Title

Statistical methods for the analysis of unbalanced matched case-control designs.

Author

Gulesserian, Sevan K.

Publication Date

2016

UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical methods for the analysis of unbalanced matched case-control designs.

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Sevan K. Gulesserian

Dissertation Committee:
Professor Daniel L. Gillen, Chair
Professor Wesley O. Johnson
Associate Professor Scott M. Bartell

2016

DEDICATION

For my parents, who have been my biggest champion from day one. Their unconditional love and encouragement carried me through this process. For that, and much more, I am deeply indebted.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge and thank my reading commitee-Daniel Gillen, Wesley Johnson, and Scott Bartell. Their input and guidance through the entire journey of completing the dissertation has been invaluable. Thank you to the additional members of my advancement committee-Babak Shahbaba and Ralph Delfino. Thank you Rosemary Busta for your ongoing administrative support. I would also like to thank Ralph Delfino and Thomas Tjoa for introducing me to the primary study that was the motivation for dissertation and permitting me to use the data.

To my advisor, Daniel Gillen, whose selfless time and care I am deeply grateful for. With every set back and breakthrough he has been hopeful and encouraging. He always inspired me to do my best. I appreciate all the advice and guidance he has given me through the years.

Many thanks to my family and friends who have been constant supporters throughout the years.

CURRICULUM VITAE

Sevan K. Gulesserian

EDUCATION

Doctor of Philosophy in Statistics **2016**
University of California, Irvine *Irvine, CA*

Masters of Science in Economics **2010**
Purdue University *West Lafayette, IN*

Bachelors of Science in Economics **2006**
California State University, Northridge *Northridge, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2010-2012**
University of California, Irvine *Irvine, California*

Graduate Research Assistant **2008-2010**
Purdue University *West Lafayette, IN*

TEACHING EXPERIENCE

Teaching Assistant **2012-2016**
University of California, Irvine *Irvine, CA*

Lecturer **2015**
University of California, Irvine *Irvine, CA*

Teaching Assistant **2008-2009**
Purdue University *West Lafayette, IN*

REFEREED JOURNAL PUBLICATIONS

Asthma morbidity and ambient air pollution: Effect modification by residential traffic related air pollution **Jan 2014**
Epidemiology

On the power of bootstrap tests for stationarity: A Monte Carlo comparison **May 2014**
Empirical Economics

WORK IN PROGRESS

On the estimation of covariate effects in matched case-control design with multiple event per cluster
Joint work with Daniel L. Gillen

On frequentist parameter estimation of matched case-control studies with unbalanced cluster sizes and effect modifcation
Joint work with Daniel L. Gillen

A semi-parametric Bayesian hierarchical model for analyzing case-control studies with unbalanced cluster sizes
Joint work with Daniel L. Gillen

PROFESSIONAL MEMBERSHIPS

American Statistical Association (ASA)
International Biometric Society (IBS)
Institute of Mathematical Statistics (IMS)

ABSTRACT OF THE DISSERTATION

Statistical methods for the analysis of unbalanced matched case-control designs.

By

Sevan K. Gulesserian

Doctor of Philosophy in Statistics

University of California, Irvine, 2016

Professor Daniel L. Gillen, Chair

The research presented in this thesis focuses on the analysis of data arising from matched case-control designs, with particular emphasis on case-crossover designs. We begin by providing a scientific example that motivates the research presented, highlighting statistical issues raised in addressing the scientific goal of study. We provide background and notation that lays the foundation for the remainder of the dissertation. The occurrence of repeated events per patient or cluster and an imbalance in cluster sizes poses statistical challenges in the analysis of case-crossover studies (or more generally in matched case-control studies). We begin with a background of existing methods, then focus on methods to estimate association parameters in matched cases control designs while accounting for within-subject correlation in the data. The methods discussed assume the willingness to break the individual matched case-control bonds within matched sets, thereby accounting for within-subject correlation directly in the estimation procedure. It is illustrated that existing estimation procedures can result in severe bias depending upon the number of repeated events per patient/cluster and the magnitude of covariate effect on the response.

Then, methods are discussed where it is no longer acceptable to break the matched case-control bonds. These methods employ substantially different weighting methods to obtain parameter estimates, and the resulting estimand consistently estimated by each procedure is investigated. We focus on the scenario of varying matched set sizes (varying cluster sizes), where effect modification exists across clusters. It is shown that currently implemented frequentist methods for analyzing case-crossover data with unbalanced cluster sizes force one to choose between weighting schemes that estimate marginal or conditionally-weighted covariate effects.

In order to directly model and contrast marginal and subject-specific estimates of association in matched case-control studies, a novel method for obtaining estimates is developed. The proposed methodology allows for simultaneous estimation of both marginal and subject-specific covariate effects by implementing a semi-parametric Bayesian hierarchical framework.

Throughout, the utility of the resulting methodology is illustrated using data obtained from a case-crossover study of children sampled from Orange County, CA seeking to quantify the effect of air pollution exposure on the risk of asthma-related hospital encounters.

Chapter 1

Introduction

1.1 Motivating Example

Past studies have shown acute adverse changes in respiratory outcomes among children with asthma from short-term increases in exposure to ambient air pollutants, including fine particulate matter measuring 2.5 microns or less in width ($PM_{2.5}$), ozone (O_3), and nitrogen dioxide (NO_2) (Trasande and Thurston [2005]). Given this background, it was hypothesized that higher exposure to residential traffic-related air pollutants would be positively associated with exacerbated asthma events.

While considered the gold-standard for establishing cause-and-effect, an interventional study of the effect of exposure to residential traffic-related air pollutants on the risk of exacerbated asthma events, where individuals are randomly assigned an exposure level and monitored for the development of exacerbated asthma, would clearly be unethical. Further, since the event of interest is rare (exacerbated asthma events occur with low frequency, even in the pediatric asthma population), a prospective observational cohort study would be infeasible as it would require a large sample prospectively followed for an extended duration in order to produce adequate statistical information for estimating the association of interest.

Due to the above constraints, a retrospective sampling design was implemented to test the hypothesis that higher exposure to residential traffic-related air pollutants would be positively associated with exacerbated asthma events. Specifically, asthma-related hospital admissions and emergency department visits (hospital encounters) for children under 18 years of age were collected from hospital records at the Children's Hospital of Orange County (CHOC) and the University of California at Irvine's Medical Center (UCIMC) through the years of 2000-2008 (Delfino et al. [2014]). In addition, time-varying traffic-related air pollutant exposures including ambient $PM_{2.5}$, O_3, nitric oxide (NO), carbon monoxide (CO), NO_2, as well as temperature and relative humidity were recorded daily from the start of 2000 to the end of 2008.

The primary objective of the study was to evaluate the effect of traffic-related air pollution exposure on the risk of an asthma-related hospital encounter. In order to estimate the association between traffic-related air pollution exposure on the risk of an asthma-related hospital encounter, a type of matched case-control design termed the case-crossover design was utilized. In the case-crossover design, matching occurs within the subject, allowing for each patient to act as his/her own control because exposures are sampled from that patient's time-varying distribution of exposure. Under this design, the exposure at a time just before the patient's event (hospital encounter) can be compared with a set of referent times representing the expected distribution of exposure for nonevent follow-up times.

	No. Hospital Encounters			
	Emer. Dept. Visit	**Hosp. Admission**	**Total Encounters**	**No. Uniq. Subj.**
Subj. Char. %	**(n=8229)**	**(n=3165)**	**(n=11394)**	**(n=7751)**
Boys	62	62	62	63
Age in years				
0-4	52	62	55	55
5-12	38	32	36	36
13-18	10	6	9	9
Race/ethnicity				
White non-hispanic	36	35	36	36
White Hispanic	54	53	54	52
African American	3	4	3	3
Asian	3	4	3	4
Other/unkown	4	4	4	5
Source of payment				
Private insurance	36	41	37	38
Government or uninsured	62	53	60	58
Unkown	2	6	3	4

Table 1.1: Summary of socio-economic factors.

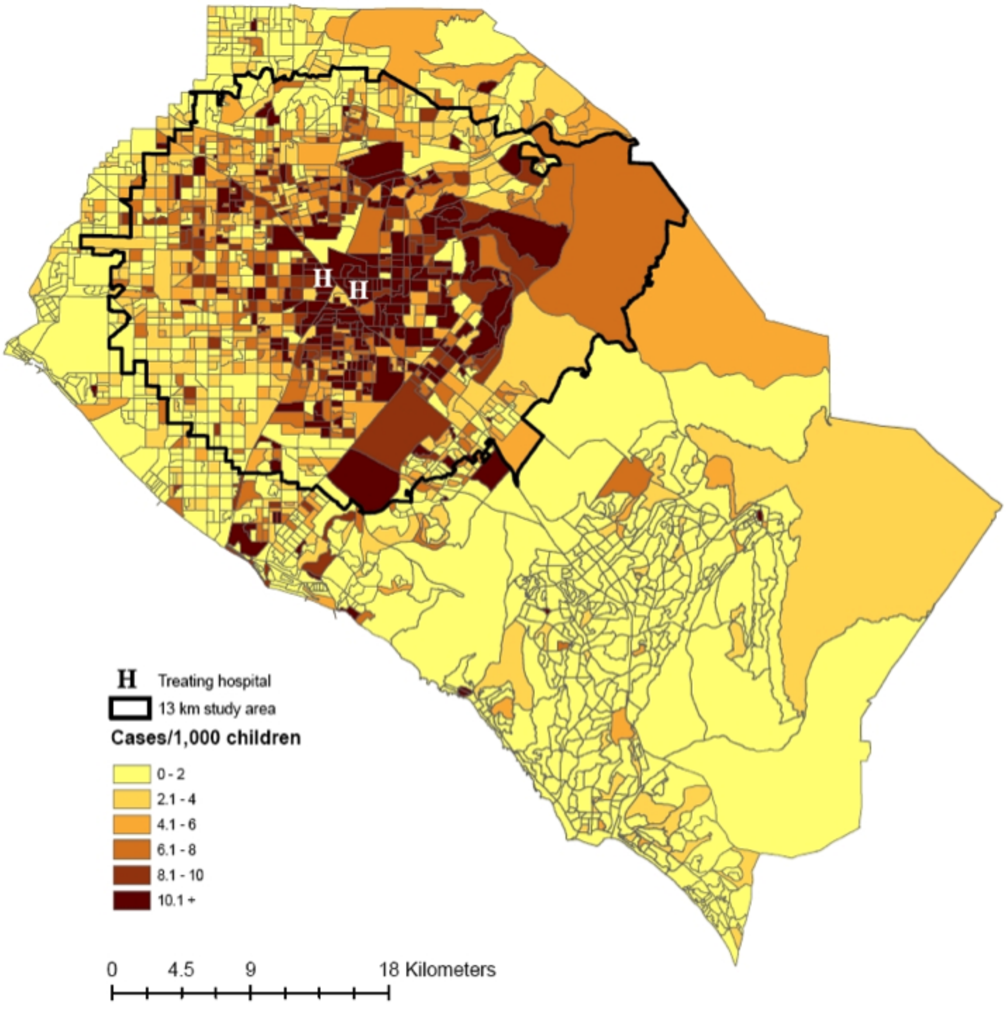

Figure 1.1: Distribution of cases across Orange County

Table 1.1 depicts basic summary statistics and socio-economic factors for patients included in the study. Figure 1.1 displays the distribution of cases across a map of Orange County. In total, $N = 7,751$ unique children were recorded to have at least one asthma-related hospital encounter between 2000 and 2008, and $11,394$ total encounters were recorded. As gleaned from comparing the total number of observed encounters to the number of unique children experiencing at least one encounter, some children experienced more than one hospital encounter over the study period. In fact, the number of encounters for a child varied from 1 to 17 over the course of the study.

1.1.1 Statistical Issues Presented by the Motivating Example

The design and resulting data described in the motivating example raise multiple statistical issues that must be considered. As is true with the asthma-related hospital encounters data, studies that utilize the case-crossover design often have an event of interest that can be experienced numerous times per subject. Additional examples include the effect of alcohol consumption on gout attacks (Zhang et al. [2006]) or the effect of medication changes on the risk of falls in elderly patients (Luo and Sorock [2008]). In the broader context of matched case-control designs, individual subjects in a case-crossover design represent a matching cluster (or matched set), and when repeated events can occur within the same subject, each cluster is comprised of the collection of the individual matched case-control pairs for each event within the subject. Further, data resulting from a case-crossover study with repeated events are likely to yield unbalanced cluster sizes since some individuals may have an inherently higher or lower propensity for the event than others. For example, in the context of the motivating example, patients with more severe asthma would likely be observed for greater numbers of asthma-related hospital encounters. The result is that the number of observed matched pairs for these patients would be greater than that of patients with less severe forms of asthma.

The occurrence of repeated events per patient or cluster and an imbalance in cluster sizes poses unique statistical challenges in the analysis of case-crossover studies (or more generally in matched case-control studies). First, one must consider how to accurately and efficiently estimate association parameters while accounting for the correlation within patients. While the conditional logistic regression model is almost universally used for the analysis of matched case-control studies, in Chapter 2 it is shown that the likelihood corresponding to this model is mathematically equivalent Cox's proportional hazard (PH) partial likelihood (Cox [1975]). Further, it is emphasized in Chapter 2 that the largest statistical software packages revert to fitting the Cox PH model when estimating the parameters in a conditional logistic regression model. One approach to accounting for within-patient correlation is to analyze the resulting data at the patient- or cluster-level. However, this approach yields multiple matched pairs within each cluster resulting in an analogous situation of multiple tied event times in the Cox PH partial likelihood. While several methods for accounting for ties in the Cox PH model have been proposed, the operating characteristics of these methods have not been considered in the setting of clustered matched case-control data.

While one can account for within-patient correlation by analyzing clustered matched case-control data at the patient- or cluster-level, beyond introducing ties into the likelihood, this approach breaks the matching bond between each index case and the corresponding control(s). The result is that parameter estimates may be biased due to the unfair comparisons between case and control exposures within the same cluster. When it is necessary to maintain the case-control pair bond, each individual index case and corresponding control can be incorporated separately into the conditional logistic regression likelihood. While variances of parameter estimates can be adjusted post-hoc in order to account for within-patient correlations, maintaining case-control bonds forces one to choose how individual index visits should be weighted during parameter estimation and the choice of weighting scheme can have substantial impacts on the scientific interpretation of the resulting parameter estimates.

1.2 Overview of the Thesis

The remainder of the thesis builds on the motivating example presented in Section 1.1 and the statistical issues raised in Section 1.1.1. The work presented here seeks (1) to describe past statistical contributions to the design and analysis of case-crossover studies, (2) to carefully examine operating characteristics of currently used methods and point out the deficiencies in these approaches, and (3) to propose a novel statistical methods that address the discovered deficiencies. To this end, in Chapter 2 a general review of the methodology to be used throughout the remaining chapters is provided. In Chapter 3, it is shown that the conditional logistic likelihood under numerous events per matched set (i.e. numerous events per subject in a case-crossover design) is mathematically equivalent to a Cox proportional hazards (PH) partial likelihood with tied event times within strata (matched sets). This is done in order to demonstrate that current software obtains parameter estimates in a matched case-control study design by maximizing the Cox PH partial likelihood using a transformation of the data. The methods discussed in this chapter assume the willingness to break the individual matched case-control bonds within matched sets, thereby accounting for within-subject correlation directly in the estimation procedure. In Chapter 4, methods are discussed where it is no longer acceptable to break the matched case-control bonds. The methods discussed in Chapter 4 employ substantially different weighting methods used to obtain parameter estimates, and the resulting estimand target by each estimation procedure is described. Throughout Chapter 4 the focus is on the scenario of varying matched set sizes (varying cluster sizes) as found in the asthma ER admissions motivating dataset, where effect modification exists across clusters.

In Chapter 4 it is shown that currently implemented frequentist methods for analyzing case-crossover data with unbalanced cluster sizes force one to choose between weighting schemes that estimate marginal or conditionally-weighted covariate effects. While both approaches address reasonable scientific goals, it would be desirable to directly model and contrast each of these estimands simultaneously. Building on the results of Chapter 4, in Chapter 5 a novel method for obtaining estimates and drawing inference in case-crossover studies with repeated events where effect modification across subjects may exist is developed. The methodology allows for simultaneous estimation of both marginal and subject-specific covariate effects. Specifically, a semi-parametric Bayesian hierarchical model to estimate subject-specific covariate effects is proposed.

Throughout the remainder of the thesis, the asthma hospital admissions study described in Section 1.1 motivates the developed research and illustrative examples are presented using data from this study. While the asthma hospital admissions study provides an excellent application for the developed methodology, the research presented throughout is presented in a general fashion that can be applied to any matched case-control study with numerous matched pairs within clusters.

