsymbol{x}_j) = P(Y_j = 1|\boldsymbol{x}_j, S_j = 1)$. However, the retrospective probability that we want is $P(\boldsymbol{x}|Y_j = 1, S_j = 1)$. i.e., we want to know the probability of the "causally" exposure given selected cases.

It will be shown that $P(\boldsymbol{x}_j|Y_j = 1, S_j = 1)$ is proportional to a logistic regression model $\pi^*(\boldsymbol{x}_j)$ with *the same effect parameter β as with the prospective study $\pi(\boldsymbol{x}_j)$, but with a new intercept α^**. This is an important first step in the theory that allows us to use prospective logistic regression on the retrospective study.

Now we have n_1 cases and n_0 controls in a sample of size n. The full likelihood function under the retrospective case-control design is

$$\prod_{i=1}^{n_1} P(\boldsymbol{x}_i|Y_i = 1, S_i = 1) \prod_{i=1}^{n_0} P(\boldsymbol{x}_i|Y_i = 0, S_i = 1). \tag{3.3}$$

To show that $P(\boldsymbol{x}_j|Y_j = 1, S_j = 1)$ is proportional to $\pi^*(\boldsymbol{x}_j)$ with the same effect parameter β but with a new intercept α^*, two applications of Bayes theorem are needed.

The first application of Bayes theorem:

$$\begin{aligned} P(A|B \cap C) &= \frac{P(A \cap B \cap C)}{P(B \cap C)} \\ &= \frac{\left[\frac{P(A \cap B \cap C)}{P(A \cap C)}\right]\left[\frac{P(A \cap C)}{P(C)}\right]}{\frac{P(B \cap C)}{P(C)}} \\ &= \frac{P(B|A \cap C)P(A|C)}{P(B|C)}. \end{aligned}$$

Let $A = \{\boldsymbol{x}_j\}, B = \{Y_j = 1\}$, and $C = \{S_j = 1\}$ be three events. Hence, the first application of Bayes theorem yields

$$\begin{aligned} P(\boldsymbol{x}_j|Y_j = 1, S_j = 1) &= \frac{P(Y_j = 1|\boldsymbol{x}_j, S_j = 1)P(\boldsymbol{x}_j|S_j = 1)}{P(Y_j = 1|S_j = 1)} \\ &= \frac{\pi^*(\boldsymbol{x}_j)P(\boldsymbol{x}_j|S_j = 1)}{P(Y_j = 1|S_j = 1)}. \end{aligned} \tag{3.4}$$

Assume that sample selection S is independent of the exposure series $\mathbf{x}$. Then, (3.4) can be written as

$$\begin{aligned} P(\boldsymbol{x}_j|Y_j = 1, S_j = 1) &= \frac{\pi^*(\boldsymbol{x}_j)P(\boldsymbol{x}_j)}{P(Y_j = 1|S_j = 1)} \\ &= \pi^*(\boldsymbol{x}_j)C_1 \end{aligned}$$

and

$$\begin{aligned} P(\boldsymbol{x}_j|Y_j = 0, S_j = 1) &= \frac{[1 - \pi^*(\boldsymbol{x}_j)]P(\boldsymbol{x}_j)}{P(Y_j = 0|S_j = 1)} \\ &= [1 - \pi^*(\boldsymbol{x}_j)]C_0 \end{aligned}$$

C_1 and C_0 can be viewed as two constants, because they do not include the parameter β. Thus, the two parts of the full likelihood function (3.3) are all proportional to $\pi^*(\boldsymbol{x}_j)$ and $1 - \pi^*(\boldsymbol{x}_j)$.

The second application of Bayes theorem can derive the relationship between $\pi^*(\boldsymbol{x}_j)$ and $\pi(\boldsymbol{x}_j)$, i.e., they share the same effect parameter β. According to the first application of Bayes theorem, we have

$$\begin{aligned} P(A'|B' \cap C') &= \frac{P(B'|A' \cap C')P(A'|C')}{P(B'|C')} \\ &= \frac{P(B'|A' \cap C')P(A'|C')}{P(A'^c \cap B'|C') + P(A' \cap B'|C')} \end{aligned}$$

Since

$$\begin{aligned} P(A'^c \cap B'|C') &= \frac{P(A'^c \cap B' \cap C')}{P(C')} \\ &= \frac{P(A'^c \cap C')}{P(C')} \times \frac{P(A'^c \cap B' \cap C')}{P(A'^c \cap C')} \\ &= P(A'^c|C')P(B'|A'^c \cap C') \end{aligned}$$

then the second application of Bayes theorem can be derived as

$$\begin{aligned} P(A'|B' \cap C') &= \frac{P(B'|A' \cap C')P(A'|C')}{P(A'^c \cap B'|C') + P(A' \cap B'|C')} \\ &= \frac{P(A'|C')P(B'|A' \cap C')}{P(A'^c|C')P(B'|A'^c \cap C') + P(A'|C')P(B'|A' \cap C')}. \end{aligned}$$

Now let $A' = \{Y_j = 1\}, B' = \{S_j = 1\}$, and $C' = \{\boldsymbol{x}_j\}$. Denote (ρ_1, ρ_0) as the probabilities of selection for the j^{th} subject.

$$\rho_1 = P(S_j = 1|Y_j = 1, \boldsymbol{x}_j) = P(S_j = 1|Y_j = 1)$$

and

$$\rho_0 = P(S_j = 1|Y_j = 0, \boldsymbol{x}_j) = P(S_j = 1|Y_j = 0).$$

Thus, the second application of Bayes theorem yields

$$\begin{aligned}
& \pi^*(\boldsymbol{x}_j) = P(Y_j = 1|\boldsymbol{x}_j, S_j = 1) \\
= & \frac{P(Y_j = 1|\boldsymbol{x}_j)P(S_j = 1|Y_j = 1)}{P(Y_j = 0|\boldsymbol{x}_j)P(S_j = 1|Y_j = 0) + P(Y_j = 1|\boldsymbol{x}_j)P(S_j = 1|Y_j = 1)} \\
= & \frac{\pi(\boldsymbol{x}_j)\rho_1}{[1 - \pi(\boldsymbol{x}_j)]\rho_0 + \pi(\boldsymbol{x}_j)\rho_1} \\
= & \left[\frac{\rho_1}{\rho_0}\frac{\pi(\boldsymbol{x}_j)}{1 - \pi(\boldsymbol{x}_j)}\right] / \left[1 + \frac{\rho_1}{\rho_0}\frac{\pi(\boldsymbol{x}_j)}{1 - \pi(\boldsymbol{x}_j)}\right] \\
= & \frac{e^{\ln(\rho_1/\rho_0)+\alpha+\beta x}}{1 + e^{\ln(\rho_1/\rho_0)+\alpha+\beta x}} \\
= & \frac{e^{\alpha^*+\beta x}}{1 + e^{\alpha^*+\beta x}} \qquad (3.5)
\end{aligned}$$

Here $\alpha^* = \ln(\rho_1/\rho_0) + \alpha$.

To sum up, The likelihood function for the retrospective case-control study (3.3) becomes:

$$l(\alpha^*, \beta) = \prod_{i=1}^{n} \pi^*(x_i)^{Y_i}[1 - \pi^*(x_i)]^{1-Y_i}.$$

So the likelihood for analysis of the retrospective study looks like the likelihood for the prospective study with $\pi^*(\boldsymbol{x})$ in place of $\pi(\boldsymbol{x})$. Now that we know what the likelihood looks like, let's take some time to carefully write out the model for the case-crossover study.