Chapter 2

Background and Review

2.1 Review: Case-Control Studies with a Binary Exposure

In a *prospective design* seeking to study the effect of a factor (or factors) on the risk of experiencing a particular outcome, a healthy cohort of subjects sampled from the target population is followed throughout the course of a study. Baseline and possibly longitudinal measures of the factor(s) they are exposed to are collected and the outcome status of each subject in the cohort is ascertained. At the completion of the study, the estimated probability of the outcome occurring conditional upon factor level can be contrasted to estimate the association between the factor and the outcome. However, if the outcome occurs rarely in the target population, then with high probability there will be a low number of subjects that are observed to experience the outcome and a large number of subjects that did not. Hence little information will be attained on the difference in the probability of the outcome occurring based on the differing levels of the factor(s) of interest.

If the outcome rate is low, a prospective study is generally infeasible for the reasons described above. A more feasible approach is to consider a *retrospective design*. A retrospective design first samples subjects based on their outcome status, and then retrospectively measures their factor levels. This type of design ensures that a sufficient number of subjects with and without the outcome are included in the analysis. When the outcome of interest is binary, the retrospective design termed a *case-control design* since cases (or subjects known to have experienced the outcome of interest) and controls (or subjects known to not have experienced the outcome of interest) are first sampled, then the risk factor (or exposure) of interest is measured and hence is random by design.

Because of the efficiency of the case-control design for studying rare outcomes, this retrospective design is used throughout medical and health related fields where the determinants of rare adverse health outcomes are often of scientific interest. Historically, the retrospective case-control design gained popularity in the early 1920's in the field of cancer research (Broders [1920], Lombard and Doering [1928], Lane-Claypon [1926]) and has roots dating back nearly one and a half centuries when researchers employed a retrospective design to compare the occupations of men with pulmonary consumption to the occupations of men having other diseases (Lileinfield and Lilienfield [1979]).

2.1.1 Inference for Case-Control Studies with a Single Binary Exposure

Consider data resulting from case-control study with binary outcome, D (commonly used to indicate disease ($D = 1$) or no disease ($D = 0$)) and a single binary exposure, X (eg. $X = 1$ for exposed and $X = 0$ for non-exposed). Let a denote the total number of diseased subjects that were exposed, b denote the number of non-diseased subjects that were exposed, c denote the number of diseased subjects that were not exposed subjects, and d denote the number of

Exposure\Disease	$D = 1$	$D = 0$	Total
$X = 1$	a	b	t
$X = 0$	c	d	$N - t$
Total	N_1	N_0	N

Table 2.1: 2 × 2 contingency table depicting case-control data with a single binary outcome and single binary exposure

non-diseased that were not exposed. Letting N denote the total number of subjects sampled in the study, these data can then be represented using a 2 × 2 contingency table of the form given in Table 2.1.

One approach to drawing inference regarding the association between exposure and outcome using the data depicted in Table 2.1 is via Pearson's chi-square test for independence (Pearson [1900]). This approach focuses on testing the null hypothesis of independence between the column and row variable versus the alternative hypothesis of an association between the column and row variables (or disease and exposure in the context of the case-control study giving rise to table Table 2.1). Pearson's chi-square test statistics is intuitively formulated by contrasting (and standardizing) the observed count in each table cell with the expected count under the null hypothesis of independence between columns and rows. More specifically, Pearson's chi-square statistic a general $I \times J$ contingency table is is given by

$$\mathcal{X}^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k},$$

where K is the number of cells in the table (Table 2.1 has $K = 4$), O_k is the observed counts in cell k, and $E_k = \frac{\text{row}_k \text{ total} \times \text{column}_k \text{ total}}{N}$ is the expected count in cell k under the null hypothesis, $k = 1, ..., K$. Under the null hypothesis and assuming independence between observations conditional upon disease status, it can be shown that $\mathcal{X}^2$ is asymptotically distributed as a chi-square random variable with degrees of freedom equal to $(I - 1) \times (J - 1)$, where I and J denotes the number of rows and columns of the table that are used to formulate the test statistic, respectively.

Cornfield [1951] aimed to address criticisms that Pearson's chi-square test for independence does not directly assess the probability of disease by exposure level, a functional of primary interest to researchers. To this end, focus shifted to the odds of a disease given by $\frac{p}{1-p}$ where p denotes the probability of disease. The corresponding odds ratio is then defined as the ratio of odds comparing two different levels of exposure. Thus in the context of a case-control study with a single binary exposure, letting $p_1 \equiv P(D=1|X=1)$ denote the probability of disease for exposed subjects and $p_2 \equiv P(D=1|X=0)$ be the probability of disease for non-exposed subjects, the odds ratio comparing exposed to unexposed subjects is given by $\frac{p_1(1-p_2)}{p_2(1-p_1)}$. Cornfield [1951] demonstrated that the exposure odds ratio comparing diseased to non-diseased subjects, (OR_E), which can be estimated using a retrospective design, is the same as the disease odds ratio comparing exposed to non-exposed subjects, (OR_D), which is what would be obtained in a prospective design. The equivalence results from a straightforward application of Bayes' theorem relating the conditional probability of exposure (conditional on disease) to the joint probability of exposure and disease and the marginal probability of disease:

$$P(X=j|D=l) = \frac{P(X=j, D=l)}{P(D=l)} \quad \text{j=0,1, \ l=0,1.}$$

From Bayes' result it is then trivial to see that

$$\begin{aligned} OR_E &= \frac{P(X=1|D=1)P(X=0|D=0)}{P(X=0|D=1)P(X=1|D=0)} \\ &= \frac{P(D=1|X=1)P(D=0|X=0)}{P(D=0|X=1)P(D=1|X=0)} \\ &= OR_D. \end{aligned} \tag{2.1}$$

It is also worth noting that if the disease is rare (the primary motivation for using a retrospective design), then $P(D=0|X=j) \approx 1$ for $j=0,1$ and thus OR_D approximates the relative risk of disease, $RR_D = \frac{P(D=1|X=1)}{P(D=1|X=0)}$. More generally, $OR_D = RR_D \frac{1-P(D=1|X=1)}{1-P(D=1|X=0)}$. Thus only in the rare outcome setting can the odds ratio obtained from a retrospective design approximate the prospective relative risk. However, in general from Eq (2.1) under the scenario of a single binary exposure and a binary disease status, the case-control study provides an estimate for inferences regarding the disease odds ratio the same as if a prospective design had been implemented.

Using the notation of Table 2.1, a consistent estimator of the odds ratio is given by

$$\widehat{OR} = \frac{ad}{cb}, \tag{2.2}$$

and statistical inference for the odds ratio proceed as follows: In a retrospective sampling design, the column totals N_1 and N_0 are fixed. Therefore, $a|N_1 \sim \text{Binomial}(N_1, p_1 = P(X = 1|D = 1))$ and $b|N_0 \sim \text{Binomial}(N_0, p_0 = P(X = 1|D = 0))$. Noting that a and b are the sum of N_1 and N_0 independent and identically distributed Bernoulli random variables, it follows from the central limit theorem that:

$$\sqrt{N}\left(\frac{a}{N_1} - p_1\right) \xrightarrow{D} \text{N}(0,\, p_1(1-p_1)/\gamma_1), \tag{2.3}$$

and similarly for $\frac{b}{N_0}$, where $\lim\limits_{N\to\infty} \frac{N_j}{N} = \gamma_j$ for $j = 0, 1$. Using the results in (2.3) a straightforward application of the delta method yields

$$\ln(\widehat{OR}_D) \dot{\sim} N(\ln(OR_D), \text{var}(\ln(OR_D)), \tag{2.4}$$

where $\text{var}(\ln(OR_D)) = \frac{1}{\gamma_1 p_1(1-p_1)} + \frac{1}{\gamma_0 p_0(1-p_0)}$ can be consistently estimated by $\widehat{\text{var}} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$.

Formal statistical testing and inference for the disease odds ratio can thus be obtained using the result in (2.4).

Exposure\Disease	$D = 1$	$D = 0$	Total
$X = 0$	a_{01}	b_{00}	t_0
$X = 1$	a_{11}	b_{10}	t_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X = K$	a_{K1}	b_{K0}	t_K
Total	N_1	N_0	N

Table 2.2: 2-way table for K levels of exposures.

2.1.2 Extensions to Ordinal Discrete Exposures

The methodology discussed above for the analysis of 2×2 tables has been extended to settings where the exposure level has $K > 1$ levels. Early work addressing this was presented by Cochran [1954] and Armitage [1955]. This led to expanding the work on the chi-square test set forth by Pearson, where the aim was to assess the presence of an association between a variable with two categories and a variable with K ordered categories. In this case, one is considered with the analysis of a $K \times 2$ table as shown in Table 2.2. The test modifies the Pearson chi-square test of independence to incorporate an assumed ordering in the effects of the K categories of the exposure variable. Taking an example of treatment level effects on disease status, doses of a treatment can be ordered as 'low', 'medium', and 'high', and it might reasonably be hypothesized that treatment benefit increases with escalating dose. Details of the Cochran-Armitage test for trends can be found in Cochran [1954] and Armitage [1955].

Finally, a common theme throughout the statistical literature focused on the analysis of contingency tables is the reconciliation in what can be inferred from a retrospective sampling design and the more natural, but often less efficient, prospective sampling design. Along these lines, Mantel and Haenszel [1959] clarified the relationship between a cohort (prospective) and case-control (retrospective) study with the observation that the primary goal with the retrospective study is to reach the same conclusions had a prospective study been performed. This fundamental goal continues to drive much of the regression methods considered in the following sections, as well as the methodology developed in the remainder of the thesis.

2.1.3 Adjustment for Confounding in Contingency Tables

Continuing the discussion from the previous section of the effect of a single binary exposure on a disease outcome, the issue of confounding arises. Multiple definitions of confounding have been proposed in the statistical and epidemiological literature. For the purposes of this thesis, a confounding variable is defined as a factor that is causally associated with both the outcome and the explanatory variable of interest. Confounding is a critical concept in the analysis and interpretation of observational data since unadjusted confounding can lead to the appearance of an association between the outcome and the explanatory variable interest, even though such an association may not truly exist. As such, methods for adjusting for potential confounding factors are critical when analyzing data arising from an observational study.

Exposure\Disease	$D = 1$	$D = 0$	Total
$X = 1$	a_k	b_k	t_k
$X = 0$	c_k	d_k	$N_k - t_k$
Total	N_{1k}	N_{0k}	N_k

Table 2.3: 2-way table with a single binary outcome and single binary exposure for the k-th level of a confounder.

Assuming a categorical confounding variable with $K > 1$ levels, Mantel and Haenszel [1959] considered the adjustment for confounding by arranging the data into a series of K independent two-way contingency tables, one at each level of the potential confounder. In this case, it is assumed that the confounder is both observable and was measured. For notational purposes, the two-way table (assuming a single binary exposure) at the k-th level of the confounder is depicted in Table 2.3. The Mantel-Haenszel (MH) summary odds ratio estimator inherently assumes a common odds ratio at each level of the confounder (ie. no interaction exists between the exposure covariate and confounding covariate), and is defined as $\hat{\psi}_{MH} = \frac{\sum\limits_{k} R_k}{\sum\limits_{k} S_k}$, where $R_k = \frac{a_k d_k}{N_k}$ and $S_k = \frac{b_k c_k}{N_k}$. Intuitively, by first stratifying on the level of the confounder, the effect of confounding is conditioned out and the conditional odds ratios can then be marginalized over the levels of the confounder. This procedure was eventually adopted for routine use by epidemiologists, who benefited from seeing their data arranged in tabular form and making comparisons of individual and summary relative risks that signaled possible heterogeneity across the k levels of the confounder (Breslow [1996]). Robins et al. [1986] and Phillips and Holland [1987] later developed an estimator for the variance of $\hat{\psi}_{MH}$, and asymptotic approximations of the distribution of the estimate were developed to allow for inference regarding the adjusted odds ratio.

To further expand on the development of the MH estimator, first note that $\mathrm{E}(R_k) = \psi_k \mathrm{E}(S_k)$, where ψ_k denotes the true odds ratio in table k . Then assuming a common value for ψ_k and setting $R = \sum\limits_{k} R_k$ and $S = \sum\limits_{k} S_k$, $\hat{\psi}_{MH}$ is the solution to the unbiased estimating equation $R - \psi S = 0$. Further, under paired binomial sampling, the variance of the k-th table's contribution to the estimating equation is given by

$$N_k^2 \operatorname{var}(R_k - \psi S_k) = E[(a_k d_k + \psi b_k c_k)(a_k + d_k + \psi(b_k + c_k))]. \tag{2.5}$$

Letting $\beta = \ln(\psi)$, it can be shown that

$$\hat{\beta}_{MH} = \ln(\hat{\psi}_{MH}) = \beta + \frac{R - \psi S}{E(R)} + o_p\left(\frac{\operatorname{var}(R)}{\mathrm{E}^2(R)} + \frac{\operatorname{var}(S)}{\mathrm{E}^2(S)}\right). \tag{2.6}$$

Combining Eqs. (2.5) and (2.6), an estimate of the variance of $\ln(\hat{\psi}_{MH})$ is given by

$$\widehat{\operatorname{var}}(\hat{\beta}_{MH}) = \frac{1}{R^2} \sum_{k} \frac{1}{N_k^2} [(a_k d_k + \hat{\psi}_{MH} b_k c_k)(a_k + d_k + \hat{\psi}_{MH}(b_k + c_k))]. \tag{2.7}$$

Further, from Eqs. (2.6) and (2.7) the asymptotic distribution of can be found $\hat{\beta}_{MH}$ and shown to be

$$\hat{\beta}_{MH} = \ln(\hat{\psi}_{MH}) \dot{\sim} \operatorname{Normal}(\ln(\psi), \widehat{\operatorname{var}}(\ln(\hat{\psi}_{MH}))). \tag{2.8}$$

From the result in (2.8), asymptotic inference regarding the confounding adjusted odds ratio can be conducted using usual frequentist methods.

As previously noted, marginalization of the conditional odds ratio across the K strata of the confounding covariate only tends to make sense when the the strata-specific odds ratio are homogeneous. To test the hypothesis of homogeneity, the Breslow test of homogeneity can be conducted (Breslow and Day [1980]). Further, if one simply wishes to test whether the association between exposure and disease exists at any level of the confounding factor (without first marginalizing over the levels of the confounder) the Cochran-Mantel-Haenszel test (Mantel [1963]) can be used. Specifically, the Cochran-Mantel-Haenszel considers a null hypothesis of the form $H_0 : \psi_1 = \psi_2 = ... = \psi_K = 1$, where ψ_k is the odds ratio comparing exposed to unexposed among subjects with confounding covariate level k, $k = 1, \ldots, K$. The Cochran-Mantel-Haenszel test statistic is then constructed as the standard sum, over all strata, of the contrasted observed and expected cell counts with asymptotic distribution given by:

$$\chi_{CMH} = \frac{\sum\limits_k (a_k - \frac{t_k N_{1k}}{N_k})}{\sum\limits_k \frac{t_k N_{1k} N_{0k} (N_k - t_k)}{N^2 (1-N)}} \dot{\sim} \chi^2_1.$$

2.2 Review: Case-Control Studies with Continuous and Categorical Exposures

The discussion up to now has primarily focused on the setting of a single binary exposure and it's association with the the probability of a binary outcome (disease or no disease). Methods to evaluate the simultaneous effects of multiple quantitative risk factors on disease rates began to appear in the 1960s. The goal was not just to differentiate between two populations (exposed and non-exposed), but rather to make inference on the risk of developing disease during a specified time period as a function of one or more exposures variables measured on each subject.

Cornfield et al. [1961] showed that if the multivariate exposure $\boldsymbol{X}$ among the diseased and the non-diseased population is normally distributed with different means but a common covariance matrix, then given a subject's exposure values $\boldsymbol{X} = \boldsymbol{x}$, the probability of developing the disease can be represented by the logistic response curve:

$$P(D = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{x\beta})}{1 + \exp(\alpha + \boldsymbol{x\beta})},$$

where the parameters ,α and $\boldsymbol{\beta}$, are simple functions of the moments of the exposure distributions and the marginal distribution of the disease. The direct use of the logistic specification above (not the normality of the exposure distribution) was recommended by Cox [1966], as it required fewer assumptions. Day and Kerridge [1967] later noted that the full likelihood based on the joint distribution of (D, X) can be factored into two pieces, the conditional likelihood specified by the logistic model, and the marginal likelihood of the exposures. Both pieces could then be maximized separately, allowing for the exposure distributions among the controls to be arbitrary.