3.3 Conditional Likelihood Applied to the Case-Crossover Study

The case-crossover study is a special case of the case-control retrospective study, except that there are extra terms α_i for each stratum in place of α.

Instead of using the model $\text{logit}[P(Y_{ik}=1)] = \alpha + \beta x_{ik}$, we are using the model

$$\text{logit}[P(Y_{ik}=1)] = \alpha_i + \beta x_{ik}, \qquad k = 1, 2. \tag{3.6}$$

Ordinary ML (maximum likelihood) estimators are consistent when the sample size is large compared with the number of parameters. Suppose we have n observations in total, then there are $n+1$ parameters (α_i and β) in the model (3.6). When there are many parameters relative to the sample size, instead of using ordinary ML, we are using the conditional ML to remedy the problem [16].

Conditional likelihoods are used to get rid of the nuisance stratum parameters α_i in the estimation problem. To do this, we need to find a sufficient statistic for α_i. By definition of sufficiency, the likelihood function conditional on the sufficient statistics for α_i will not depend on α_i, but will only depend on β. The sufficient statistic for α_i is the stratum total and the total number of cases observed [17].

Suppose we have an exposure series (including dust storm status), $\mathbf{x_{it}}$, defined at times $t = 1, 2, \ldots, T$ common to all $i = 1, 2, \ldots, n$ subjects (or individuals). Denote the index time for subject i by t_i, and let W_{t_i} represent the *referent window* for subject i. The referent window includes the index time and all referent times. Then for subject i, x_{ij} represents the exposures when time j is in the referent window W_{t_i}.

To derive the estimator of the parameter of interest (β), the conditional likelihood is

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

where $l_i(\beta)$ is the conditional likelihood for the i^{th} stratum. Suppose each stratum has 1 case and m controls. According to the concept of combination, there are $\binom{m+1}{1}$ ways of assigning the status of CASE to one of the $m+1$ subjects in the stratum. Let's call these combinations $P_1, P_2, \ldots, P_{m+1}$. For instance, if $t_i = 1$, then we can denote the case and controls in the i^{th} stratum as:

$$P_1\text{: } (x_{i1}, Y=1) \text{ and } \{(x_{ij}, Y=0)\}_{j \in W_{t_i}, j \neq 1}$$

where x_{i1} now corresponds to the exposure when a case occurred at index time $t_i = 1$. The conditional likelihood for the i^{th} stratum is

$$\begin{aligned} l_i(\beta) &= P(P_1|data) \\ &= \frac{P(data|P_1)P(P_1)}{\Sigma_{j\in W_{t_i}} P(data|P_j)P(P_j)} \\ &= \frac{P(data|P_1)}{\Sigma_{j\in W_{t_i}} P(data|P_j)}. \end{aligned}$$

Given that we have 1 case and m independent controls for each stratum, the conditional likelihood for the i^{th} stratum is:

$$l_i(\beta) = \frac{P(x_{i1}|Y_j = 1)\Pi_{j\in W_{t_i}, j\neq 1} P(x_{ij}|Y_j = 0)}{\Sigma_{j_0\in W_{t_i}} \left[P(x_{ij_0}|Y_{j_0} = 1)\Pi_{j\in W_{t_i}, j\neq j_0} P(x_{ij}|Y_j = 0)\right]}$$

$$= \frac{\pi^*(x_{i1})\Pi_{j\in W_{t_i}, j\neq 1} \left[1 - \pi^*(x_{ij})\right]}{\Sigma_{j_0\in W_{t_i}} \left[\pi^*(x_{ij_0})\Pi_{j\in W_{t_i}, j\neq j_0}[1 - \pi^*(x_{ij})]\right]}$$

Upon substituting previous expression for $\pi^*(\boldsymbol{x}_j)$ in (3.5), we have

$$\begin{aligned} l_i(\beta) &= \frac{\frac{e^{\alpha_i^* + \beta x_{i1}}}{1+e^{\alpha_i^* + \beta x_{i1}}} \Pi_{j\in W_{t_i}, j\neq 1} \frac{1}{1+e^{\alpha_i^* + \beta x_{i1}}}}{\Sigma_{j_0\in W_{t_i}} \left[\frac{e^{\alpha_i^* + \beta x_{ij_0}}}{1+e^{\alpha_i^* + \beta x_{ij_0}}} \Pi_{j\in W_{t_i}, j\neq j_0} \frac{1}{1+e^{\alpha_i^* + \beta x_{ij_0}}}\right]} \\ &= \frac{e^{\alpha_i^* + \beta x_{i1}}}{\Sigma_{j_0\in W_{t_i}} \left(e^{\alpha_i^* + \beta x_{ij_0}}\right)} \\ &= \frac{e^{\beta x_{i1}}}{\Sigma_{j_0\in W_{t_i}} e^{\beta x_{ij_0}}}. \end{aligned} \tag{3.7}$$

Thus, the full conditional likelihood for the case-crossover study is:

$$l(\beta) = \Pi_{i=1}^n \left(\frac{e^{\beta x_{i1}}}{\Sigma_{j\in W_{t_i}} e^{\beta x_{ij}}}\right).$$

Take the derivative of the log likelihood $\ln l(\beta)$, with respect to β, to obtain the estimating equation of β [13]:

$$\frac{d \ln l(\beta)}{d\beta} = \sum_{i=1}^n x_{i1} - \sum_{i=1}^n \sum_{j\in W_{t_i}} x_{ij} \frac{e^{x_{ij}\beta}}{\sum_{j_0\in W_{t_i}} e^{x_{ij_0}\beta}}. \tag{3.8}$$

For continuous exposures, $e^{\triangle x\beta}$ approximates the relative risk of an event associated with a short-term increase $\triangle x$ in exposure. For dust storm ($x = 1$), e^{β}, the odds ratio (i.e., hazard ratio), approximates the relative risk of hospitalization for dust storms. The procedure PROC PHREG in SAS will maximize this likelihood and solve for $e^{\hat{\beta}}$.

Chapter 4

Data Description

The effect of dust storms on respiratory health in El Paso county is our main application in this thesis. Three sources of data were utilized in this study: hospitalization data, weather data and pollution data.

4.1 Hospitalizations Data

The hospitalization data are collected by the Texas Health Care Information Council (THCIC) in Austin, Texas. We are utilizing the data for the years from 2000 to 2005. Hospitalization counts are initially collected by hospitals and then collected by the states. As the name implies, the data base identifies hospitalizations of people with very severe cases of common respiratory diseases. We have four diseases in our data set: asthma, bronchitis, sinusitis, and upper respiratory infections. The hospitalization data do not include visits to the emergency room or to a primary care provider for these diseases, so it cannot be used for calculation of respiratory prevalence rates in the general population. Patients in our dataset are those who were hospitalized (spent at least one night in the hospital) in El Paso with one of those four acute respiratory diseases as primary diagnoses between January 1 2000 and December 31 2005.

Variables used include patient's age, sex, hospitalization date, payer and diagnosis. "Payer" is referring to the insurance status for patients. There are five types of payers for the patients (Table 4.1). The "diagnosis" are simply the ICD-9 codes indicating hospitalizations for respiratory complaints. For specificity, ICD-code 493.X indicates asthma, ICD-code 466.X indicates bronchitis, ICD-code 461.X indicates sinusitis, ICD-code 465.X

indicates upper respiratory infections as primary diagnoses for the patients. Table(4.2) lists sample records for a few hospitalizations; each row corresponds to one such admission. We have 12,554 hospitalizations in our dataset within the six years, of which 27 have missing data.

Table 4.1: Payer Types in Hospitalization Dataset

Payer	Insurance Status
1	Private Insurance
2	Medicare
3	Medicaid
4	No Insurance
5	Other Public

Table 4.2: Hospitalization Dataset Records

Record_ID	Sex	Age	Admission Date	Payer	Diagnosis
1	F	0	20000101	3	466.11
2	F	23	20000101	4	493.90
3	M	0	20000101	3	466.19
4	M	0	20000101	1	466.19
5	M	1	20000102	3	493.90
6	F	1	20000102	3	466.11
7	F	2	20000102	3	466.11
8	F	0	20000102	3	493.90
9	M	70	20000102	1	465.90
10	F	0	20000103	3	466.11
11	M	0	20000103	3	466.11
...	...	...	...	...	...

4.2 Weather-Pollution Data

The weather and pollution data is collected at El Paso International Airport (ELP) by the National Weather Service (NWS). Several monitors (CAMS) throughout El Paso compile a continuous record of data 24 hours per day, every day of the year. Usually, we take the 24-hour average for each monitor, and then take the average over all monitors. This is called "average, average", abbreviated as "AvAv". If instead we take the 24-hour maximum for each monitor and then take the average over all the monitor maxima, we call it "maximum, average", abbreviated as "MxAv". The available variables are storm, hourly PM2.5, temperature, dew point, lowwind, NO_2 and Ozone.

The dust storm data were defined by the National Weather Service's careful analysis

of hourly weather observer's records in El Paso, Texas. The raw data were coded by three types: "0" (no dust storm), "C" (Convective/Haboob dust storm), and "D" (dust storm). The Arabic word "haboob" means "monsoon". It is dust caused by convection in the atmosphere (a thunderstorm). However, we are ignoring the haboob/convective dust storms in our analysis, and consider them as "0"s. According to Dr. Tom Gill, Department of Geological Sciences at UTEP, the reason why we are ignoring the haboob/convective events is two fold: First, the convective dust storms are much smaller in spatial and temporal scale/effect that the more widespread (synoptic) dust storms. A cloud of dust caused by the downdrafts of a thunderstorm will generally cover only a small area at a given time, and only last for a short time. The dust caused by a convective (haboob) event will generally pass through a given area very quickly - within minutes, or less than an hour - and cover only perhaps one neighborhood or part of town at a time. So it is very transient. A "D" dust event will cover the entire cities of El Paso, Texas and Juarez, Mexico at the same time for many hours at a time. Thus the expectation is it will give a prolonged dose of dust to the entire population. Second, the study is based on the official weather records from the airport, which will not be impacted by every haboob that hits the city. In other words, there could be a convective dust storm out on the east side or west side of town that never hits the airport, and thus it would not be reflected in that data. To sum up, there will be only two categories of dust storms in our analysis: "not a dust storm" (0 and C) and "dust storm" (D). In total, 142 dust storm days occurred in El Paso region between 2000 to 2005. The corresponding variable in the model is called "storm":

$$\text{storm} = \begin{cases} 1 & \text{dust storm day} \\ 0 & \text{not a dust storm day} \end{cases} \tag{4.1}$$

The measurement "AvAv" is applied to the hourly PM2.5, temperature and dew point. "AvAv" is abbreviated from "average, average". That means we take the 24-hourly average for each monitor and then take the average over all the monitor averages. The corresponding names in the models to be fit are "PM2_5", "Temp", and "Dew_Point" in the model.

However, when it comes to NO_2 and Ozone, there is an issue about choosing "AvAv"

or "MxAv". Similar as "AvAv", "MxAv" means that we take the 24-hourly maximum for each monitor and then take the average over all the monitor maxima. Experience said that the maximum level in a 24-hour period across CAMS is the best measure for NO_2 and Ozone[1]. To decide this in a scientific way, the correlation coefficient between NO_2_AvAv and Ozone_AvAv, also NO_2_MxAv and Ozone_MxAv was computed for the overall data, and between NO_2_Av and Ozone_Av, NO_2_Mx and Ozone_Mx with data from CAM_12. See Table(4.3) and Table(4.4). Notice that average across monitors still preserves the sign and magnitude of the correlations at CAM_12. The "AvAv"s, have much higher correlation coefficients than those of the "MxAv"s. So we would prefer the "MxAv" to avoid the colinearity in the model. Also, according to chemistry knowledge, NO_2 and Ozone are supposed to be positively correlated, because they both are oxidants. Thus, choosing NO_2_MxAv and Ozone_MxAv to be our covariates is reasonable.

Table 4.3: Pearson Coefficients between NO_2_AvAv and Ozone_AvAv

OVERALL	NO_2_AvAv	Ozone_AvAv	CAM_12	NO_2_Av	Ozone_Av
NO_2_AvAv	1.00000	-0.61992	NO_2_Av	1.00000	-0.64800
Ozone_AvAv	-0.61992	1.00000	Ozone_Av	-0.64800	1.00000

Table 4.4: Pearson Coefficients between NO_2_MxAv and Ozone_MxAv

OVERALL	NO_2_MxAv	Ozone_MxAv	CAM_12	NO_2_Mx	Ozone_Mx
NO_2_MxAv	1.00000	0.06333	NO_2_Mx	1.00000	0.06100
Ozone_MxAv	0.06333	1.00000	Ozone_Mx	0.06100	1.00000

In the study, the ambient condition of low wind is another weather variable that is important to consider. Low wind conditions are believed to be associated with urban air

[1]Personal communication has been made with Dr. Carol Atkinson-Palombo, University of Connecticut.

pollution, trapped by atmospheric inversions, whereas dust storm days are associated with coarse particulate matter such as entrained sand. The low wind condition is entered as an indicator variable in the model. The threshold to be considered as "lowwind" is 4.5mph, the 10^{th} percentile of daily average from the overall records for wind within the six years. In total, 221 low wind days occurred in El Paso region between 2000 to 2005.

$$\text{lowwind} = \begin{cases} 1 & \text{AvWind} \leq 4.5 \text{ mph (10th percentile)} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

Table(4.5) gives the descriptive statistics for the hospitalization counts associated with lowwind and storm. Figure(4.1) is the boxplot for the hospitalization counts on days with low wind, or dust storm weather conditions. It seems that high hospital admissions occur on lowwind and storm days, when examining quartiles. Further model results will be given in the next chapter.