Let $\boldsymbol{x}$ be a 1 by p vector of covariates and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. Based on the work of Cox [1966], Seigel and Greenhouse [1973] noted a key feature of the logistic model for case-control studies. Namely, they noted that

$$\frac{P(D = 1|\boldsymbol{X} = \boldsymbol{x}_1)P(D = 0|\boldsymbol{X} = \boldsymbol{x}_0)}{P(D = 0|\boldsymbol{X} = \boldsymbol{x}_1)P(D = 1|\boldsymbol{X} = \boldsymbol{x}_0)} = \exp[(\boldsymbol{x}_1 - \boldsymbol{x}_0)\boldsymbol{\beta}], \tag{2.9}$$

which is to say the disease odds ratio for comparing exposure levels of $\boldsymbol{x}_1$ to $\boldsymbol{x}_0$ has an exponential form, and thus $(\boldsymbol{x}_1 - \boldsymbol{x}_0)\boldsymbol{\beta}$ is the log odds ratio for comparing $\boldsymbol{x}_1$ to $\boldsymbol{x}_0$. The issue is that the case-control sampling design will contain terms of the form $P(\boldsymbol{X}|D)$, not $P(D|\boldsymbol{X})$.

Again, noting the need to reconcile inference obtained from a retrospective sampling design to that from a prospective design, Prentice and Pyke [1979] set out to show how the parameter estimates in the directly specified logistic case-control model can be obtained from a retrospective design and related to the corresponding prospective estimands. Let $\boldsymbol{x}_0$ denote a baseline $1 \times p$ vector of covariates (i.e. vector of 0s) and $\boldsymbol{\beta}_j$ a $p \times 1$ vector of parameters. The prospective model for the probability of disease conditional on $\boldsymbol{x}_0$ is given by

$$p(D = j|\boldsymbol{x}) = \frac{\exp(\alpha_j + \boldsymbol{x}\beta_j)}{\sum\limits_{j=0}^{1}\exp(\alpha_j + \boldsymbol{x}\beta_j)}, \quad j = 0, 1, \tag{2.10}$$

with $\boldsymbol{\beta}_0 = \boldsymbol{0}$ and $\gamma_0 = 0$ for uniqueness. Then the odds ratio of comparing exposure levels $\boldsymbol{x}$ to the baseline or referent level is calculated to be:

$$\exp\left[(\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{0}})\boldsymbol{\beta}_{\boldsymbol{j}}\right] \tag{2.11}$$

Additionally, (2.10) can be recovered by beginning with (2.11) and defining α as follows:

$$\alpha_j = \log\left(\frac{p(D = j|\boldsymbol{x})}{p(D = 0|\boldsymbol{x}_0)}\right) - \boldsymbol{x}_{\boldsymbol{0}}\boldsymbol{\beta}_{\boldsymbol{j}}.$$

Utilizing the equality between the prospective and retrospective odds ratios, and the representation in (2.11), the following can be calculated for $j = 0, 1$:

$$P(\boldsymbol{X} = \boldsymbol{x}|D = j) = c_j\text{exp}[\gamma(\boldsymbol{x}) + \boldsymbol{x}\boldsymbol{\beta}_j], \tag{2.12}$$

where $\gamma(\boldsymbol{x}_1) = \log\left(\frac{P(\boldsymbol{x}|D=0)}{P(\boldsymbol{x}_0|D=0)}\right)$ for all $\boldsymbol{x}$ and $c_j = c_j(\gamma, \boldsymbol{\beta}_j)$, a normalization factor.

Now consider a retrospective design and suppose n_0 controls and n_1 cases are sampled from their respective subpopulations. Let $\boldsymbol{x}_{ji}$ for $i = 1, 2, ..., n_j$ denote the n_j regressor variables (a $1 \times p$ vector) in disease group j, $j = 0, 1$, and set $n = n_0 + n_1$. The likelihood function resulting from the retrospective sampling design is then given by

$$\prod_{j=0}^{1}\prod_{i=1}^{n_j} P(\boldsymbol{x}_{ji}|D=j) = \prod_{j=0}^{1}\prod_{i=1}^{n_j} c_j \exp[\gamma(\boldsymbol{x_{ji}}) + \boldsymbol{x_{ji}}\boldsymbol{\beta}_j]. \tag{2.13}$$

As noted by Prentice and Pyke [1979], re-parameterization can clarify the estimation problem. To this end, let $q(\boldsymbol{x}) = [\exp(\gamma(\boldsymbol{x})]\sum_{l=0}^{1}\frac{n_l}{n}c_l\exp(\boldsymbol{x}\boldsymbol{\beta}_l)$. Solving for $\exp(\gamma(\boldsymbol{x}))$, and plugging in the result into (2.12):

$$P(\boldsymbol{x}|D=j) = \left[\exp(\delta_j + \boldsymbol{x}\boldsymbol{\beta}_j)/\sum_{l=0}^{1}\exp(\delta_l + \boldsymbol{x}\boldsymbol{\beta}_l)\right] q(\boldsymbol{x})\frac{n}{n_j}, \tag{2.14}$$

where $\delta_j = \log(c_j n_j/n)$.

The likelihood in (2.13) can then be written as

$$L \propto \left[\prod_{j=0}^{1}\prod_{i=1}^{n_j} p_j(\boldsymbol{x}_{ji})\right]\left[\prod_{j=0}^{1}\prod_{i=1}^{n_j} q(\boldsymbol{x}_{ji})\right] = L_1 L_2,$$

where $p_j(\boldsymbol{x}) = \exp(\delta_j + \boldsymbol{x}\boldsymbol{\beta}_j)/\sum_{l=0}^{1}\exp(\delta_l + \boldsymbol{x}\boldsymbol{\beta}_l)$ for $j = 0, 1$.

The parameters and $q(.)$ are restricted by the constraint that (2.12) is a probability distribution for each j and hence must satisfy

$$\frac{n_j}{n} = \int p_j(\boldsymbol{x})q(\boldsymbol{x})d\boldsymbol{x}, \tag{2.15}$$

where if $\boldsymbol{x}$ is discrete the integration becomes a summation.

Maximizing L without considering the constraints, the parameter estimates are solutions to the following estimating equations:

$$\frac{\partial \log L_1}{\partial \delta_j} = n_j - \sum_{m=0}^{1}\sum_{i=1}^{n_m} p_j(\boldsymbol{x}_{mi}) = 0.$$

$$\frac{\partial \text{log} L_1}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n_j} \boldsymbol{x}_{ji}^{'} - \sum_{m=0}^{1} \sum_{i=1}^{n_m} \boldsymbol{x}_{mi}^{'} p_j(\boldsymbol{x}_{mi}) = 0.$$

However, Prentice and Pyke [1979] note that the non-parametric maximum likelihood estimator of $q(.)$ is the empirical probability function $\hat{q}(.)$ that assigns mass s/n to any value of $\boldsymbol{x}$ that is observed with multiplicity s and 0 elsewhere. Prentice and Pyke [1979] then go on to show that unconstrained maximum likelihood estimators $\hat{\delta}_1$, $\hat{\boldsymbol{\beta}}_1$ (again setting $\delta_0 = 0$ and $\boldsymbol{\beta}_0 = 0$ for uniqueness) and $\hat{q}(.)$ satisfy the constraint in (2.15). Along with the fact that $\int \hat{q}(\boldsymbol{x}) d\boldsymbol{x} = 1$, the conclusion is reached that $((\hat{\delta}_1, \hat{\boldsymbol{\beta}}_1, \hat{q}(.))$ are the desired constrained maximum likelihood estimators.

The implication of the above results from Prentice and Pyke [1979] is that if the prospective model were applied to retrospectively sampled case-control data, the likelihood equations would be identical, with α in the place of δ. Most importantly, this implies that the disease odds ratio parameters can be obtained by maximizing the retrospective likelihood and hence usual logistic regression techniques are justified for estimating the disease odds ratio based upon retrospectively sample data.

Standard maximum likelihood theory can be used to obtain asymptotic distribution of the MLEs obtained above, namely that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I^{-1})$ where I denotes the information matrix. Hypothesis tests for $\boldsymbol{\beta}$ can utilize methods like the score test or likelihood ratio tests.

2.3 Matched Case-Control Designs and the Conditional Logistic Regression Model

In Section 2.1.3 the need for confounding adjustment in observational studies was reviewed. In that section, the Mantel-Haenszel estimator was introduced as a means to adjust for a single discrete confounding factor in the analysis of contingency tables that may arise from a retrospective study design. In Section 2.2 the logistic regression model was discussed, which allows for adjustment of multiple continuous and/or discrete confounding factors and the seminal results of Prentice and Pyke [1979] that established the justification for use of standard logistic regression modeling of retrospectively sampled data was reviewed. However, regression-based adjustment for confounding still has limitations. First, it is generally infeasible, in terms of estimating parameters given a fixed sample size, to control for too many confounding factors. In addition, regression-based adjustment forces the decision of functional form when modeling a confounding factor, and if the functional form of the confounder is mis-specified, residual confounding can still exist. Given these potential drawbacks, it can be advantageous to adjust for confounding by design through matching. In the context of a case-control study, matching cases to controls with respect to confounding factors will fully adjust for confounding since since both case and controls will have the same value for the factor. In this section, the statistical analytic techniques that have been proposed for the analysis of data stemming from a retrospective matched design are reviewed.

2.3.1 Analysis of Matched Data with a Single Binary Exposure

For a moment, return to the 2×2 contingency table setting and consider the matched pair data depicted in Table 2.4. Suppose interest lies in a single binary exposure and that N cases are sampled with a matched control to each case. Due to the matching, which introduces

	Control pair member		
Case pair member	Exposed	Non-exposed	Total
Exposed	a	b	t
Non-exposed	c	d	$N-t$
Total	N_1	N_0	N

Table 2.4: 2-way table for matched pairs with a single binary outcome and single binary exposure

correlation among observations, the observations within the sample are not independent and hence the previously reviewed methods for drawing inference between exposure and outcome do not apply. Instead, McNemar's test (McNemar [1947]) can be used to test for the association of interest. In order to test if exposure is associated with outcome, McNemar's test uses only the number discordant pairs (b and c in Table 2.4), since the other pairs provide no information regarding the differential status of exposure among cases and controls. Intuitively, if the difference between b and c is large, this would imply that cases to be more (or less) exposed than their matched controls. This lays the foundation for the test. More precisely, McNemar's test statistic and the asymptotic distribution of the statistic are given by:

$$\chi^2_{MN} = \frac{(b-c)^2}{b+c} \,\dot{\sim}\, \chi_1$$

In addition, it can be easily shown that a consistent estimator of the odds ratio based upon data arranged as in Table 2.4 is given by

$$\widehat{OR} = \frac{b}{c}.$$

2.3.2 Analysis of Matched Data with Multiple Adjustment Covariates

Returning to the scenario of multiple continuous and categorical exposures, Breslow et al. [1978a] and Breslow and Day [1980] began to investigate what they termed the stratified logistic regression in the context of matched case-control data. Suppose that each case is matched to $M \geq 1$ controls base upon some set of potential confounding factors. Such a design is referred to as a 1-to-M $(1 : M)$ matched design. Further suppose that the population at risk is stratified on such a fine grid that each case and its matched controls are drawn from the same stratum. Letting $\boldsymbol{S_i}$ denote the observed and unobserved matching factors that define the i^{th} stratum $(i = 1, 2, ..., n)$, a prospective model for the probability of disease can be written as

$$P(D = 1 | \, \boldsymbol{S_i}, \boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\alpha_i + \boldsymbol{x\beta})}{1 + \exp(\alpha_i + \boldsymbol{x\beta})}. \tag{2.16}$$

The model in (2.16) involves a separate parameter, α_i, for each of the n strata and allows for inclusion of possible interactions between matching variables and exposures included in the explanatory vector $\boldsymbol{x}$. In this model specification, the curse of dimensionality is evident since as the sample size increases, the number of parameters to estimate increases proportionally. As such, it would be inefficient to estimate each α_i separately. To highlight this issue, consider the extreme case of a single case and a single matched control (a 1:1 matched design) in each strata. In this scenario, a sample of only size 2 would be available for estimation of each stratum specific parameter regardless of how large n is, resulting in the Neyman-Scott paradox. Fortunately, as pointed out by Breslow and Day [1980], the stratum specific α_i can be eliminated from the likelihood by conditioning on an ancillary statistic. In this case, the conditioning is done with the unordered set of exposures for the cases and controls in each stratum, which is equivalent to the number of cases in each stratum. This conditioning is elaborated for the remainder of this section.

Consider a single stratum's contribution to the full conditional likelihood. Continuing with the setting of $1 : M$ matching in each stratum, let $\boldsymbol{Y}_i$ be a vector of size $M + 1$, where each element can be a 0 to denote no disease and 1 for disease. Further, let j denote the j^{th} observation in the i^{th} strata, $i = 1, \ldots, n$, $j = 1, \ldots, M + 1$. Then the unconditional probability for the j^{th} observation in strata i can be modeled as

$$P(Y_{ij} = y_{ij}|\boldsymbol{S_i}) = \frac{\exp(y_{ij}[\alpha_i + \boldsymbol{x_{ij}\beta}])}{1 + \exp(\alpha_i + \boldsymbol{x_{ij}\beta})},$$

and noting independence within strata we have

$$P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{S_i}) = \exp\left(\sum_{j=1}^{M+1} y_{ij}[\alpha_i + \boldsymbol{x_{ij}\beta}]\right) / \prod_{j=1}^{M+1}(1 + \exp(\alpha_i + \boldsymbol{x_{ij}\beta})). \tag{2.17}$$

If the likelihood contribution from strata i is conditioned on the number of cases in that strata, namely $\sum\limits_{j=1}^{M+1} y_{ij} = 1$, then (2.17) will factor into

$$\begin{aligned}
P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\sum_{j=1}^{M+1} y_{ij} = 1, \boldsymbol{S_i}) &= \frac{P(\boldsymbol{Y}_i = \boldsymbol{y}_i) \times I(\sum\limits_{j=1}^{M+1} y_{ij} = 1)}{\sum\limits_{\{\boldsymbol{y}_i^*:\sum\limits_j y_{ij}^*=1\}} P(\boldsymbol{Y}_i = \boldsymbol{y}_i^*)} \\
&= \begin{cases} \frac{\exp\{\alpha_i + \sum\limits_j y_{ij} x_{ij} \beta\}}{\sum\limits_{\{\boldsymbol{y}_i^*:\sum\limits_j y_{ij}^*=1\}} \prod\limits_j P(Y_{ij}=y_{ij}^*)} & , \sum\limits_{j=1}^{M+1} y_{ij} = 1 \\ 0 & , \sum\limits_{j=1}^{M+1} y_{ij} \neq 1 \end{cases} \\
&= \begin{cases} \frac{\exp(\alpha_i + \sum\limits_j y_{ij} \boldsymbol{x}_{ij} \boldsymbol{\beta})}{\sum\limits_{\{\boldsymbol{y}_i^*:\sum\limits_j y_{ij}^*=1\}} \exp(\alpha_i + \sum\limits_j y_{ij}^* \boldsymbol{x}_{ij} \boldsymbol{\beta})} & , \sum\limits_{j=1}^{M+1} y_{ij} = 1 \\ 0 & , \sum\limits_{j=1}^{M+1} y_{ij} \neq 1 \end{cases} \\
&= \begin{cases} \frac{\exp(\sum y_{ij} \boldsymbol{x}_{ij} \boldsymbol{\beta})}{\sum\limits_{\{\boldsymbol{y}_i^*:\sum\limits_j y_{ij}^*=1\}} \exp(\sum\limits_j y_{ij}^* \boldsymbol{x}_{ij} \boldsymbol{\beta})} & , \sum\limits_{j=1}^{M+1} y_{ij} = 1 \\ 0 & , \sum\limits_{j=1}^{M+1} y_{ij} \neq 1. \end{cases} \qquad (2.18)
\end{aligned}$$

From (2.18) it can be seen that upon conditioning on the strata and number of cases within the strata, the stratum specific parameters α_i are factored out of the likelihood contribution for strata i, and hence do not require estimation. Without loss of generality, let $\boldsymbol{x}_{i1}$ be the exposure values for the case and $\boldsymbol{x}_{ij}$ $,j = 2, 3, ..., M + 1$ be the values for the controls. Then the above likelihood contribution for the i^{th} strata from (2.18) is given by

$$\frac{\exp(\boldsymbol{x}_{i1}\beta)}{\exp(\boldsymbol{x}_{i1}\boldsymbol{\beta}) + \sum\limits_{j=2}^{M+1}\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})}. \tag{2.19}$$

Note that $\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})$ is the odds ratio for comparing exposure levels of $\boldsymbol{x}_{ij}$ to exposure levels of 0 for all exposures, $\boldsymbol{x} = \boldsymbol{0}$. Intuitively, the likelihood contribution for the i^{th} stratum in (2.19) reduces to comparing the covariate values for the case to the covariate values of all members within the same strata, which includes both the case and the matched controls. The above model specification and likelihood contribution gives rise to the full conditional logistic likelihood based on the full study sample:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \sum_{j=1}^{M+1} y_{ij} = 1, \boldsymbol{S}_i). \tag{2.20}$$

The conditional logistic likelihood shown in (2.20) can be maximized to obtain parameter estimates for $\boldsymbol{\beta}$, which can in-turn be used to obtain disease odds ratio estimates.