Table 4.5: Descriptive Statistics for Hospitalization Counts under Different Weather Conditions

Group	Mean	StDev	Minimum	Median	Maximum	Number of Days
Storm= 0 & Lowwind= 1	7.550	5.066	1.000	7.000	29.000	221
Storm= 0 & Lowwind= 0	5.768	4.842	1.000	4.000	33.000	1826
Storm= 1	7.599	5.269	1.000	6.000	26.000	145

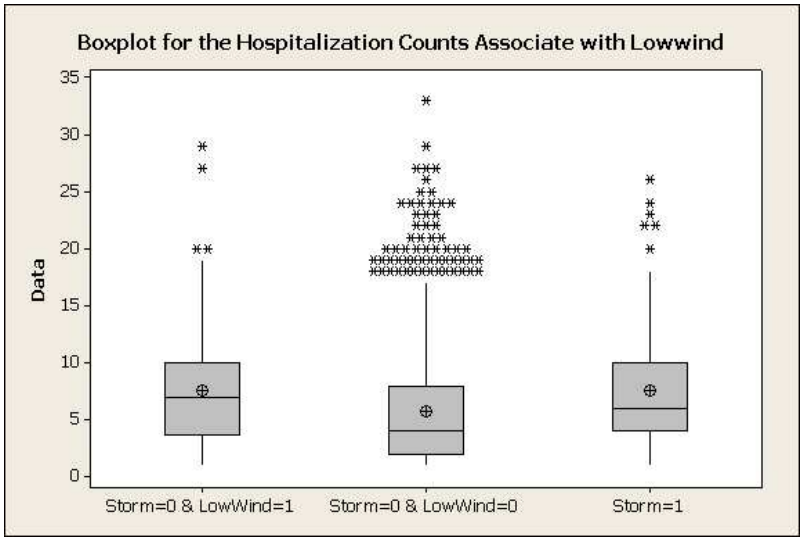

Figure 4.1: Boxplot for the Hospitalization Counts on Days with Lowwind and Dust Storm

In summary, the available variables that we are utilizing as the weather-pollution variables are NO_2_MxAv, Ozone_MxAv, PM2_5, lowwind, storm, Dew_Point and Temperature. The summary statistics and the correlation matrix of the weather-pollution variables are given in the Table(4.6) and Table(4.7). Time-series plots for these weather-pollution variables (except for lowwind and storm) are listed from Figure (4.2) to Figure (4.4).

Table 4.6: Distributions of Weather and Pollution Variables

				Percentile points						
Variable	Units	N	Minimum	10%	25%	50%	75%	90%	Maximum	Mean
NO_2_MxAv	ppb	2192	7.00	19.00	25.00	33.00	41.00	48.00	99.00	33.602
Ozone_MxAv	ppb	2192	6.00	33.00	39.00	48.00	59.00	70.00	109.00	50.162
PM2_5	$\mu g/m^3$	2192	1.34	5.57	7.48	10.63	14.91	21.02	119.07	12.428
lowwind	mph	2192	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.101
Dew_Point	F	2192	-6.00	19.00	25.00	36.00	49.00	56.00	68.00	36.704
Temp	F	2192	24.00	44.00	52.00	67.00	80.00	85.00	93.00	65.587

Table 4.7: Pearson Correlation Coefficients among Weather and Pollution Variables

Variable	NO_2_MxAv	Ozone_MxAv	PM2_5_AvAv	Dew_Point	Av_Temp
NO_2_MxAv	1.00000	0.06333	0.33440	-0.41852	-0.25164
Ozone_MxAv		1.00000	0.06930	0.38613	0.65968
PM2_5			1.00000	-0.25191	0.00243
Dew_Point				1.00000	0.69362
Temp					1.00000