One of the difficulties of retrospective sampling designs is appropriate matching of controls to cases. For example, consider matching a case child admitted to the hospital for a disease outcome to a similarly aged control child identified at home. The matching factor of age will be controlled for by design, but there is potential for confounding to other socio-economic factors, such as health insurance status. As a result, unmatched factors would need to be included in the model via regression adjustment (assuming they were observed). This will result in the estimation of additional parameters and the potential for residual confounding as previously mentioned. Unfortunately, due to logistical consideration, it is often infeasible to identify appropriate controls when matching on many criteria. However, in some studies it is possible to use a case as his/her own control, hence creating a matching scheme that controls for all within-subject invariant factors by design. In the next section, such a design is elaborated on.

2.3.3 The Case-Crossover Design

The case-crossover design can be viewed as a hybrid between a matched case-control design and a traditional crossover design. As previously noted, in a retrospective matched case-control design inference is based on the comparison of the exposures between the case and control(s) within each matching set, where controls are matched to cases based on a specified set of criteria. In a traditional crossover design investigating two treatments, each subject receives each treatment once, in a randomized order, and the outcome following each treatment regime is contrasted within the subject. Merging these two designs results in the case-crossover design. In this setting, each case subject serves as his/her own control, but is from a different time period where the event that defines case status was not experienced. Thus, since the same subject is both the case and control, observed and unobserved time in-variant matching factors are controlled for by design.

Historically, the case-crossover design was initially developed to study the effects of transient, short-term exposures on the risk of acute events (Maclure [1991]). This type of design represents a valid and efficient design for studying relationships between exposures and events with the following characteristics: 1) the individual exposure varies within short time intervals; 2) the disease has abrupt onset and short latency for detection; and 3) the induction period is short (Jaakola [2003]). The above implies that subject level exposures must vary within short time periods and induce the event of interest within a short time period. Additionally, the event's onset must be able to be defined and observed. As previously noted, by comparing exposures near event periods to exposures during non-event periods, each subject is able to act as his/her own control.

Implementation of the case-crossover design does require careful thought. Once the data for cases are obtained, the case subject's exposure from a different time period (past or future) sufficiently distant from the case exposure period is used as the controls exposure (Navidi [1998], Navidi and Weinhandl [2002]). This window of time for the control is known as the *referent* period. Since controls are created from the case subjects, there is no risk of selection bias, which occurs when controls are not representative of the population from which the cases arise from. However, there is a risk of *overlap bias*, meaning that the score equations resulting from the likelihood based on the case-crossover data do not have mean zero. Improperly choosing referent times will lead to overlap bias, which in turn will lead to increased bias in coefficient estimates. As such, the choice of referent periods is further elaborated on in the coming paragraphs.

Let $\boldsymbol{x}$ be a shared exposure series, defined at times $t = 1, 2, ..., T$ common to all $i = 1, 2, ..., n$ subjects. Let the index (case) time for subject i be denoted by t_i, the exposure at the index time be denoted by $\boldsymbol{x}_{t_i}$, and let W_i represent the referent window for subject i (which includes the index and all referent periods).

To begin, consider the estimating equations from a conditional logistic regression likelihood. Because the case-crossover design represents a special case of the matched case-control design, the conditional logistic regression likelihood derived earlier still applies here, and the resulting estimating equations obtained from the conditional logistic regression likelihood in the current context are given by

$$\sum_{i=1}^{n} U_i(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n} (\boldsymbol{x}_{it_i} - \sum_{t \in W_i} \boldsymbol{x}_t \frac{\exp(\boldsymbol{x}_t \boldsymbol{\beta})}{\sum\limits_{s \in W_i} \exp(\boldsymbol{x}_s \boldsymbol{\beta})}) = \mathbf{0}.$$

Typically, the underlying model for the case-crossover design is the proportional hazards model for a rare disease (Navidi and Weinhandl [2002], Janes et al. [2005a]), with a constant baseline hazard (ie. an exponential model for time). In this case, the hazard for subject i at time t, given time-varying covariates $\boldsymbol{x}_{it}$ is given by

$$\lambda_i(t;\, \boldsymbol{x}_{it}) = \lambda_i \exp(\boldsymbol{x}_{it}\boldsymbol{\beta}).$$

The goal in choosing a referent selection scheme that will pick referent window W_i for subject i is to obtain a *localizable* and *non-ignorable* design. A localizable referent design means there exists an unbiased estimating equation restricted to the referent windows. Within localizable referent selection schemes, an ignorable referent selection scheme means that the referent sampling scheme can be ignored in conducting the analysis (i.e. the data likelihood does not depend on the referent scheme). Therefore the derivatives of the true conditional log-likelihood of the data with respect to the model parameters, $\boldsymbol{\beta}$, will be equal to the conditional logistic likelihood estimating equations, which are unbiased, and thus the conditional logistic likelihood estimating equations can be used to obtain consistent parameter estimates.

The likelihood function in terms of a defined referent selection scheme is constructed as follows. Assume a single event within each matched set i and let Y_{it} be an indicator of whether subject i's index time was on day t. If a localizable and ignorable referent selection scheme is chosen, then the likelihood of the data conditioning on the referent window, exposure series, and number of cases from subject i is:

$$\begin{aligned} P(T_i = t_i | \boldsymbol{x}, W_i, \sum_{s=1}^{T} Y_{is} = 1) &= \frac{P(T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1 | \boldsymbol{x}, W_i)}{\sum_{t=1}^{T} P(T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1 | \boldsymbol{x}, W_i)} \\ &= \frac{\lambda_i \exp(\boldsymbol{x}_{t_i}\boldsymbol{\beta})}{\sum_{t \in W_i} \lambda_i \exp(\boldsymbol{x}_t \boldsymbol{\beta})} \\ &= \frac{\exp(\boldsymbol{x}_{t_i}\boldsymbol{\beta})}{\sum_{t \in W_i} \exp(\boldsymbol{x}_t \boldsymbol{\beta})}. \qquad (2.21) \end{aligned}$$

The likelihood in (2.21) only depends on exposures at times within the referent windows, and the derivative of the natural log of this likelihood set to zero gives the conditional logistic likelihood estimating equations. Thus, the conditional logistic likelihood can be used to estimate the parameters $\boldsymbol{\beta}$, provided that a localizable and ignorable referent selection scheme is implemented.

An example of a non-localizable referent selection is the symmetric bi-directional design, which picks referent times based upon a fixed number of days before and after the index time. In this design the likelihood of the index times conditional on the referent windows reduces to equalling 1. This is a direct result of the fact that, based on this design, the referent window will determine the index time, since the index time is always the center of the referent window. Alternatively, the semi-symmetric bi-directional design picks referent times to be at either a pre-determined number of days, δ, before the index, or a pre-determined number of days after the index, with equal probability of picking the early or later referent time. In this referent selection design, $W_i = \{t_i \pm \delta\}$ with probability 0.5 in each direction (+ or -), and the conditional likelihood obtained from this referent selection scheme is given by

$$\begin{aligned} P(T_i = t_i | \boldsymbol{x}, W_i, \sum_{s=1}^{T} Y_{is} = 1) &= \frac{P(W_i = w_i | \boldsymbol{x}, T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1)\, P(T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1 | \boldsymbol{x})}{\sum_{t=1}^{T} P(W_i = w_i | \boldsymbol{x}, T_i = t, \sum_{s=1}^{T} Y_{is} = 1)\, P(T_i = t, \sum_{s=1}^{T} Y_{is} = 1 | \boldsymbol{x})} \\ &= \frac{P(W_i = w_i | \boldsymbol{x}, T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1)\, P(T_i = t_i, \sum_{s=1}^{T} Y_{is} = 1 | \boldsymbol{x})}{\sum_{t \in W_i} P(W_i = w_i | \boldsymbol{x}, T_i = t, \sum_{s=1}^{T} Y_{is} = 1)\, P(T_i = t, \sum_{s=1}^{T} Y_{is} = 1 | \boldsymbol{x})} \\ &= \begin{cases} \frac{\pi(W_i|t_i)\lambda_i \exp(\boldsymbol{x}_{t_i}\boldsymbol{\beta})}{\pi(W_i|t_i)\lambda_i \exp(\boldsymbol{x}_{t_i}\boldsymbol{\beta}) + \pi(W_i|t_i - \delta)\lambda_i \exp(\boldsymbol{x}_{t_i - \delta}\boldsymbol{\beta})} & , W_i = \{t_i - \delta, t_i\} \\ \frac{\pi(W_i|t_i)\lambda_i \exp(\boldsymbol{x}_{t_i}\boldsymbol{\beta})}{\pi(W_i|t_i)\lambda_i \exp(\boldsymbol{x}_{t_i}\boldsymbol{\beta}) + \pi(W_i|t_i + \delta)\lambda_i \exp(\boldsymbol{x}_{t_i - \delta}\boldsymbol{\beta})} & , W_i = \{t_i, t_i + \delta\}, \end{cases} \end{aligned} \tag{2.22}$$

where $\pi(W_i|t) = P(W_i = w_i|T_i = t_i, \boldsymbol{x}, \sum Y_{is} = 1)$. In a semi-symmetric bi-directional design $\pi(W_i|t_j) = 0.5$ if t_j is in the middle of the exposure series, which is to say that both before and after referent periods are available to be selected. If $\pi(W_i|t_j) = \pi(W_i|t_k)\ \forall\ (t_j, t_k) \in W_i$, then the likelihood in (2.22) reduces to that of the conditional logistic likelihood. However, subjects who have index times at the beginning or end of the exposure time series have only one referent period available for choosing. For example, a subject that has an event at the start of the study will not have a referent period from the past available, as the exposures for that time were not recorded. This results in overlap bias unless the likelihood in (2.22) is modified. Specifically, if an offset of ln(2) is added for days at the beginning and end of the exposure series, (2.22) reduces to the conditional logistic likelihood as a result of the π's canceling out (since $0.5 * \exp(\ln(2) + \boldsymbol{x}_t\boldsymbol{\beta}) = \exp(\boldsymbol{x}_t\boldsymbol{\beta})$).

Using the above results, Janes et al. [2005a] deduced that this referent selection scheme, called the adjusted semi-symmetric bi-directional design, is both localizable and ignorable, and can be used to select referents for a given case index. No overlap bias will occur in this design as the estimating equations of the conditional logistic likelihood will be used to estimate parameters. Alternatively, if referent times are randomly chosen for cases at the beginning and end of the exposure time series, and those referent times are outside the exposure series, those cases can be dropped from the analysis. This is equivalent to weighting cases at the beginning or end of the exposure series as 0.5, which means the above will again reduce to the conditional logistic likelihood.

For the remainder of the thesis the adjusted semi-symmetric bi-directional referent design will be utilized to create the case-crossover dataset that will be used in applied analyses. Referent times that are available to be chosen for a specific index day will be multiples of 7, which will control for day of week confounding (Bateson and Schwarts [1999], Bateson and Schwarts [2001]). Specifically, using a lagged referent time of 14 days prior to the index day and a lead referent time of 14 days past the event day will control for seasonal trend (Levy and Lumley [2000]). Since these confounders are controlled for by design, they do not need to be included in the model. To illustrate, if the case index time is set to be a day before the event date (a lag of 1), then the past referent time will be 15 days before the event time (14 days before the lag 1 time) and the future referent time will be 13 days ahead of the event time (14 days ahead of the event time). A similar approach will be used when computing moving averages. If the day of the event and the previous 6 days are used to create an average exposure for the case, then the control moving averages are defined to be the average of the 14th-20th day exposures in the past or the average of 8th-14th day exposures in the future. The symmetry about the day(s) used for the case index is maintained in all cases. Note that the previous derivations can be easily extended to include $1 : M$ matching per event. For example, given an index day, the referent times can be picked to be the 14th and 21st day prior or the 14th and 21st day after the index day. This will result in $1 : 2$ matching.

It is worth noting that a second type of localizable and ignorable referent selection strategy is called the time-stratified design. This method of picking referent times divides time into disjoint strata (for example stratifying a year into disjoint sets of 28 days). The index day is then used to determine which strata to pick, and all of the same days of the week as the index day that belong to the strata are chosen as referent times (or a random sample of these days is chosen). This, like the semi-symmetric method mentioned earlier, will control for day of week and season confounding.

Here, the time-stratified referent design having score equation with mean 0 is shown. First, note that the expectation of the estimating equation for a single subject (with respect to t_i), regardless of the referent selection is:

$$\mathrm{E}_{t_i}(U_i(\boldsymbol{\beta})) = \sum_{t_i=1}^{T} \frac{\exp\{\boldsymbol{x}_{t_i}\boldsymbol{\beta}\}}{\sum\limits_{s=1}^{T}\exp\{\boldsymbol{x}_s\boldsymbol{\beta}\}} (\boldsymbol{x}_{t_i} - \sum_{u\in W_i} \boldsymbol{x}_u \frac{\exp\{\boldsymbol{x}_u\boldsymbol{\beta}\}}{\sum\limits_{v\in W_i}\exp\{\boldsymbol{x}_v\boldsymbol{\beta}\}}) \tag{2.23}$$

since $P(T = t|\boldsymbol{x}, \sum_{s=1}^{T} Y_{is} = 1) = \frac{\exp\{\boldsymbol{x}_t\boldsymbol{\beta}\}}{\sum_{s=1}^{T}\exp\{\boldsymbol{x}_s\boldsymbol{\beta}\}}$. Setting $t = t_i$ and $W_i = W_t$, (2.23) factors into:

$$\mathrm{E}_{t_i}(U_i(\boldsymbol{\beta})) = \sum_{t_i=1}^{T} v_t(\boldsymbol{x}_t - \overline{\boldsymbol{x}(W_t)}) \tag{2.24}$$

where $v_t = \frac{\exp\{\boldsymbol{x}_t\boldsymbol{\beta}\}}{\sum_{s=1}^{T}\exp\{\boldsymbol{x}_s\boldsymbol{\beta}\}}$, W_t is the referent window to which t belongs to, and

$$\overline{x(W_t)} = \sum_{u \in W_t} \frac{\boldsymbol{x}_u * \exp\{\boldsymbol{x}_u\boldsymbol{\beta}\}}{\sum_{v \in W_t} \exp\{\boldsymbol{x}_v\boldsymbol{\beta}\}}.$$

In a time-stratified design, the sum over all times can rewritten as a sum over strata $s = 1, 2, ..., S$ and time within strata, since time is partitioned into a complete set of disjoint strata. Then (2.24) becomes

$$\begin{aligned}
\mathrm{E}_{t_i}(U_i(\boldsymbol{\beta})) &= \sum_{s=1}^{S} \frac{1}{\sum_{u=1}^{T}\exp\{\boldsymbol{x}_u\boldsymbol{\beta}\}}(\sum_{t \in s} \boldsymbol{x}_t \exp\{\boldsymbol{x}_t\boldsymbol{\beta}\} - \sum_{u \in s} \boldsymbol{x}_u \exp\{\boldsymbol{x}_u\boldsymbol{\beta}\} \frac{\exp\{\boldsymbol{x}_u\boldsymbol{\beta}\}}{\sum_{v \in W_i}\exp\{\boldsymbol{x}_v\boldsymbol{\beta}\}}) \\
&= \mathbf{0}
\end{aligned}$$

Thus the time-stratified design yields a score equation with mean 0, and therefore is a localizable and ignorable referent selection strategy as well.

2.4 Review: Conditional Logistic Likelihood Parameter Estimation via Cox Proportional Hazard Partial Likelihood Maximization

Here a brief review of the Cox proportional hazards partial likelihood from survival analysis is provided as it will later be shown to have a connection to the conditional logistic likelihood function, and will play a key role in the research presented in Chapters 3 and 4. Begin by assuming time, T, is a continuous, positive random variable denoting survival time. In the context of a survival study, subjects experiencing events during the study have their event times recorded. Subjects that do not experience an event during the study, nor do they dropout of the study, have what is termed a right-censored event time. What is known about a censored subject's event time is that it is beyond their censoring time, but the subject is considered at risk throughout the observed follow-up.