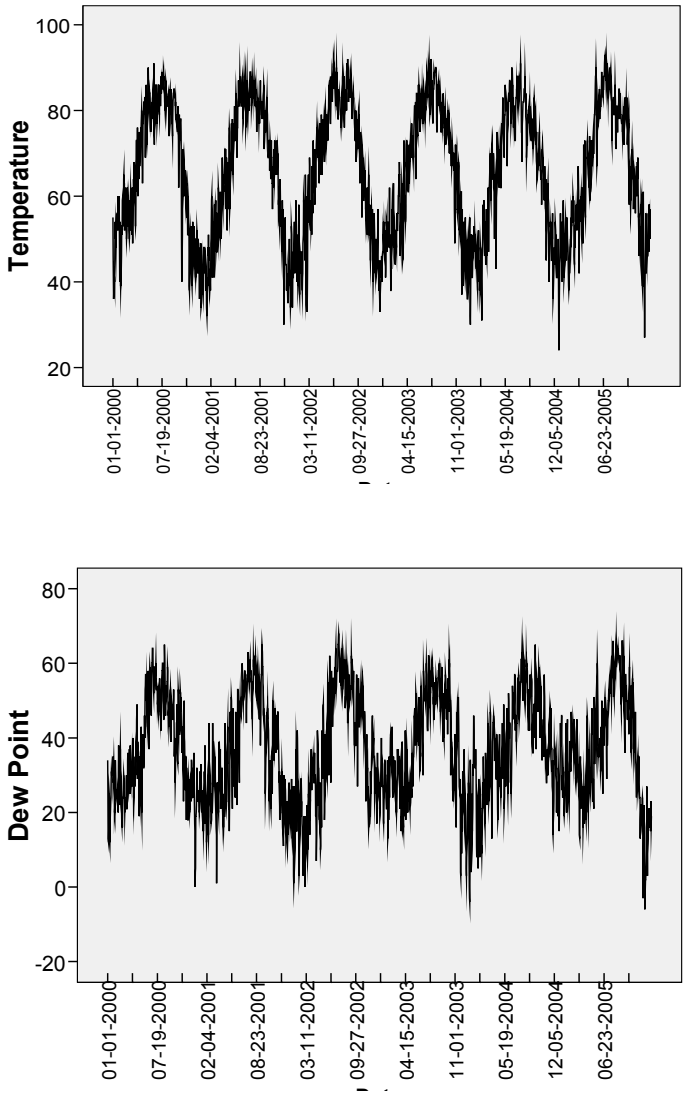

Figure 4.2: Time-series Plot for Temperature and Dew Point in 2000-2005

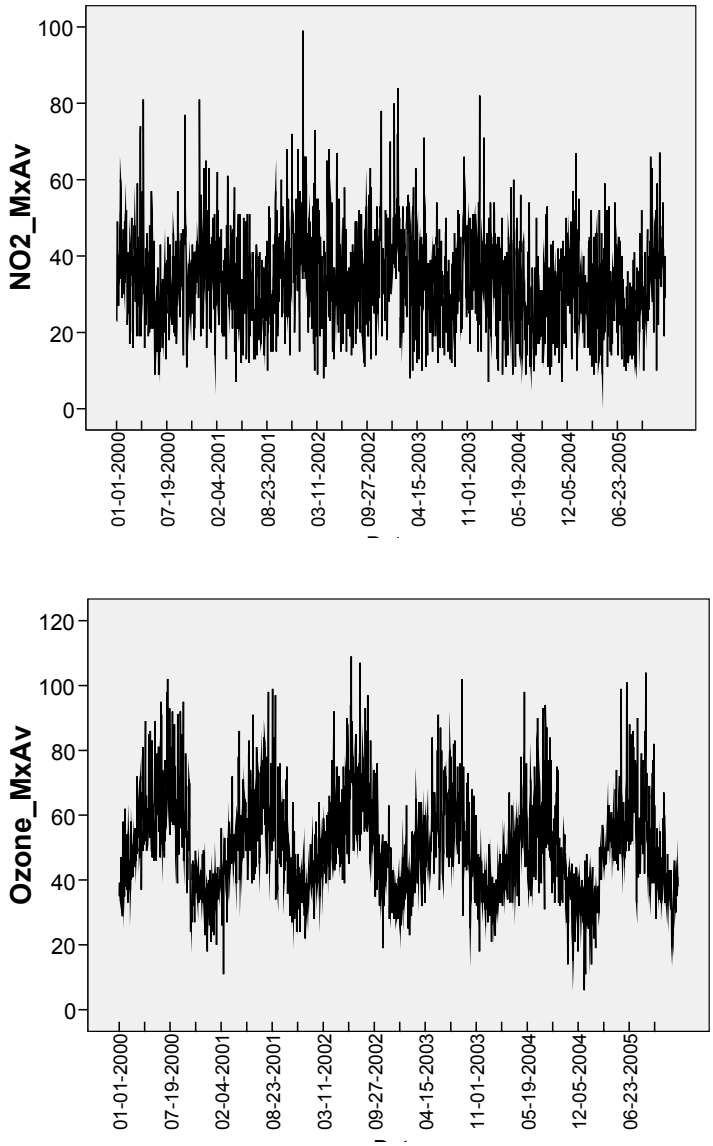

Figure 4.3: Time-series Plot for NO_2_MxAv and Ozone_MxAv in 2000-2005

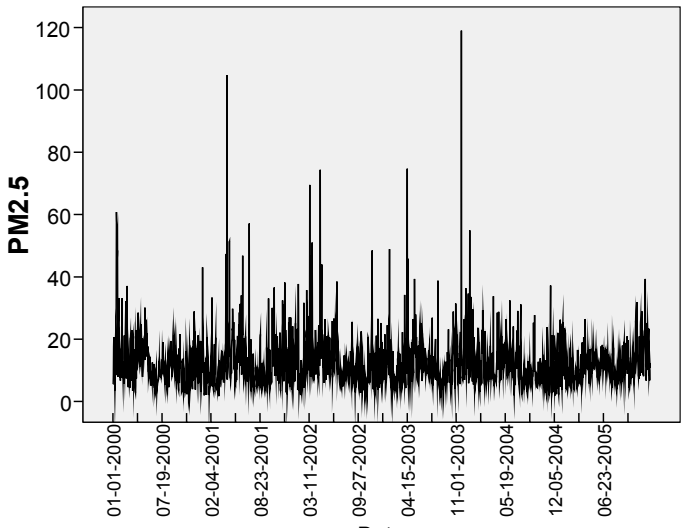

Figure 4.4: Time-series Plot for PM2_5 in 2000-2005

Chapter 5

Modeling Results

This study leads us to pose three primary related research questions (1-3) and three exploratory subgroup analyses (4-6):

1. Does the probability of hospitalization depend on dust storms, without adjustment for covariates?

2. What is the effect of dust storm after adjustment for weather and pollution variables?

3. Does dust storm have the same effect on different age groups of the patients?

4. Which age group appears to be more prone to hospitalization after dust storms?

5. Do insurance status and dust storm interact in a model for probability of hospitalization?

6. Do gender and dust storm interact in a model for probability of hospitalization?

Backward variable selection is utilized in this analysis. We are using the threshold level of 0.2, that is, the variable with the largest p-value from among those with a p-value bigger than 0.2 will be excluded from the model at each selection step. The final variables will be the most significant ones with p-values less than 0.2, and certain variables forced into the model[1]. The *hazard ratio* (i.e., odds ratio for the conditional logistic regression model) provides an estimate of relative risk of events of interest.

[1]Forced variables mean these variables will not be excluded from the model no matter what their p-values are. They will always stay in the model.

Two sections follow to answer the questions listed above. It is, however, necessary that we explain the model building process before we start answering any of the questions. We are using the nesting strategy to build our models as in Staniswalis 2005 [18]. There are 3 steps in this process: Step 1, without forcing any variables, use backward selection to find the significant variables from the weather-pollution pool, and to construct the "Baseline" model. Here "weather-pollution pool" includes Temperature, Lowwind, Dew Point, PM2_5, NO_2_MxAv and Ozone_MxAv and all their three lags, but not "storm". Step 2, forcing the variables from the baseline model, apply backward variable selection to the storm variables: storm lags 0-3. Step 3, for dealing with the interaction effects, force the baseline model variables, and the storm variables with the smallest p-value in step 2 for construction of the interaction terms. The interaction terms are forced in the model.

5.1 Primary Research Analyses

Three subsections will be used to address each of the first three research questions.

5.1.1 The Effect of Dust Storms Alone

Results for the first research question (Does the probability of hospitalization depend on dust storms, without adjustment for covariates?) can be found in Table(5.1). Dust storm for the current day (lag 0) and at lags 1-3 were included in the model using backward variable selection. Dust storm (lag 0) is the only significant variable (p-value = 0.0196), with a hazard ratio (i.e., relative risk) of 1.103. Thus, on any day with dust storm (lag 0), people are 1.103 times more likely to be hospitalized than those on any day without dust storm. This corresponds to a 10% increase in expected number of hospitalizations on dust storm days among the El Paso population.

Table 5.1: Analysis of Maximum Likelihood Estimates for Storm Alone

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Storm	1	0.09803	0.04201	5.4440	0.0196	1.103

5.1.2 The Effect of Dust Storm Adjusted for Weather-Pollution Covariates

The second research question is: What is the effect of dust storm after adjustment for weather and pollution variables? The covariates for the adjustment include all the weather-pollution variables on the current day together with the three lags of each variable. The weather-pollution variables are Temperature, Lowwind, Dew Point, PM2_5, NO_2_MxAv and Ozone_MxAv. Of these, backward variable selection identified temperature at lags 1 and 3, dew point lag 3, and lowwind as significant predictors of hospitalization in the baseline model (Table 5.2).

Table 5.2: Analysis of Maximum Likelihood Estimates for Baseline Model Including Only Weather-Pollution Covariates

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lowwind	1	-0.05752	0.03536	2.6461	0.1038	0.944
Lag1temp	1	-0.01378	0.00801	2.9597	0.0854	0.986
Lag3temp	1	0.01431	0.00799	3.2045	0.0734	1.014
Lag3dpt	1	-0.00281	0.00142	3.9036	0.0482	0.997

After adjusting for the weather-pollution covariates, dust storm on the current day and at lags 1-3 were added to the baseline model. Table(5.3) displays the results of the backward variable selection, dust storm (lag 0) is significant (p-value = 0.03) at the 0.05

level of significance. The hazard ratio for dust storm is 1.095, meaning that on any day with dust storm (lag 0), people are 1.095 times more likely to be hospitalized than those on any day without dust storm. This corresponds to a 9.5% increase in expected number of hospitalizations on dust storm days among the El Paso population. This finding is similar to the effect of dust storm reported without the weather and pollution variables adjustment.

Table 5.3: Analysis of Maximum Likelihood Estimates for Storm Adjusted with Weather-Pollution Covariates

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lowwind	1	-0.04988	0.03545	1.9791	0.1595	0.951
Lag1temp	1	-0.01451	0.008	3.291	0.0697	0.986
Lag3temp	1	0.01447	0.00799	3.2846	0.0699	1.015
Lag3dpt	1	-0.00257	0.00142	3.2448	0.0717	0.997
Storm	1	0.09053	0.04273	4.4877	0.0341	1.095

Exploratory plots of PROC PHREG residuals versus the continuous covariates, lag1tmep, lag3temp and lag3dpt, did not indicate any lack-of-fit.

5.1.3 The Effect of Dust Storm Interacting with Age

After analyzing the significant predictors of probability of hospitalization, we want to compare the dust storm effects within different ages of the patients. We divide the age into three groups: children (0-17 years), adult (18-64 years), and elderly (65+ years). Table(5.4) gives the observation numbers of each group of age. Children will be used as the reference group, because it is the largest group. The model compares the probability of hospital admission for the adult and elderly with the child. Two dummy variables were created to

indicate the adult and the elderly groups.

$$y1 = 1 \quad \text{and} \quad y2 = 0, \quad \text{if the patient is adult;}$$

$$y1 = 0 \quad \text{and} \quad y2 = 1, \quad \text{if the patient is elderly.}$$

Letting these two dummy variables interact with the storm variable, we can analyze the effect of dust storm in different age groups by comparing with the reference group (children). So starting with the model in Table(5.3), and forcing the two interaction terms, Table(5.5) gives the answer to the third research question (Does dust storm have the same effect on different age groups of the patients?). Since the p-values of the two interaction terms are 0.4779 and 0.5152, the adult and elderly are not significantly different than children in their response to dust storms. Note that y1, y2 were not included as terms in the model because they are stratum covariates that could be canceled out from the conditional likelihood function (see equation (3.7) in chapter 3).

Table 5.4: Frequency Table for Hospitalizations With Respect to Different Age Groups

Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 (0-17 years)	8696	69.38	8696	69.38
2 (18-64 years)	2172	17.33	10868	86.72
3 (65+ years)	1665	13.28	12533	100.00

Frequency Missing = 21

Table 5.5: Analysis of Maximum Likelihood Estimates for Storm Interacting with Age

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lowwind	1	-0.05108	0.0355	2.0705	0.1502	0.95
Lag1temp	1	-0.01453	0.008	3.2964	0.0694	0.986
Lag3temp	1	0.01446	0.00799	3.2784	0.0702	1.015
Lag3dpt	1	-0.00267	0.00143	3.5144	0.0608	0.997
Storm	1	0.08916	0.04873	3.3473	0.0673	1.093
$y1$*storm	1	0.06967	0.09817	0.5036	0.4779	1.072
$y2$*storm	1	-0.06764	0.10395	0.4234	0.5152	0.935

adult: $y1$*storm

elderly: $y2$*storm

5.2 Exploratory Subgroup Analyses

In this section, we will explore some interesting subgroup analyses by using some interaction terms with dust storm. The main reason for studying interactions is because they may help to identify high risk groups. In certain subgroups, the hazard ratio (i.e., relative risk) associated with exposure might be higher than in the rest of the population. This can be implied by the corresponding interaction. This section includes three parts in order to answer the last three research questions which is stated at the beginning of the chapter.

5.2.1 The Most Sensitive Age Groups

From the last section, we do not see different effects of dust storms within the three age groups. But we still wonder which weather-pollution covariates and dust storm lags are the significant predictors for each of the groups. By using the backward variable selection with threshold 0.2, Table(5.6) - Table(5.11) present a separate analysis for determination

of the baseline model and dust storm lags for each of the three age groups.

According to the model selection process, we should (1) construct the baseline model first from the weather-pollution variables, (2) and then use backward variable selection on the dust storm variables. For the child group, Table(5.6) is the model result of step (1), and Table(5.7) is the result of step (2). In Table(5.6), lowwind, dew point lag 3 and ozone lag 1 are in the baseline model for child patients. We can find storm lag 1 is selected in Table(5.7) as the most significant storm variable for children. Although, 1-day lag of dust storm was not a significant predictor of hospitalization in children with a p-value of 0.1015 and a hazard ratio of 1.085.

Table 5.6: Child Group (0-17 years): Analysis of Maximum Likelihood Estimates for Baseline Model Including Only Weather-Pollution Covariates

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lowwind	1	-0.05907	0.04127	2.0484	0.1524	0.943
Lag3dpt	1	-0.00523	0.00163	10.2942	0.0013	0.995
Lag1oz	1	-0.00232	0.00163	2.0436	0.1528	0.998

Table 5.7: Child Group (0-17 years): Analysis of Maximum Likelihood Estimates for Storm Adjusted with Weather-Pollution Covariates

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lowwind	1	-0.05165	0.04143	1.5539	0.2126	0.95
Lag3dpt	1	-0.00512	0.00163	9.8382	0.0017	0.995
Lag1oz	1	-0.00212	0.00163	1.6867	0.194	0.998
Lag1storm	1	0.08153	0.04978	2.6819	0.1015	1.085

Next Table(5.8) and Table(5.9) give the modeling results for the adult group. In Table(5.8), lowwind (lag 0) is dropped from the model. Instead, the third lag of this variable is selected. Besides this, temperature, PM2.5 lag 0 and lag 3, NO2_MxAv, and the third lag of dew point are selected for the baseline model. At step (2), we have storm lag3 left in the model (Table 5.9), with a p-value of 0.0489, and a hazard ratio of 1.234. Hence, for adults, the third day after the dust storm is a significant predictor of respiratory health.

Table 5.8: Adult Group (18-64 years): Analysis of Maximum Likelihood Estimates for Baseline Model Including Only Weather-Pollution Covariates

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Temp	1	0.01001	0.0044	5.1686	0.023	1.01
PM2_5	1	0.00812	0.004	4.1256	0.0422	1.008
NO2_MxAv	1	-0.00887	0.00261	11.5265	0.0007	0.991
Lag3wind	1	0.19977	0.08332	5.749	0.0165	1.221
Lag3dpt	1	-0.0076	0.00345	4.8409	0.0278	0.992
Lag3pm	1	-0.00549	0.00404	1.8457	0.1743	0.995

Table 5.9: Adult Group (18-64 years): Analysis of Maximum Likelihood Estimates for Storm Adjusted with Weather-Pollution Covariates

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Temp	1	0.0104	0.00441	5.5608	0.0184	1.01
PM2_5	1	0.00832	0.004	4.3119	0.0378	1.008
NO2_MxAv	1	-0.00771	0.00268	8.3009	0.004	0.992
Lag3wind	1	0.21715	0.08379	6.7163	0.0096	1.243
Lag3dpt	1	-0.00734	0.00346	4.5017	0.0339	0.993
Lag3pm	1	-0.00799	0.00424	3.5486	0.0596	0.992
Lag3storm	1	0.21019	0.10672	3.879	0.0489	1.234

However, things are not the same when it comes to the elderly. The only weather-pollution covariates left in the baseline model are the first lag of temperature and the third lag of lowwind (Table 5.10). In the step of selecting dust storm variables, none of them is left in the model after backward variable selection (Table 5.11). But we will keep storm lag 2 as the predictor for elderly in the next analyses, since it has the smallest p-value among all the dust storm variables (p-value= 0.381).

Table 5.10: Elderly Group (65+ years): Analysis of Maximum Likelihood Estimates for Baseline Model Including Only Weather-Pollution Covariates

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lag1temp	1	0.00598	0.00455	1.7234	0.1893	1.006
Lag3wind	1	-0.17578	0.09662	3.3094	0.0689	0.839

Table 5.11: Elderly Group (65+ years): Analysis of Maximum Likelihood Estimates for Storm Adjusted with Weather-Pollution Covariates

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lag1temp	1	0.00678	0.00461	2.164	0.1413	1.007
Lag3wind	1	-0.17105	0.09662	3.1343	0.0767	0.843
Storm	1	0.07064	0.16966	0.1734	0.6771	1.073
Lag1storm	1	-0.02386	0.21756	0.012	0.9127	0.976
Lag2storm	1	-0.19264	0.21991	0.7673	0.381	0.825
Lag3storm	1	0.06391	0.1697	0.1418	0.7065	1.066

Note: No storm variable is significant.