First consider functionals of the survival distribution that are often of scientific interest. Define the probability density function $f(t)$, cumulative distribution function $F(t)$, survival function $S(t)$, hazard function $\lambda(t)$ and cumulative hazard function $\Lambda(t)$ as:

$$f(t) = \lim_{\triangle t \to 0^+} \frac{1}{\triangle t} P[t \le T < t + \triangle t]$$

$$F(t) = P[T \le t]$$

$$S(t) = P[T > t] = 1 - F(t) = 1 - \int\limits_0^t f(s)ds$$

$$\lambda(t) = \lim_{\triangle t \to 0^+} \frac{1}{\triangle t} P[t \le T < t + \triangle t | T \ge t] = f(t)/S(t)$$

$$\Lambda(t) = \int\limits_0^t \lambda(s)ds$$

The hazard at time t is as the instantaneous rate of an event at time t given no event up to time t. In a proportional hazards model, the effect of an increase in a covariate is multiplicative with respect to the hazard rate, where the multiplicative effect remains constant at all times. Cox [1972] developed a model used for survival data based on the proportional hazards assumption. Specifically, the Cox proportional hazards model is given by

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\text{exp}(\boldsymbol{x\beta}) = \lambda_0(t)\text{exp}\{\beta_1 x_{i1} + ... + \beta_p x_{ip}\},$$

where $\boldsymbol{x}$ is a $1 \times p$ vector of covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters and $\lambda_0(t)$, which need not be specified, represents the baseline hazard function at time t. That is to say $\lambda_0(t)$ is the hazard at time t for a subpopulation with all covariate values equal to 0, $\boldsymbol{x} = \mathbf{0}$. β_k denotes the log relative risk comparing $x_{ik} + 1$ to x_{ik} for $k = 1, 2, .., p$, assuming all other x's are held constant. Since $\lambda_0(t)$ is not specified, $\boldsymbol{\beta}$ cannot be estimated using a standard parametric likelihood, as that would require full specification of the underlying probability model. To avoid modeling the baseline hazard function, Cox [1975] derived what is termed the partial likelihood, which can be used to obtain estimates for $\boldsymbol{\beta}$.

To construct the partial likelihood, the notation of Fleming and Harrington [1991] is used. Begin by assuming that for a sample size of n, $\{t_{(1)}, ..., t_{(K)}\}$ denote the ordered failure times with no ties, $k = 1, \ldots, K$. Further, let $(A_1, B_1), (A_2, B_2), ..., (A_K, B_K)$ be a collection of $2K$ events, where A_k be the event specifying the labels that failed at time $t_{(k)}$ and B_k be the event describing the observed times of censoring in the interval $[t_{(k-1)}, t_{(k)})$, along with the labels associated with the censoring times, given the fact that an observation failed at time $t_{(k)}$, $k = 1, \ldots, K$. Then the likelihood for these $2K$ events is

$$\begin{aligned} P(A_1, B_1, ..., A_K, B_K) &= \left[\prod_{k=2}^{K} P(A_k, B_k | A_{k-1}, B_{k-1}, ..., A_1, B_1)\right] P(A_1, B_1) \\ &= \left[\prod_{k=2}^{K} P(A_k | B_k, A_{k-1}, B_{k-1}, ..., A_1, B_1)\right] P(A_1 | B_1) \\ &\quad \times \left[\prod_{k=2}^{K} P(B_k | A_{k-1}, B_{k-1}, ..., A_1, B_1)\right] P(B_1) \end{aligned} \tag{2.25}$$

All four terms in the right hand side of (2.25) depend on parameters defining the survival distribution of interest. Thus taking only the first two terms out of the four constitutes a partial likelihood. Let $R(t)$ denote the labels of the observations still at risk at time just prior to t, t^-. In many scenarios it is a reasonable assumption that the censoring times are independent of the actual failure times, and hence do not provide additional information about the failure distribution that would not be available if censoring were not present. Thus, it is reasonable to assume that the events B_k contain little information about $\boldsymbol{\beta}$. This is the initial reasoning on neglecting the last 2 terms out of the 4 in (2.25) when constructing the partial likelihood.

The partial likelihood is constructed by comparing the risk of the subjects that experienced an event, given their covariate values, to the risk of subjects still at risk for the event, given their covariate values. More specifically, for the subject failing at time $t_{(k)}$, the likelihood contribution is

$$\begin{aligned}
L_{(k)}(\boldsymbol{\beta}) &= P(A_k|B_k, A_{k-1}, B_{k-1}, ..., A_1, B_1) \\
&= P[\text{subject with } \boldsymbol{x}_{(k)} \text{ fails at } t_{(k)}|\text{some subject failed at } t_{(k)}] \\
&= \frac{P[\text{subject with } \boldsymbol{x}_{(k)} \text{ fails at } t_{(k)}]}{P[\text{some subject fails at } t_{(k)}]} \\
&= \frac{\left[\lambda_k(t_{(k)})(\triangle t) \prod_{j \in R(t_{(k)})-(k)} \{1 - \lambda_j(t_{(k)})(\triangle t)\}\right]}{\left[\sum_{l \in R(t_{(k)})-l} \lambda_l(t_{(k)})(\triangle t) \prod_{j \in R(t_{(k)})-l} \{1 - \lambda_j(t_{(k)})(\triangle t)\}\right]} \\
&= \frac{\left[\lambda_k(t_{(k)})(\triangle t)\right]}{\left[\sum_{l \in R(t_{(k)})} \lambda_l(t_{(k)})(\triangle t)\{1 - \lambda_k(t_{(k)})(\triangle t)\}/\{1 - \lambda_l(t_{(k)})(\triangle t)\}\right]} \qquad (2.26)
\end{aligned}$$

Now, since $\triangle t$ is small, $\frac{1-\lambda_k(t_{(k)})(\triangle t)}{1-\lambda_l(t_{(k)})(\triangle t)} \approx 1$, and (2.26) factors as

$$\begin{aligned}
L_{(k)}(\boldsymbol{\beta}) &= \frac{\lambda_k(t_{(k)})(\triangle t)}{\sum_{l \in R(t_{(k)})} \lambda_l(t_{(k)})(\triangle t)} \\
&= \frac{\lambda_k(t_{(k)})}{\sum_{l \in R(t_{(k)})} \lambda_l(t_{(k)})} \\
&= \frac{\lambda_0(t_{(k)})\exp(\boldsymbol{x}_{(k)}\boldsymbol{\beta})}{\sum_{l \in R(t_{(k)})} \lambda_0(t_{(k)})\exp(\boldsymbol{x}_l\boldsymbol{\beta})} \\
&= \frac{\exp(\boldsymbol{x}_{(k)}\boldsymbol{\beta})}{\sum_{l \in R(t_{(k)})} \exp(\boldsymbol{x}_l\boldsymbol{\beta})} \qquad (2.27)
\end{aligned}$$

Let δ_i be an event indicator for subject i, such that $\delta_i = 1$ if the true event time was observed for subject i and 0 otherwise, $i = 1, \ldots, n$. Then using (2.27), the partial likelihood across all observed failure time becomes

$$L_P(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\exp(\boldsymbol{x}_{(k)}\boldsymbol{\beta})}{\sum_{l \in R(t_{(k)})} \exp(\boldsymbol{x}_l\boldsymbol{\beta})} = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{l \in R(t_i)} \exp(\boldsymbol{x}_i\boldsymbol{\beta})} \right]^{\delta_i}.$$

If the data are stratified into $G > 1$ independent strata, and if L_g represents the likelihood contribution corresponding to strata g, then the partial likelihood across all strata is given by

$$L_P(\boldsymbol{\beta}) = \prod_{g=1}^{G} L_g = \prod_{g=1}^{G}\prod_{k=1}^{K} \frac{\exp(\boldsymbol{x}_{g(k)}\boldsymbol{\beta})}{\sum_{l \in R(t_{(gk)})} \exp(\boldsymbol{x}_{gl}\boldsymbol{\beta})}$$

The log likelihood, $l(\boldsymbol{\beta})$, and score equation, $U_j(\boldsymbol{\beta})$, for the Cox proportional hazard likelihood are as follows:

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{K} \boldsymbol{x}_{(k)}\boldsymbol{\beta} - \log\left(\sum_{i \in R_{(k)}} \exp(\boldsymbol{x}_i\boldsymbol{\beta}) \right)$$

$$U_j(\boldsymbol{\beta}) = \sum_{k=1}^{K} \boldsymbol{x}_{(k)j} - \bar{\boldsymbol{x}}_{(k)j}$$

for $j = 1, 2, ..., p$, where K is the total number of distinct observed event times, $\bar{\boldsymbol{x}}_{(k)j} = \sum_{i \in R_{(k)}} \boldsymbol{x}_{ij} w_{(k)i}(\boldsymbol{\beta})$ and $w_{(k)i}(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{i \in R_{(k)}} \exp(\boldsymbol{x}_i\boldsymbol{\beta})}$.

Additionally, the observed information matrix, $I(\boldsymbol{\beta})$, is defined by elements

$$I_{jh}(\boldsymbol{\beta}) = I_{hj}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left[\sum_{i \in R_{(k)}} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_{(k)j})(\boldsymbol{x}_{ih} - \bar{\boldsymbol{x}}_{(k)h}) w_{(k)i}(\boldsymbol{\beta}) \right], \; j, h = 1, 2, ..., p.$$

Parameter estimates for $\boldsymbol{\beta}$ are obtained by setting the score equations to 0 and solving. Cox [1975] noted that since the sets B_k provide little information about $\boldsymbol{\beta}$, that the partial likelihood can be maximized to obtain reasonably efficient estimates for $\boldsymbol{\beta}$. Efron [1977] showed that the maximum partial likelihood estimator is asymptotically locally efficient under mild conditions when the proportional hazards assumption holds. Presenting survival data in a counting process setting and borrowing results from martingale theory, Andersen and Gill [1982] derived the asymptotic distribution of the maximum partial likelihood, namely that

$$\boldsymbol{\beta} \dot{\sim} N(\boldsymbol{\beta}, I_{\boldsymbol{\beta}}^{-1}),$$

under regularity conditions analogous to those assumed under standard likelihood theory.

The review on time-to-event analysis via the Cox proportional hazards partial model is presented in order to establish a key relationship between Cox's partial likelihood and the conditional logistic likelihood. More specifically, Breslow and Day [1980] noted an equivalence between the conditional logistic likelihood and the Cox's partial likelihood. This derivation is recreated below.

For ease of notation, assume 1 : 1 matching and a single event in each matched set. From a conditional logistic likelihood viewpoint, subject i will have $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$, $i = 1, \ldots, n$. Without loss of generality, assume the first observation, Y_{i1} denotes the event, and the second, Y_{i2} is the matched control. It was shown in (2.19) that the likelihood contribution for this subject is

$$\begin{aligned} L_i &= P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \, S_i, \sum_j y_{ij} = 1) \\ &= \frac{\exp(\boldsymbol{x}_{i1}\boldsymbol{\beta})}{\exp(\boldsymbol{x}_{i1}\boldsymbol{\beta}) + \exp(\boldsymbol{x}_{i2}\boldsymbol{\beta})} \end{aligned} \tag{2.28}$$

Now, from a time to event analysis viewpoint, assume the data are stratified, and each stratification contains a single matched pair. Thus, each strata has only a single event. For a specific stratum (i) the likelihood contribution was shown in (2.27) to be:

$$L_{(i)} = \frac{\exp(\boldsymbol{x_{(1)i}\beta})}{\sum\limits_{j \in R_i} \exp(\boldsymbol{x_{ji}\beta})} \tag{2.29}$$

Given the conditional logistic likelihood data, if the event time within each subject is created and set to a constant, and control times are created and set to a number higher than the event time, (2.28) and (2.29) will be equivalent. Therefore taking the product of the individual subject/strata likelihood contributions across all i subjects will maintain equality and thus the conditional logistic likelihood is mathematically equivalent to Cox's partial likelihood. Because of this equivalence, the most common statistical software packages used to analyze matched case-control data via the conditional logistic model proceed to do so by transforming the data to an equivalent Cox partial likelihood maximization procedure. The more general case involving $1 : M$ matching and numerous events within subject is discussed in Chapter 3.

2.5 Review: Bayesian Analysis of Case-Control Studies

Exposure\Disease	$D = 1$	$D = 0$	Total
$X = 1$	a	b	t
$X = 0$	c	d	$N - t$
Total	N_1	N_0	N

Table 2.5: 2-way table with a single binary outcome and single binary exposure

The earliest Bayesian inference on case-control studies began with Zelen and Parker [1986], Nurminen and Mutanen [1987], and Marshall [1988]. Each of these authors considered a Bayesian model formulation of a case-control model with a single binary exposure, X. To summarize the model, let p_1 denote the probability of exposure in the control population and let p_2 denote the probability of exposure in the case population. Recall the two-way table as presented in Table 2.5. Since the individual cell counts given the column totals are binomially distributed, the likelihood is proportional to

$$L(p_1, p_2) \propto p_1^b (1-p_1)^d \, p_2^a (1-p_2)^c \tag{2.30}$$

One approach is to specify independent conjugate priors for p_i, $i = 1, 2$, where $p_1 \sim \text{Beta}(\mu_1, \mu_2)$ and $p_2 \sim \text{Beta}(\nu_1, v_2)$, where the parameter of scientific interest is again the log odds ratio, namely $\beta = \log\left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right)$. After re-parameterization and a change of variables via a Jacobian transformation, a prior on β is induced and the posterior of β based on this prior and the likelihood in (2.30) is given by

$$p(\beta|a, b, c, d) \propto \exp([a + \nu_1]) \int_0^1 \frac{p_1^{a+c+\nu_1+\mu_2-1}(1-p_1)^{b+d+\nu_2+\mu_2-1}}{(1-p_1+p_1\exp(\beta))^{a+b+\nu_1+\nu_2}} dp_1 \tag{2.31}$$

The posterior in (2.31) is not a closed form density and numerical methods can be applied to evaluate the integral, and thus sample from the posterior distribution of β.

Müller and Roeder [1997] investigated Bayesian modeling of case-control studies by considering continuous exposures with measurement error. This scenario expanded on the single binary exposure approaches. Let D denote the event outcome (0 or 1), X denote the exposure of interest with possible measurement error, and a completely observed covariate Z. Error in variables occurs when for a subset of the data (referred to as reduced data or R) a proxy W is measured instead of the true covariate X. For the complete data (C), both the true X and W is recorded. For the reduced data let X_R denote the missing exposure and let X_C denote the observed exposures for the complete data. The retrospective sample is chosen as follows: $n_1 = n_{1R} + n_{1C}$ cases are sampled and $n_0 = n_{0R} + n_{0C}$ controls are sampled.

For the prospective probability of event, a logistic link is assumed i.e. $P(D = 1|X, Z, W) = \frac{\exp(\beta_0+\beta_1 T_1(X)+\beta_2 T_2(Z))}{1+\exp(\beta_0+\beta_1 T_1(X)+\beta_2 T_2(Z))}$, where $T_1(.)$ and $T_2(.)$ are monotonic transformations. Let $\boldsymbol{\beta}$ denote the vector of parameters in the logistic link (the log odds ratio parameters) and let $\boldsymbol{\theta}$ denote the parameters for the marginal distribution of the exposure X and covariates Z and W. Assume D and W are conditionally independent given X ,which is to assume non-differential measurement error (Carroll 1993):

$$P(D|X,Z,W,\beta) = P(D|X,Z,\beta).$$

Choosing a prior, $p(\boldsymbol{\beta,\theta})$, for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the joint posterior is of the form

$$p(\boldsymbol{\beta,\theta}|X_C,W,D,Z) \propto p(\boldsymbol{\beta,\theta})\prod_{i\in C}p(X_i,Z_i,W_i|D_i,\boldsymbol{\beta,\theta})\prod_{i\in R}\left[\int p(X_i,Z_i,W_i|D_i.\boldsymbol{\beta},\boldsymbol{\theta})dX_i\right].$$

Augmenting the parameters with the latent vector X_R, the joint posterior of the parameters and the reduced data exposures is given by

$$p(\beta,\boldsymbol{\theta},X_R|X_C,W,D,Z) \propto p(\beta,\boldsymbol{\theta})\prod_{i=1}^{n}p(X_i,Z_i,W_i|D_i.\beta,\boldsymbol{\theta})$$

Invoking the assumption of non-differential measurement error, the joint posterior of the parameters and reduced data exposures becomes

$$p(\boldsymbol{\beta},\boldsymbol{\theta},X_R|X_C,W,D,Z) \propto p(\boldsymbol{\beta},\boldsymbol{\theta})\prod_{i=1}^{n}p(X_i,Z_i,W_i|\boldsymbol{\theta})p(D_i|X_i,Z_i,\boldsymbol{\beta})/p(D_i|\boldsymbol{\beta},\boldsymbol{\theta})$$

where $P(D_i|\beta,\boldsymbol{\theta}) = \int P(D_i|X_i,Z_i,W_i,\boldsymbol{\beta})dP(X_i,Z_i,W_i|\boldsymbol{\theta})$ since the distribution of D_i conditional on X_i,Z_i,W_i does not depend on $\boldsymbol{\theta}$, and the distribution of X_i,Z_i,W_i only depends on $\boldsymbol{\theta}$.

In the work of Müller and Roeder [1997], the joint distribution of (X_i,Z_i,W_i) assumes a mixture model with a multivariate normal kernel $\phi_{\boldsymbol{\theta}_i}$ using a Dirichlet process mixture model such that

$$p(X_i,Z_i,W_i|\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i,\Sigma_i)) \sim N(\boldsymbol{\mu}_i,\Sigma_i)$$

$$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)|G \sim G$$

$$G \sim DP(\alpha, G_0)$$

where α is the concentration parameter and G_0 is the base measure (a more in-depth review of the Dirichlet process and its properties follows in the next section). This non-parametric approach models the joint distribution of (X_R, Z, W) for the reduced data and (X_C, Z, W) for the complete data. Using a mixture of normal models with a Dirchlet process prior on the mixing measure (Ferguson [1973], Escobar [1994], Escobar and West [1995]), a class of flexible mixture distributions is obtained for the joint distribution of the exposure and the covariates.

Müller and Roeder [1997] complete their Bayesian semi-parametric hierarchical model by setting $G_0 \equiv N(\boldsymbol{\mu}_0, \Sigma_0)$, specifying a gamma prior for α, and specifying a diffuse prior on $\boldsymbol{\beta}$. This approach assumes a mixture of multivariate normal models for the distribution of (X, Z, W), with a Dirichlet process prior model on the unknown mixture measure. Note that since X_R is latent, it is sampled according to $p(X_R|\boldsymbol{\beta}, \boldsymbol{\theta}, W, D, Z, X_C)$. Finally, note that the above model setup can easily be extended to multivariate covariates, $\boldsymbol{X}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$.

Markov chain Monte Carlo methods are implemented for sampling from the posterior distribution of the model parameters. When the the dimension of the space of $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z})$ increases, computation becomes intensive since $p(D_i|\boldsymbol{\beta}, \boldsymbol{\theta})$ is numerically integrated over $(\boldsymbol{X, Z, W})$.

The work by Müller and Roeder [1997] introduced the idea of incorporating continuous exposures and flexible nonparametric modeling of the exposure distribution for the Bayesian analysis of case-control data. In this formulation, categorical exposures required a different treatment as the normal kernel of the Dirichlet process implicitly assumed continuous exposures.

In related work, Seaman and Richardson [2001] replaced the binomial likelihood of the two-way table with a multinomial likelihood, thus extending the binary exposure model of Zelen and Parker [1986]. In this work, the set of multinomial probabilities that would correspond to the exposure categories in case and control populations was assumed to have a discrete Dirichlet prior. A diffuse Dirichlet(0,0,...,0) prior for the exposure probabilities is assumed, which then implies an induced improper uniform prior on β. A continuous exposure can also be used, but only by discretizing them into groups.

2.5.1 Bayesian Analysis of Matched Case-Control Studies

The first Bayesian approach for matched case-control studies was proposed by Diggle et al. [2000], where a nuisance parameter to represent the separate effects of matching in each matched set is introduced. Of consideration in this study is an exposure of interest that is defined by the spatial location of an individual relative to a point or line source of pollution.

For an unmatched design, the set up is as follows. Let $g(x)$ denote the intensity function of a non-homogenous Poisson process that defines the locations of individuals in the population at risk. Given a subject's location x , let $p^*(x)$ be the probability of becoming a case. The odds of disease amongst sampled subjects is

$$r(x) = \frac{a}{b}\frac{p^*(x)}{1 - p^*(x)}$$

where a and b are the sampled proportions of cases and controls. Commonly, $r(x) \equiv r(x, \theta)$, denoting that the odds, involves an unknown parameter θ. This is modeled as $r(x, \theta) = \rho h(x, \theta)$ where $h(x, \theta)$ involves the interpretable parameter of interest θ.

Extending the above setup to J matched case-control pairs proceeds as follows. Let x_{0j} and

x_{1j} denote the J locations for the cases and controls respectively, $j = 1, 2, ..., J$. Given a subject at location x in stratum j, the probability of becoming a case is given by

$$p_j(x, \theta) = \frac{r_j(x, \theta)}{1 + r_j(x, \theta)} = \frac{\rho_j h(x, \theta)}{1 + \rho_j h(x, \theta)},$$

where the ρ_j's represent the baseline odds that vary among the matched pairs, and are considered nuisance parameters. The probability of disease for a subject at distance x in stratum j conditional on the unordered set of exposures within that matched pair is:

$$p_c(x_{j0}, \theta) = \frac{h(x_{jo}, \theta)}{h(x_{jo}, \theta) + h(x_{j1}, \theta)}.$$

In the case of $1 : M$ matching and q additional covariates, $z_k(x_{ji})$ $(k = 1, 2, ..., q)$, measured for subject i at the j-th stratum $(i = 1, 2, ..., M+1; j = 1, 2, ..., J)$, the conditional probability is of the form

$$p_c(x_{j0}, \theta, \boldsymbol{\phi}) = \frac{h(x_{jo}.\theta)\text{exp}(\sum_{k=1}^{q} z_k(x_{j0})\phi_k)}{\sum_{i=1}^{M+1} h(x_{jo}.\theta)\text{exp}(\sum_{k=1}^{q} z_k(x_{j0})\phi_k)}, \tag{2.32}$$

where $\boldsymbol{\phi} = (\phi_1, ..., \phi_q)$. The conditional likelihood based on (**??**) is then given by

$$L(\theta, \boldsymbol{\phi}) = \sum_{j=1}^{J} \log(p_c(x_{j0}, \theta, \boldsymbol{\phi})).$$

In the type of spatial studies that motivated the work of Diggle et al. [2000], there is a point source of interest at location x^*, and interest is on how risk changes with locations in relation to x^* . Letting $d = ||x - x^*||$ be the distance to the source from location x. Diggle [1990] suggests h be of the form:

$$h(x) = 1 + \alpha\text{exp}(-(\frac{d}{\beta})^2) \text{ for fixed } x^*$$

The parameter α represents the proportional increase in disease odds at the source and β measures the rate of decay with increasing distance from the source in units of distance.

Diggle [1990] conduct the Bayesian analysis by putting independent priors on ϕ_k's, $\phi_k \overset{iid}{\sim} N(\mu, \sigma^2)$ and uniform priors on α and β. Posterior draws are obtained via Markov chain Monte Carlo sampling that incorporated a component wise Metropolis-Hastings algorithm. This approach represented an alternative to likelihood methods as the likelihood in this model is highly irregular.

More recently Ghosh and Chen [2002] developed a Bayesian technique for matched case-control studies with one or more binary exposures. In this model, they work with the unconditional likelihood. As a result, the matched set specific parameters γ_i are in the likelihood, as well as the parameter of interest β. In their approach, they assume independent priors for each of the γ_i and the β. Mukherjee et al. [2007] apply a semi-parametric approach to case-control data in a study of gene-environment association with disease status. Similar to Müller and Roeder, the covariate distribution is assumed to come from a mixture of normals, where the mixing proportions are specified a Dirichlet process prior. The most recent work on matched case-control studies via Bayesian modeling will be discussed in chapter 5 (Sinha et al. [2004] ,Sinha et al. [2005])

2.5.2 Equivalence of Retrospective and Prospective Analysis in a Bayesian Framework

Seaman and Richardson [2004] showed that the posterior for $\boldsymbol{\beta}$ obtained using a retrospective likelihood is the same as the posterior obtained using a prospective likelihood. This is the Bayesian analogue of the results of Prentice and Pyke [1979]. Seaman and Richardson [2004] first revisit the multinomial-Poisson transformation presented by Baker [1994].

Let $\boldsymbol{X}$ be a discrete exposure with J support points $\boldsymbol{z}_1, ..., \boldsymbol{z}_J$. Let n_{0j} and n_{1j} denote the number of cases and controls with values $\boldsymbol{X} = \boldsymbol{z}_j$ respectively, for $j = 1, ..., J$. If

$P(\boldsymbol{X} = \boldsymbol{z}_j|D = 0) = \frac{\theta_j}{\sum_j \theta_j}$ and odds of disease for $\boldsymbol{X} = \boldsymbol{x}$ is $\exp(\boldsymbol{x\beta})$ then the natural retrospective likelihood is:

$$L_{MR} = \prod_{d=0}^{1}\prod_{j=1}^{J}\left[\frac{\theta_j \exp(d\boldsymbol{z_j\beta})}{\sum_{j=1}^{J}\theta_j \exp(d\boldsymbol{z_j\beta})}\right]^{n_{dj}}$$

In a cohort study, the natural prospective likelihood is:

$$L_{MP} = \prod_{j=1}^{J}\prod_{d=0}^{1}\left[\frac{\alpha^d \exp(d\boldsymbol{z_j\beta})}{\sum_{k=0}^{1}\alpha^k \exp(d\boldsymbol{z_j\beta})}\right]^{n_{dj}} .$$

The parameter α represents the odds of disease when exposure is $\mathbf{0}$, i.e. the baseline odds. Now let Y_{dj} $(d = 0, 1;\ j = 1, ..., J)$ be independently distributed as $Y_{dj} \sim \text{Poisson}(\lambda_{dj})$ where $\log(\lambda_{dj}) = \log(\mu) + d\log(\alpha) + \log(\theta_j) + d\boldsymbol{z}_j\boldsymbol{\beta}$ with $\theta_1 = 1$ for identifiability. The likelihood for $(\mu, \alpha, \boldsymbol{\beta, \theta})$ where $\boldsymbol{\theta} = (\theta_1, ..., \theta_J)$ is then given by

$$L_{\text{po}}(\mu, \alpha, \boldsymbol{\beta, \theta; y}) = \prod_{d=0}^{1}\prod_{j=1}^{J}(\lambda_{dj})^{y_{dj}}\exp(-\lambda_{dj})$$

The profile likelihood of a parameter $\boldsymbol{\beta}$ based on a likelihood involving two parameters $(\alpha, \boldsymbol{\beta})$,$L(\boldsymbol{\beta}, \alpha)$, is $L_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = L(\boldsymbol{\beta}|\hat{\alpha})$ where $\hat{\alpha} = \arg\max_{\alpha} L(\alpha, \boldsymbol{\beta}|\boldsymbol{\beta})$. Baker [1994] points out that L_{MR} is the profile likelihood for $(\boldsymbol{\beta, \theta})$ after maximizing L_{po} with respect to (μ, α) and similarly the profile likelihood for $(\boldsymbol{\beta}, \alpha)$ after maximizing L_{po} with respect to $(\mu, \boldsymbol{\theta})$ is L_{MP}. Since the order of maximization is arbitrary, it follows that the profile likelihood for $\boldsymbol{\beta}$ after maximizing over the nuisance parameters is equivalent regardless if the starting likelihood is L_{po}, L_M, or L_{MR}. This is an analogous result of derivation of Prentice and Pyke [1979], which showed that under this construction, one can start with either a prospective likelihood or retrospective likelihood to obtain equivalent estimates for $\boldsymbol{\beta}$. From the perspective of maximization, the L_{MP} model is easier to fit as it only involves a single nuisance parameter, α. On the other hand the L_{MR} model contains $J > 1$ many nuisance parameters.

The Bayesian analogue of this involves integration as opposed to maximization. Moving to a Bayesian framework with the same initial data generating setup, specify independent improper priors for α and $\boldsymbol{\theta}$, $p(\alpha) \propto \alpha^{-1}$ and $p(\theta_j) \propto \theta_j^{a_j - 1}$. Specify a prior on $\boldsymbol{\beta}$, $p(\boldsymbol{\beta})$ such that $\boldsymbol{\beta}$ is independent of α and $\boldsymbol{\theta}$, and that for some q and r such that $y_{0q} \geq 1$ and $y_{0r} \geq 1$, $\mathrm{E}(\boldsymbol{z_q \beta})$ and $\mathrm{E}(\boldsymbol{z_r \beta})$ exist and are finite.

Let $y_{+j} = y_{0j} + y_{1j}$ and $y_{d+} = \sum\limits_{j=1}^{J} y_{dj}$. Then the following statements hold:

1. Letting $\omega = \log(\alpha)$ the posterior of $(\omega, \boldsymbol{\beta})$ is

$$p(\omega, \boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{\beta}) \prod_{j=1}^{J} \frac{[\exp(\omega + \boldsymbol{z}_j \boldsymbol{\beta})]^{y_{1j}}}{[1 + \exp(\boldsymbol{z}_j \boldsymbol{\beta})]^{y_{+j} + a_j}} \tag{2.33}$$

2. The posterior of $(\boldsymbol{\delta}, \boldsymbol{\beta})$ where $\boldsymbol{\delta} = (\delta_1, ..., \delta_J)$ and $\delta_j = \theta_j / \sum\limits_{j=1}^{J} \theta_j$ is

$$p(\boldsymbol{\delta}, \boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{\beta}) \prod_{j=1}^{J} \delta_j^{a_j - 1} \prod_{d=0}^{1} \frac{\prod\limits_j [\delta_j \exp(d \boldsymbol{z}_j \boldsymbol{\beta})]^{y_{dj}}}{[\sum\limits_j \delta_j \exp(\boldsymbol{z}_j \boldsymbol{\beta})]^{y_{d+}}} \tag{2.34}$$

To show (1.), begin with the posterior of $(\alpha, \boldsymbol{\theta}, \boldsymbol{\beta})$ given by

$$p(\alpha, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{\beta}) \frac{1}{\alpha} \prod_{j=1}^{J} \theta_j^{a_j - 1} \prod_{d=0}^{1} \prod_{j=1}^{J} (\lambda_{dj})^{y_{dj}} \exp(-\lambda_{dj}) \tag{2.35}$$

Integrating (2.35) with respect to $\boldsymbol{\theta}$ and then conducting a transformation of variable from α to ω yields (2.33).

To show (2.), begin with (2.35) and transform $\boldsymbol{\theta}$ to $(\boldsymbol{\delta}, \psi)$ where $\psi = \sum\limits_{j=1}^{J} \theta_j$. Then integrate out α followed by ψ . The Jacobian going from $\boldsymbol{\theta}$ to $(\boldsymbol{\delta}, \psi)$ is $|\frac{\partial(\theta_1, ..., \theta_J)}{\partial(\delta_1, ..., \delta_{J-1}, \psi)}| = \psi^{J-1}$. Using this transformation in (2.35), the posterior becomes:

$$\begin{aligned} p(\alpha, \boldsymbol{\delta}, \psi, \boldsymbol{\beta}|\boldsymbol{y}) &\propto p(\boldsymbol{\beta})\alpha^{y_{1+}-1}\psi^{y_{++}+a_{+}-1}\prod_{j=1}^{J}\delta_j^{y_{+j}+a_j-1} \\ &\times \exp(\sum_{j=1}^{J} y_{1j}\boldsymbol{z}_j\boldsymbol{\beta})\exp\left(-\psi\left[\sum_{j=1}^{J}\delta_j(1+\alpha\exp(\boldsymbol{z}_j\boldsymbol{\beta}))\right]\right) \end{aligned}$$

where $y_{++} = \sum_{d=0}^{1}\sum_{j=1}^{J} y_{dj}$ and $a_{+} = \sum_{j=1}^{J} a_j$. Integrating this over α and ψ , (2.34) is obtained.