To answer the fourth research question (Which age group appears to be more prone to hospitalization after dust storms?), we need to review the model results again. The elderly group has no significant storm variable at all. Besides that, the storm lag 1 was not a significant predictor of hospitalization for children, with a p-value of 0.1015 and a hazard ratio of 1.085. While with a smaller p-value of 0.0489 and a bigger hazard ratio of 1.234, the third day after the dust storm was significant for the adult. Thus, surprisingly, instead of the child and elderly, the adult group is the most prone to hospitalization after dust storms in our analysis.

5.2.2 The Effect of Dust Storm Interacting with Insurance Status

Based on the previous discussion, there is evidence that not all groups are impacted by dust storms to the same degree. Besides age, we are also interested in patients using different insurances. Generally speaking, insurance status is an indicator of socio-economic status. In this section, we will investigate whether certain insurance groups are impacted differently by dust storms.

There are 5 types of insurance status ("Payer") used in the hospitalization data set. Table (5.12) gives the frequency table for the counts of each group. Using insurance group "Private" as the reference, we define four dummy variables to indicate the other different insurance status:

$$z1 = 1, z2 = 0, z3 = 0, z4 = 0 \quad \text{if the patient is using Medicare;}$$

$$z1 = 0, z2 = 1, z3 = 0, z4 = 0 \quad \text{if the patient is using Medicaid;}$$

$$z1 = 0, z2 = 0, z3 = 1, z4 = 0 \quad \text{if the patient is using No Insurance;}$$

$$z1 = 0, z2 = 0, z3 = 0, z4 = 1 \quad \text{if the patient is using Other Public;}$$

Letting these four dummy variables interact with the storm variable, we can analyze the effect of dust storm in different insurance groups by comparing with the reference group (Private). Tables (5.13-5.15) give the answer of the fifth research question (Do insurance status and storm interact in a model for probability of hospitalization?).

Table 5.12: Frequency Table for Insurance Groups

Payer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Private	3513	28.37	3513	28.37
Medicare	1916	15.47	5429	43.85
Medicaid	6456	52.14	11885	95.99
No Insurance	259	2.09	12144	98.08
Other Public	238	1.92	12382	100.00

Frequency Missing = 172

Table 5.13: Model Results for Children (0-17 years): Storm Interacting with Insurance Status

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lowwind	1	-0.04824	0.04159	1.3455	0.2461	0.953
Lag3dpt	1	-0.005	0.00164	9.3361	0.0022	0.995
Lag1oz	1	-0.00203	0.00164	1.5447	0.2139	0.998
Lag1storm	1	0.07066	0.08824	0.6412	0.4233	1.073
z2*Lag1storm	1	0.01098	0.09851	0.0124	0.9112	1.011
z3*Lag1storm	1	0.08064	0.36135	0.0498	0.8234	1.084
z4*Lag1storm	1	0.25651	0.33062	0.6019	0.4378	1.292

Medicaid: z2*Lag1storm

No Insurance: z3*Lag1storm

Other Public: z4*Lag1storm

Reference Group: Private

The results for modeling storm interacting with insurance, just for children, can be found in Table(5.13). Note that we excluded the "Medicare" category for children, since Medicare basically is only for adult and elderly. In this table, we see the p-values corresponding to Medicaid, No Insurance, and Other Public are 0.9112, 0.8234, and 0.4378 respectively. So none of the interaction terms are significant for children. Furthermore, we can not find any significant interaction term either for adult (Table 5.14), nor for elderly (Table 5.15). Thus, the answer to the fifth research question should be: Insurance status and dust storm do not interact in a model for probability of hospitalization by age.

Table 5.14: Model Results for Adult (18-64 years): Storm Interacting with Insurance Status

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Temp	1	0.01144	0.00449	6.4983	0.0108	1.012
PM2_5	1	0.00822	0.00404	4.1343	0.042	1.008
NO2_MxAv	1	-0.00673	0.00271	6.1468	0.0132	0.993
Lag3wind	1	0.21739	0.08521	6.5094	0.0107	1.243
Lag3dpt	1	-0.00861	0.00352	5.9923	0.0144	0.991
Lag3pm	1	-0.00865	0.00429	4.0582	0.044	0.991
Lag3storm	1	0.16416	0.14549	1.2731	0.2592	1.178
z1*Lag3storm	1	0.41975	0.24463	2.9442	0.0862	1.522
z2*Lag3storm	1	-0.15844	0.23974	0.4368	0.5087	0.853
z3*Lag3storm	1	0.51162	0.35106	2.1239	0.145	1.668
z4*Lag3storm	1	0.2031	0.63957	0.1008	0.7508	1.225

Medicare: z1*Lag3storm

Medicaid: z2*Lag3storm

No Insurance: z3*Lag3storm

Other Public: z4*Lag3storm

Reference Group: Private

Table 5.15: Model Results for Elderly (65+ years): Storm Interacting with Insurance Status

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lag1temp	1	0.00783	0.0046	2.8988	0.0886	1.008
Lag3wind	1	-0.16226	0.09684	2.8075	0.0938	0.85
Lag2storm	1	0.16615	0.4615	0.1296	0.7188	1.181
z1*Lag2storm	1	-0.26929	0.47019	0.328	0.5668	0.764
z2*Lag2storm	1	-12.97821	274.99531	0.0022	0.9624	0

Medicare: z1*Lag2storm

Medicaid: z2*Lag2storm

Reference Group: Private

5.2.3 The Effect of Dust Storm Interacting with Gender

Besides insurance status, we are also interested in patients' gender. The sixth research question is to explore the interaction effects between dust storm and gender. Table(5.16) to Table(5.18) give the results for answering this research question.

According to the modeling process of step 3, we are matching the interaction terms to the storm variables that we selected for each of the groups in section 5.2.1. In Table(5.16), the p-value for the interaction term, female*Lag1storm, is 0.1028, and the hazard ratio is 1.15. Thus, female children are more likely to be admitted to the hospital than the male, although it is not a significant finding at the level of significance 0.05. Also, the p-values for the interaction terms of either adult (0.5238 in Table 5.17) or elderly (0.6603 in Table 5.18) are not significant. Thus, the interaction effect between dust storm and genders is not significant for child, adult or elderly.

Table 5.16: Model Results Just for Child: Gender Interacts Storm

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lowwind	1	-0.05177	0.04143	1.561	0.2115	0.95
Lag3dpt	1	-0.0051	0.00163	9.7674	0.0018	0.995
Lag1oz	1	-0.00212	0.00163	1.6957	0.1928	0.998
Lag1storm	1	0.02252	0.06197	0.1321	0.7162	1.023
Female*Lag1storm	1	0.13973	0.08564	2.6622	0.1028	1.15

Reference Group: Male

Table 5.17: Model Results Just for Adult: Gender Interacts Storm

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Temp	1	0.01036	0.00441	5.5104	0.0189	1.01
PM2_5	1	0.00836	0.004	4.3557	0.0369	1.008
NO2_MxAv	1	-0.00772	0.00268	8.3254	0.0039	0.992
Lag3wind	1	0.2172	0.08379	6.7192	0.0095	1.243
Lag3dpt	1	-0.00732	0.00346	4.4792	0.0343	0.993
Lag3pm	1	-0.008	0.00424	3.5563	0.0593	0.992
Lag3storm	1	0.31168	0.19083	2.6677	0.1024	1.366
Female*Lag3storm	1	-0.13334	0.20916	0.4064	0.5238	0.875

Reference Group: Male

Table 5.18: Model Results Just for Elderly: Gender Interacts Storm

Variable	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Lag1temp	1	0.00693	0.00457	2.3038	0.1291	1.007
Lag3wind	1	-0.17428	0.09648	3.2627	0.0709	0.84
Lag2storm	1	-0.05622	0.17742	0.1004	0.7513	0.945
Female*Lag2storm	1	-0.09228	0.20999	0.1931	0.6603	0.912

Reference Group: Male

5.3 Modeling Summary

By using the case-crossover study, this analysis reveals the effects of dust storm on hospitalization from common respiratory diseases (asthma, bronchitis, sinusitis and upper respiratory infections) within the range of El Paso county. The data suggests that the probability of hospitalization indeed depends on dust storms, whether adjusted for weather-pollution covariates or not. Together with the dust storm on the current day, we also find that lowwind, the first and third lags of temperature, and the third lag of dew point were chosen in the baseline model, but none are significant predictors of hospitalizations (Table 5.3). For different age groups, the child, adult and elderly are not significantly different in their response to dust storms. Although a subgroup exploratory analysis suggests that the adult group appears to be the most prone to hospitalization after dust storms (section 5.2.1).