Finally, $p(\boldsymbol{\beta}|\boldsymbol{y})$ can be obtained by integrating the joint posterior of $(\alpha, \boldsymbol{\theta}, \boldsymbol{\beta})$, $p(\alpha, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y})$, over $\alpha, \boldsymbol{\theta}$. Since the order of integration does not matter, one can integrate (2.33) with respect to ω or integrate (2.34) with respect to $\boldsymbol{\delta}$ and obtain the same $p(\boldsymbol{\beta}|\boldsymbol{y})$. With respect to a case-control study, the following interpretation can be made. The equation in (2.33) corresponds to a prospective model for the data where ω is the baseline odds of disease. Similarly, equation (2.34) can be viewed as the retrospective model for the data in which δ_j is the probability that a control has exposure level $\boldsymbol{z}_j$. Although this derivation is constrained as it only applies to a discrete exposure with a Dirichlet prior, it is a significant contribution to the Bayesian analysis of case-control studies as it lays the foundation for the justifying the use of the retrospective likelihood to obtain the same posterior distribution of β had a prospective likelihood been used.

2.5.3 The Dirichlet Process

This chapter is concluded with a review of the Dirichlet process (DP), as it will play an integral role in the methods developed in Chapter 5. The Dirichlet process is a stochastic process whose realizations are probability distribution. The most common intuitive explanation of the Dirichlet process is that it is a probability distribution whose domain is itself a set of probability distributions. Moreover, it is a random probability measure. Draws from a Dirichlet process are distributions themselves.

A review of the Dirichlet distribution follows. A vector, $\boldsymbol{x}$, of size $K > 1$ is said to follow a Dirichlet distribution with parameters $\alpha_1, ..., \alpha_K$ if $f(x|\alpha) \propto \prod\limits_{i=1}^{K} x_i^{\alpha_i - 1}$ where $x_1, ..., x_K \in (0, 1)$ and $\sum\limits_{i=1}^{K} x_i = 1$.

There are a few useful properties of the Dirichlet distribution. First, if $z_i \overset{ind.}{\sim} \text{Gamma}(\alpha_i, \theta)$ then

$$\left(\frac{z_1}{\sum\limits_{i=1}^{K} z_i}, ..., \frac{z_K}{\sum\limits_{i=1}^{K} z_i} \right) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K).$$

Additionally, the additive property of the Dirichlet distribution is such that if $(x_1, ..., x_K) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K)$ then

$$(x_1, ..., x_i + x_j, ..., x_K) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_i + \alpha_j, ..., \alpha_K).$$

The technical definition of a Dirichlet process is as follows. Let $G \sim \text{DP}(\alpha, G_0)$ where $\alpha > 0$ is the concentration parameter and G_0 is the base distribution. Consider a sample space Ω and $\mathcal{F}$ a σ-algebra on Ω. Then $G \sim \text{DP}(\alpha, G_0)$ if for all measurable partitions of $\Omega(A_1, , , ., A_K)$:

$$(G(A_1), ..., G(A_K)) \sim \text{Dirichlet}(\alpha G_0(A_1), ..., \alpha G_0(A_K)) \tag{2.36}$$

This collection of finite dimensional distributions defines a stochastic process whose sample path is a probability distribution over Ω. Ferguson [1973] introduced the Dirichlet process and using properties of the Dirichlet distribution proved its existence by showing the definition in (2.36) satisfies the Kolmogorov consistency criteria.

Additional useful properties of the Dirichlet process are follows. The posterior of a Dirichlet process is also a Dirichlet process. To see this, let $\theta_i|G \sim G$ and $G \sim \mathrm{DP}(\alpha, G_0)$, then

$$G|\theta_1, ..., \theta_n \sim \mathrm{DP}\left(\alpha + n, (\alpha + n)^{-1}(\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}(\cdot)\right).$$

The mean and the variance of G can be shown to be $\mathrm{E}[G(A)] = G_0(A)$ and $\mathrm{Var}[G(A)] = \frac{G_0(A)[1-G_0(A)]}{\alpha+1}$. The expected shape of the random distribution G is G_0 and α controls the variability of the realizations around G_0.

Let $y_i|\theta_i \sim f(\theta_i)$, $\theta_i|G \sim G$ and where $G \sim \mathrm{DP}(\alpha, G_0)$ and $f(\cdot)$ is some parametric distribution. Blackwell and MacQueen [1973] obtained a representation of the marginal distribution of θ_i's in terms of the successive conditional distributions by integrating over G. The resulting characterization is

$$\theta_i|\theta_1, ..., \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta_{\theta_j}(\cdot) + \frac{\alpha}{i-1+\alpha} G_0(\cdot) \quad \text{for } i = 1, 2, 3, ... \tag{2.37}$$

where $\delta_{\theta_j}(.)$ is the point mass distribution concentrated at θ_j (the Dirac delta function, an indicator that equals 1 when $\theta = \theta_j$ and 0 otherwise) and $\theta_1 \sim G_0$. The process described in (2.37) is also known as the Pólya urn scheme.

The Dirichlet process can also be obtained by taking the limit as K goes to infinity in the following model:

$$\begin{aligned} y_i|c_i, \boldsymbol{\phi} &\sim f(\phi_{c_i}) \\ c_i|\boldsymbol{p} &\sim \mathrm{Discrete}(p_1, ..., p_K) \\ \boldsymbol{p} &\sim \mathrm{Dirichlet}(\alpha/K, ..., \alpha/K) \\ \phi_{c_i} &\sim G_0 \end{aligned}$$

where $\boldsymbol{p} = (p_1, ..., p_K)$, $\boldsymbol{\phi} = (\phi_1, ..., \phi_K)$ and c_i is a latent label indicating which of the ϕ_c's are equal. This is known as the latent Dirichlet allocation. Let $G_K = \sum_{i=1}^{K} p_i \delta_{\phi_i}$ where $\phi_i \overset{iid}{\sim} G_0$.

Ishwaran and Zarepour [2002] show that $G_K \overset{D}{\to} G$ where $G \sim DP(\alpha, G_0)$.

If i and j, $i \neq j$, have the same label ($c_i = c_j$) then $\phi_{c_i} = \phi_{c_j}$. Integrating over the proportion $\boldsymbol{p}$, the marginal probability that label $c_i = c$ is given by

$$P(c_i = c|c_1, ..., c_{i-1}) = \frac{n_{i,c} + \alpha/K}{i - 1 + \alpha}$$

where $n_{i,c}$ is the number of $c_j = c$ for $j < i$. Letting K go to infinity, the probabilities have the following limits:

$$\begin{aligned} P(c_i = c|c_1, ..., c_{i-1}) &\to \frac{n_{i,c}}{i - 1 + \alpha} \\ P(c_i \neq c_j \,\text{for all} j < i|c_1, ..., c_{i-1}) &\to \frac{\alpha}{i - 1 + \alpha} \end{aligned}$$

As a result, the conditional probability distribution of θ_i where $\theta_i = \phi_{c_i}$ is given by

$$\theta_i|\theta_1, ..., \theta_{i-1} \sim \frac{1}{i - 1 + \alpha} \sum_{j<i} \delta_{\theta_j}(\cdot) + \frac{\alpha}{i - 1 + \alpha} G_0(\cdot). \tag{2.38}$$

Note (2.37) and (2.38) have the same form. After integrating over G, the observations θ_i are exchangeable but not independent. Therefore for ease of notation can set $i \equiv n$, and (2.38) becomes:

$$\theta_i|\theta_{-i} \sim \frac{1}{n - 1 + \alpha} \sum_{j \neq i} \delta_{\theta_j}(\cdot) + \frac{\alpha}{n - 1 + \alpha} G_0(\cdot). \tag{2.39}$$

Given data $y_1, ..., y_n$ from the distribution $f(y|\theta)$, the posterior of (2.38) is given by

$$p(\theta_i|\theta_{-i}, y_i) \propto \frac{1}{n - 1 + \alpha} \sum_{j \neq i} \delta_{\theta_j}(\cdot) \times f(y_i|\theta_j) + \frac{\alpha}{n - 1 + \alpha} \left(\int G_0(\cdot) * f(y_i|\theta) d\theta \right) \pi(\theta_i|y_i).$$

Since the posterior of θ_i based on the prior $\pi(\theta_i) = G_0$ is

$$\pi(\theta_i|y_i) = \frac{\pi(\theta_i)f(y_i|\theta_i)}{\int \pi(\theta_i)f(y_i|\theta_i)d\theta_i},$$

this leads to a distribution of the form:

$$\theta_i|\theta_{-i}, y_i \sim \sum_{j\neq i} q_{ij}\delta_{\theta_j}(\cdot) + r_i H_i(\cdot) \tag{2.40}$$

where $q_{i,j} = bf(y_i|\theta_j)$ and $r_i = b\alpha \int f(y_i|\theta)dG_0(\theta)$. H_i is the posterior distribution for θ_i with prior G_0 and a single observation y_i, and b is such that $\sum\limits_{j\neq i} q_{i,j} + r_i = 1$.

An example of using (2.40) follows. Let the data be generated as:

$$\boldsymbol{Y}_i = \beta_{0i} + \beta_1 \boldsymbol{X}_i + \boldsymbol{e}_i, \tag{2.41}$$

where $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{im_i})^T$ has m_i many observations, $\boldsymbol{X}_i = (1, 2, .., m_i)^T$ and $\boldsymbol{e}_i \sim \mathrm{N}(\boldsymbol{0}, \Sigma = \sigma^2 I_{m_i})$ where I_{m_i} is a m_i by m_i identity matrix. $\boldsymbol{X}$ can be viewed as a covariate. Assume the following true values, $n = 30$, $m_i \sim \mathrm{Uniform}\{5, 6, ..., 10\}$, $\beta_1 = 1$, $\sigma^2 = 0.5$, and $\beta_{0i} = \{-5, 0, 5\}$ with 10 subjects (i's) allocated to each. Finally, consider the following prior specification for the parameters in (2.41):

$$\begin{aligned} \beta_1 &\sim \mathrm{N}(\mu_{\beta_1} = 0, \sigma^2_{\beta_1} = 5) \\ \beta_{0i}|G &\sim G \\ G &\sim \mathrm{DP}(\alpha = 1.5, G_0 \equiv \mathrm{N}(\mu_0 = 0, \sigma_0^2 = 10)) \end{aligned}$$

Let σ^2 be fixed. To estimate the model in (2.41), the full augmented data likelihood is needed and is given by

$$L(Y|\beta_{0i},\beta_1) = \prod_{i=1}^{n}\prod_{k=1}^{m_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y_{ij}-\beta_{0i}-x_{ij}\beta_1)^2\right].$$

Further, the q_{ij}'s and r_i's are as follows:

$$\begin{aligned} q_{ij} &\propto f(Y_i = y_i|\beta_1, \beta_{0j}; i \neq j) \\ &= \prod_{k=1}^{m_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y_{ik}-\beta_{0j}-\beta_1 x_{ik})^2\right] \end{aligned}$$

$$\begin{aligned} r_i &\propto \alpha\int f(y_i|\beta_{0i},\beta_1)dG_0(\beta_{0i}) \\ &= \alpha\int\prod_{k=1}^{m_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y_{ij}-\beta_{0i}-x_{ij}\beta_1)^2\right]\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left[-\frac{1}{2\sigma_0^2}\beta_{0i}^2\right]d\beta_{0i} \\ &= \alpha\left(\frac{1}{2\pi\sigma^2}\right)^{m_i/2}\sqrt{\frac{\sigma^2}{m_i\sigma_0^2+\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{\sigma^2+\sigma_0^2(m_i-1)}{\sigma^2(\sigma^2+\sigma_0^2 m_i)}\right)\sum_{j=1}^{m_i}(y_{ij}-x_{ij}\beta_1)^2\right] \end{aligned}$$

Based on the above specification, the full conditional posterior distribution of β_{0i} is given by

$$f(\beta_{0i}|\beta_1, Y_i, X_i) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{m_i}(\beta_{0i}-(y_{ij}-x_{ij}\beta_1))^2-\frac{1}{2\sigma_0^2}\beta_{0i}^2\right],$$

and completing the squares yields

$$\beta_{0i}|\beta_1, Y_i, X_i \sim \mathrm{N}\left(\left(\frac{1}{\sigma^2}+\frac{m_i}{\sigma_0^2}\right)^{-1}\left[\frac{\sum\limits_{j=1}^{m_i}(y_{ij}-x_{ij}\beta_1)^2}{\sigma^2}\right], \left(\frac{1}{\sigma^2}+\frac{m_i}{\sigma_0^2}\right)^{-1}\right).$$

To obtain H_i, the posterior of β_{0i}, it is necessary need to integrate the full conditional posterior with respect to β_1:

$$\int f(\beta_{0i}|\beta_1, Y_i, X_i)f(\beta_1|Y)d\beta_1,$$

Alternatively, one can use the full conditional posterior distribution (i.e. Gibbs sampling).

Finally, the full conditional posterior distribution of β_1 is given by

$$f(\beta_1|\beta_{0i}, Y_i, X_i) \propto \left[\prod_{i=1}^{n}\prod_{k=1}^{m_i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y_{ij}-\beta_{0i}-x_{ij}\beta_1)^2\right]\right]\exp\left(-\frac{1}{2\sigma_{\beta_1}^2}\beta_1^2\right).$$

The model is fit by implementing a Metropolis-Hastings algorithm to sample β_1 with a N(0,1) proposal distribution and samples of β_{0i} are obtained by (2.40). 5,000 samples are taken and a burn-in period of 2,000 samples is used. Based upon these draws, the posterior mean of β_1 is computed as 1.045 and Figure 2.1 shows the fitted values of β_{0i}. The Figure 2.1 also shows appropriate clustering of the β_{0i}'s.

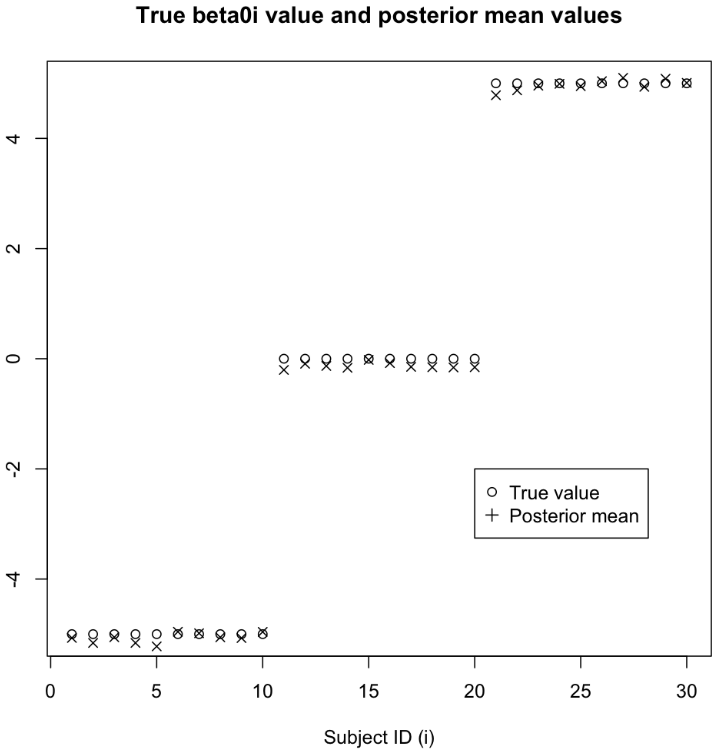

Figure 2.1: Plot of true value of β_{0i} and posterior means, by i. Circles are the true value and X's are the fitted values.

Sethuraman [1994] introduced a constructive definition of G. This definition allows for the direct construction of G. To begin, assume $G \sim \text{DP}(\alpha, G_0)$. Let $v_j \overset{iid}{\sim} \text{Beta}(1, \alpha)$ for $j = 1, 2, ...,$ and $\pi_l = v_l \prod_{j=1}^{l-1}(1 - v_j)$ for $l = 1, 2, ..., .$ The construction of $G^{'}$ is then as follows:

$$\begin{aligned} G^{'} &= \sum_{l=1}^{\infty} \pi_l \delta_{\theta_l}(\cdot) \\ \theta_l &\overset{iid}{\sim} G_0 \end{aligned} \tag{2.42}$$

Sethuraman showed that for all partitions of Ω $(A_1, , , ., A_K)$

$$(G^{'}(A_1), ..., G^{'}(A_K)) \sim \text{Dirichlet}(\alpha G_0(A_1), ..., \alpha G_0(A_K)) \tag{2.43}$$

Thus $G^{'} \overset{D}{=} G$ and hence $G^{'} \sim \text{DP}(\alpha, G_0)$. (2.39) and (2.43) show that realizations from a Dirichlet process are almost surely discrete distributions. The probability that any two draws from a Dirichlet process random measure are equal is non-zero. If the parameter of interest is believed to be from a continuous distribution, then putting a Dirichlet process prior on this parameter will result in discrete draws. Referring back to the example model given in (2.41), assume that $\beta_1 = 0$. The β_{0i} had a Dirichlet process prior, and therefore there is the possibility of equality of the posterior draws of β_{0i} $(i = 1, 2, ..., n)$ at any given iteration of the Markov chain, hence the clustering characteristic of the Dirichlet process prior.