When a dust storm occurs, we were originally concerned that the socio-economic conditions might have some effect on the hospitalization for patients. In our analysis, insurance status is considered as the proxy for patient's socio-economic condition, which is very common in many epidemiology reports. But it turns out that insurance status and dust storm do not interact in a model for probability of hospitalization by age (section 5.2.2). In the field of epidemiology, experts are also concerned that females might be more likely to be

admitted to the hospital than males. So we investigated the interaction effect between dust storm and gender by age in an exploratory analysis. But it turns out that the gender is not significant for the child, adult or elderly, although we have the highest relative risk (1.15) among female children.

Finally, a significant association between dust storm and the respiratory health was found in the region of El Paso, Texas. On the current day of dust storm, usually in spring and sometime in November, which suggests that outdoor activities should be restricted during dust storms. Strong dust storm might cause severely acute respiratory response, such as asthma attack in people who have not been diagnosed with asthma. This analysis uses a large data set (four monitoring sites, large hospital admission counts, and 6 years of data), this is a strength. Our findings contribute to the knowledge about the dust storm's respiratory health effects.

Chapter 6

Strengths and Limitations of Case-Crossover Design

In a case-crossover study, there are four advantages compared with studies based on daily counts and air pollution levels. First, confounding is reduced because cases serve as their own controls. Secondly, the case-crossover design allows for the use of routinely monitored air pollution information and at the same time makes it possible to study individuals rather than days as the unit of observation. Thirdly, it is also possible to study effect modification (i.e., to identify individuals susceptible to the effects of exposure of interest), based on the information on individual characteristics such as age, sex, health status, and socio-econometric levels. Lastly, restricted bi-directional selection of control periods allows individual adjustment for seasonal and secular trends, such as day of week within strata defined by month and year ([14] and [15]).

Above all, however, a disadvantage of this approach is that, compared with Poisson regression time-series analysis, it has approximately as low as 66% efficiency, as shown by Bateson and Schwartz [19]. The case-crossover study design is also limited within certain applications. It is only fit for estimating the effects of serious outcomes with an abrupt onset to merely short-term exposures, whereas other types of study are needed for studying effects of long-term exposures.

References

[1] Dockery D. An association between air pollution and mortality in six U.S. cities. *N. Engl. J. Med.*, **329**, 1753-1759, 1993

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

[3] Lu Y., Zeger SL. On the equivalence of case-crossover and time series methods in environmental epidemiology. *COBRA*, **49**, 803-821, 2006

[4] Comstock GW. Evaluating vaccination effectiveness and vaccine efficacy by means of case-control studies. *Epidemiol Rev.*, **16**, 77-89, 1994

[5] Selby JV. Case-control evaluations of treatment and program efficacy. *Epidemiol Rev.*, **16**, 90-101, 1994

[6] Weiss NS. Application of the case-control method in the evaluation of screening. *Epidemiol Rev.*, **16**, 102-108, 1994

[7] Dwyer DW., Strickler H., Goodman RA., Armenian HK.. Use of case-control studies in outbreak investigations. *Epidemiol Rev.*, **16**, 109-123, 1994

[8] Grobbee DE., Hoes AW. *Arno W. Hoes. Cilinical epidemiology.* Jones and Bartlett Publishers, LLC.. Sudbury, Massachusetts U.S.A., 2007

[9] Breslow NE., Day NE. *Statistical methods in cancer research.* IARC Scientific Publications No. 32. Lyon, United Kingdom, 1980

[10] Friis RH., Sellers TA. *Epidemiology for public health practice.* Jones and Bartlett Publishers, LLC.. Sudbury, Massachusetts U.S.A., 2007

[11] Lopez-Carrillo L., Avila MH., Dubrow R. Chili pepper consumption and gastric cancer in Mexico: A case-control study. *American Journal of Epidemiology*, **139**, 263-271, 1994

[12] Broders AC. Squamouscell epithelioma of the lip. *JAMA*, **74**, 647-656, 1920

[13] Janes H., Sheppard L., Lumley T. Overlap bias in the case-crossover design, with application to air pollution exposures. *Statistics in Medicine*, **24**, 285-300, 2005

[14] Xu X., Zborowski J.V., Arena V.C., Rager J., Talbott E.O.. Case-crossover analysis of air pollution and cardiorespiratory hospitalizations: using routinely collected health and environmental data for tracking: science and data. *J Public Health Manag Pract.*, **14(6)**, 569-76, 2008

[15] Levy D., Sheppard L., Checkoway H., Kaufman J., Lumley T., Koenig J., Siscovick D.. A case-crossover analysis of particulate matter air pollution and out-of-hospital primary cardiac arrest. *Epidemiology*, **12(2)**, 193-9, 2001

[16] Cox DR., Hinkley DV. *Theoretical Statistics*. Chapman and Hall, Inc.. U.S.A., 1974

[17] Hosmer DW., Lemeshow S. *Applied Logistic Regression*. John Wiley and Sons, Inc.. U.S.A., 1989

[18] Staniswalis JG., Parks NJ., Bader JO., Maldonado YM. Temporal analysis of airborne particulate matter reveals a dose-rate effect on mortality in El Paso: Indications of differential toxicity for different particle mixtures. *Air and Waste Manage. Assoc.*, **55**, 893-902, 2005

[19] Bateson TF., Schwartz J. Control for seasonal variation and time trend in case-crossover studies of acute effects of environmental exposures. *Epidemiology*, **10**, 539-544, 1999

Curriculum Vitae

Yanlei Peng, the only daughter of Huimin Wang and Liang Peng, was born on November 9, 1984 in Handan, People's Republic of China. She attended the First High School in Handan. After three years, she was admitted in Tianjin University of Commerce when she was 18 years old. She completed her Bachelor's of Science degree in applied mathematics in June 2007. Right after that, she left China, and flew to the U.S.A. to enroll at the University of Texas at El Paso (UTEP), and began her graduate studies in Fall 2007. She was majored in statistics in the Department of Mathematical Sciences. During the course of her master's studies she used to work as a Teaching Assistant for the first semester. From spring 2008 to present, she worked as a Research Assistant in the department at UTEP. She plans to continue studying for the Ph.D. in statistics after completing her work on M.S.. She has obtained an opportunity to pursue a doctor's degree in statistics in the University of South Carolina at Columbia in Fall 2009.

Present address: 1700 Hawthorne St. Apt. 130

El Paso, Texas 79902

This thesis was typed by Yanlei Peng.