Antoniak [1974] showed that the expected number of unique clusters is $\sum\limits_{i=1}^{n} \frac{\alpha}{\alpha+i-1}$ which can be approximated by $\alpha \text{log}\left(\frac{n+\alpha}{\alpha}\right)$. If the interest is in modeling the $\boldsymbol{Y_i}$'s, the Dirichlet process prior on the prior mean of the $\boldsymbol{Y_i}$'s normal distribution allows for a flexible class of continuous distributions for $\boldsymbol{Y_i}$. This is referred to as a Dirichlet process mixture model, and was first discussed by Antoniak [1974]. The idea was to use a Dirichlet process as a mixing distribution. Ferguson [1983] showed how density estimation could be peformed by mixtures of normals, where the parameters of the normal distribution are given a Dirichlet process prior. Lo [1984] shows that a Dirichlet process location-scale mixture of normals has full support on the space of absolutely continuous distributions. Dirichlet process mixtures are countable mixtures with an infinite number of components and a specific prior on the weights and the component-specific parameters. Modeling Y_i as being from mixture of normals, where the number of components is unknown, proceeds as follows (letting $\theta_i = (\mu_i, \sigma_i^2)$) :

$$\begin{aligned} Y_i &\sim \text{N}(\mu_i, \sigma_i^2) \\ \theta_i | G &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \end{aligned}$$

where an appropriate G_0 is chosen to have support on $R \times R_+$.

Escobar [1994] and Escobar and West [1995] implement such a model which implies that the data Y_i come from a Dirichlet process mixture of normals. They set G_0 to be an appropriate distribution for θ_i, namely the normal-inverse-gamma distribution. Sampling is done via a Gibbs sampler on the full conditionals. The conjugacy of the normal likelihood and the normal-inverse-gamma distribution allow for full conditionals to be analytically derived.

Additionally, Escobar and West [1995] introduce an approach to update the concentration parameter α. Let k denote the number of unique parameters drawn according to the DP prior and the prior on α is $\mathrm{G}(a, b)$. If $\eta|\alpha, k \sim \mathrm{B}(\alpha+1, n)$ then $\alpha|\eta, k \sim \pi\mathrm{G}(a+k, b-\log(\eta))+(1-\pi)\mathrm{G}(a+k-1, b-\log(\eta))$ where $\frac{\pi}{1-\pi} = \frac{a+k-1}{n(b-\log(\eta))}$. MacEachern and Muller [1998] expand on this DP mixture model by considering the case of non-conjugate base distributions and likelihood.

Finally, Neal [2000] summarizes several algorithms that can be used to generate draws from a DP, with (2.39) being the first of the algorithms, and expands to cases where several θ_i are updated at once. The subsequent algorithms presented in Neal [2000] address non-conjugacy of the priors, such as the sampling methods set forth by MacEachern and Muller [1998].

Chapter 3

Estimation of Covariate Effects in Matched Case-Control Designs with Multiple Events per Cluster

3.1 Introduction

As noted in Chapter 2, the case-crossover design is a hybrid between a matched case-control design and a traditional crossover design. In a matched case-control design, controls are matched to cases based upon a pre-specified set of criteria (i.e. age, location, etc.). In this case, inference is based on the comparison of the the exposure between the case and control(s) in each matching set. In a traditional crossover design investigating two treatments, each subject receives each treatment once, in a randomized order. The outcome following each treatment regime is then contrasted within the subject. Combining these two designs in the context of an observational study yields the case-crossover design.

The case-crossover design was originally developed to study the effects of transient, short-term exposures on the risk of acute events (Maclure [1991]). This type of design represents a valid and efficient design for studying relationships between exposures and an event of interest with the following characteristics: (1) the individual exposure varies within short time intervals; (2) the disease has abrupt onset and short latency for detection; and (3) the induction period is short (Jaakola [2003]). This is to say that subject level exposures vary within short time periods, the event of interest's onset can be defined and observed, and the exposures induce the event of interest within a short time period. By comparing exposures near the event period to exposures during non-event periods, each subject is able to act as their own control. As such, the case-crossover design is a powerful method for studying the effect of an exposure on the risk of a rare event because the study design allows for inherent control of within subject invariant confounding factors. Since all subjects in the data set are cases, and have therefore experienced the event of interest at least once, inference is based on a comparison of the exposure distribution among cases and controls (ie. a retrospective analysis) rather than a comparison of risk conditional on exposure (ie. a prospective analysis).

Implementation of the case-crossover design proceeds by first obtaining data on case subjects, or those subjects known to have experienced the event of interest. Once the data for the cases are obtained, the case subject's exposure from a different time period (past or future) sufficiently distant from the case exposure period is used as the control's exposure (Navidi [1998], Navidi and Weinhandl [2002]). This window of time for the control is known as the *referent period.* Since controls are created from the case subjects, there is no risk of selection bias, which occurs when controls are not representative of the population from which the cases arise from. However, there is risk of overlap bias, which occurs when the score equations from the likelihood based on the case-crossover data are biased, implying in that they do not have expectation zero (see Chapter 2). As a result, the estimating equations used to obtain parameter estimates will not have mean zero and hence the resulting estimates obtained by solving the estimating equations may show larger bias than estimators obtained from solving an unbiased estimating equation. Improperly choosing referent times will lead to overlap bias.

In a case-crossover design, the event (case) index day is taken to be the day of event. The case exposure will be that of the case index day and possibly the few days before the index day to account for lingering effects of the exposure and also to incorporate information from several days of exposures (this is typically done by taking the moving average of the case index day and a specified number of days before). To obtain the control exposures, a modified semi-symmetric bidirectional control referent selection scheme can be implemented. (Levy and Lumley [2000], Janes et al. [2005b], Janes et al. [2005a]). This control referent selection scheme picks control index times to be from either some time before or after the case index time, with equal probability. It is assumed that no event was experienced by the subject at either of the two possible times. As reviewed in Chapter 2, for those subjects at the beginning or end of the exposure time series, an offset term of log(2) is added, since these subjects will have only only one choice of control referent time available for selecting.

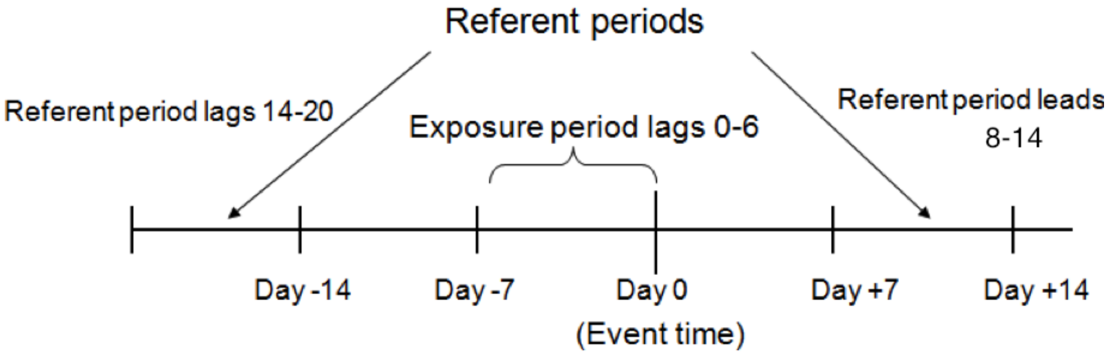

Figure 3.1: Semi-symmetric bidirectional design (Delfino et al. [2014]).

To further explain the modified semi-symmetric bidirectional control referent selection scheme, consider the example depicted in Figure 3.1 that considers a moving average of 7 days as the exposure of interest. Here we assume that the control referent index days come from the same day of the week as the case index day either 14 days before the case index day or 14 days after the case index day. By picking the same day of the week for the control index as is for the case index, confounding by day of week is mitigated, and by choosing referent times close to the case time, confounding by seasonality trends in the exposure are hopefully avoided (Bateson and Schwarts [1999]).

The modified semi-symmetric bidirectional control referent selection scheme described above gives rise to what is termed a localizable and ignorable design (Janes et al. [2005b]). A localizable referent selection scheme means there exists an unbiased estimating equation restricted to the referent windows and an ignorable referent selection scheme means that the referent sampling scheme can be ignored in conducting the analysis (the likelihood of the data does not depend on the referent sampling scheme). As such, localizable and ignorable referent selection schemes will result in no overlap bias.

Once the case-crossover dataset is created using a localizable and ignorable referent selection scheme, conditional logistic regression (CLR) can be used to obtain parameter estimates and inferences. Utilizing the mathematical equivalence between the CLR likelihood and the Cox PH partial likelihood, most current software implements the CLR model by fitting a Cox proportional hazards (Cox PH) model on a transformation of the data. The derivation of the equivalence of the conditional logistic likelihood and the Cox PH partial likelihood under the scenario of a single event experienced by each subject was shown in Chapter 2, Section 2.4. In Section 3.2 of the current chapter this result is further expanded on and the notion that repeated events within a matched set (subject) results in a CLR likelihood that is mathematically equivalent to tied survival times within strata in the Cox PH partial likelihood is shown.

In many studies that utilize a case-crossover design, there are repeated events observed on each subject over the course of the study. Examples of recurrent events studied in a case-crossover design are the number of falls in elderly people after changes in medication (Luo and Sorock [2008]) or asthma related hospital visits after exposure to environmental elements

(Delfino et al. [2014]). In this chapter, the discussion will focus on methods used to obtain parameter estimates that take into account the correlation structure among the clustered observations in the estimation procedure. These methods maintain the clustering of the data by combining all cases and controls for a given subject into a single strata with the likelihood. When simultaneously using all events and their accompanying controls from a subject to obtain parameter estimates the issue of dealing with tied event times under a Cox PH partial likelihood setting in terms of computational intensity arises. Additionally, the issue of breaking the bond between the matched case-control pairs arises. When accounting for the correlation among the data in the estimation procedure, all methods discussed in this chapter assume the willingness to break the individual matched case-control bonds. This implies that within each subject, all the case exposure values and the control values create a single risk set for this subject, which will thus be a set of the unordered exposures from that subject. In doing so, it will no longer be apparent which control was matched to which case.

The issue in breaking the bonds between each matched case-control pairing causes concern when there is evidence of seasonality trends in the exposure time series. For example, assume a subject experiences two events, one in summer and one in winter. The summer case will have a matched control which will also be from summer, based on the control selection explained previously. Similarly, the winter case will have a control matched to it from winter. If exposures in summer are generally higher than winter, then it is likely the summer control will have a value higher than that of the winter control. If both matched case-control pairs from summer and winter were to be combined to create a single risk set, the matched pair bonds would be broken. As noted in Chapter 2, Section 2.4, using the Cox PH partial likelihood to obtain parameter estimates will result in each strata's (subject's) likelihood contribution comparing the events exposures to the risk sets exposures. If the bond between matched case-control pairs is maintained, the case exposure will be strictly compared to only its risk set, which includes the case itself and its matched control. If the bonds are broken, then each case exposure will be compared to the entire risk set within that subject, which will contain case and control exposures from other seasons. In the hypothetical example presented, the case exposure from summer will be compared to the matched case-control pair from winter (and similarly the winter case exposure will be compared to the summer matched pair). If seasonality trend is evident in the exposure series, breaking of the bond will give rise to the potential of obtaining inaccurate parameter estimates.

The four most widely available methods to obtain parameter estimates for a Cox PH partial likelihood with tied event times are investigated in this chapter. The use of such methods implies that it has been deemed scientifically reasonable to break the bonds between the individually matched case-control pairs and will account for the correlation among the data (i.e. the clustering of the data) in the estimation procedure. The four methods considered are the Breslow method (Breslow [1975]), the Efron method (Efron [1977]), the Kalbfleisch and Prentice (KP) method (Kalbfleisch and Prentice [1976]), and the discrete method (Cox [1972]). Previous literature has partially studied the operating characteristics among some of the different methods (Hertz-Picciotto and Rockhill [1997], Fung et al. [2007]). Hertz-Picciotto and Rockhill [1997] investigate parameter estimation strictly from time to event viewpoint, not a case-crossover approach. In their paper they induce tied event times among the subjects by grouping together the truly continuous times of the events. Using only simulation studies based on just a single true value for the coefficient parameter, they investigate the Breslow, Efron, and discrete methods. In the simulation studies under a moderate sample size ($n = 500$), all methods range in bias from -2% to 3%, and they conclude the Efron method is the most appropriate. They do not investigate when the true parameter value approaches a null effect. It is mentioned that the discrete method could be appropriate if the number of ties is large, but this is not investigated. Fung et. al. investigate parameter estimation from a case-crossover viewpoint and only discuss the discrete method.

This chapter expands on previous work by investigating the operating characteristics, in terms of bias and mean squared error, of all four of the stated methods commonly used to compute parameter estimates in a case-crossover study with numerous events per subject (large number of ties in the Cox PH partial likelihood). Section 3.2 provides a review of the methodology for fitting data stemming from a case-crossover study and shows the equivalence between the CLR likelihood and Cox's partial likelihood when numerous events are observed per person. Section 3.3 highlights issues with methods used to obtain parameter estimates under tied event times in the Cox PH partial likelihood. Simulation studies under a variety of parameter settings and number of tied event times are discussed. The bias among the methods is shown to exhibit a distinct ordering the further the true parameter value deviates from 0. Section 3.4 presents an illustrated example using applied results from the study of air pollution exposure effects on asthma-related hospital encounters that was introduced in Chapter 1 and presented in Delfino et al. [2014].

3.2 The Conditional Logistic Regression

In a case-crossover study, all observed subjects experience the event of interests at least once since each subject acts as both the case and the control subject. A standard approach to estimating the association between the exposure and outcome in a case-crossover design is to use a conditional logistic regression (CLR) model. This method expands a simple logistic model by conditioning on the matching set and the number of events known to happen in each matched set (subjects in a case-crossover design).

We begin by showing the equivalence between the CLR likelihood with numerous matched pairs for each subject and Cox's partial likelihood with tied event times within strata. First, the likelihood contribution for the individual matched sets in the conditional logistic likelihood will be derived. $\boldsymbol{S}_i$ denote the observed and unobserved matching covariates used to define strata/subject i, $i = 1, \ldots, n$. To this end, let $\boldsymbol{Y}_i$ denote the vector of binary observations indicating occurrence of the event for each observation of subject i. Assuming each subject may experience the event more than once, the size of the vector $\boldsymbol{Y}_i$ will depend on how many events a subject experienced and the number of controls matched to each of the events. If subject i experiences d_i many events, and has 1:M matching for each event, $\boldsymbol{Y}_i$ will be a vector length $d_i(1+M) = d_i + d_iM$ $(j = 1, 2, ..., d_i + d_iM)$, where each element either takes on a value of 1 for an event or 0 for no event. For ease of exposition, this chapter will focus on the case of 1:1 matching (1 control matched to each event) in terms of simulation output and illustrations, but the derivations will be for a general $1 : M$ matching design.

We first specify the prospective probability of an event given covariates to be of the logistic form. That is to say the probability, π_{ij}, of an event for the j^{th} observation for subject i is given by $\pi_{ij} = \frac{e^{\beta_{0i}+\boldsymbol{x_{ij}\beta}}}{1+e^{\beta_{0i}+\boldsymbol{x_{ij}\beta}}}$ and $(1-\pi_{ij}) = \frac{1}{1+e^{\beta_{0i}+\boldsymbol{x_{ij}\beta}}}$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})$. Observe that each subject has there own specific intercept, β_{0i}, which varies across i.

To obtain the conditional logistic likelihood contribution for subject i, begin with a standard logistic regression likelihood contribution for subject i, $P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{S_i})$, and condition on the d_i-many events known to have occurred within subject i (ie. condition on the quantity $\sum\limits_{j=1}^{d_i(M+1)} y_{ij} = d_i$). Then the likelihood contribution for subject i is given by

$$
\begin{aligned}
P(\boldsymbol{Y}_i = \boldsymbol{y}_i | \sum_j y_{ij} = d_i, \boldsymbol{S}_i) &= \frac{P(\boldsymbol{Y}_i = \boldsymbol{y_i} \,\text{and} \sum_j y_{ij} = d_i \,|\boldsymbol{S}_i)}{P(\sum_j y_{ij} = d_i \,|\boldsymbol{S}_i)} \\
&= \frac{\prod_{j=1}^{d_i(M+1)} P(Y_{ij} = y_{ij}) * I(\sum_j y_{ij} = d_i)}{\sum_{\{\boldsymbol{y}_i^*:\sum_j y_{ij}^* = d_i\}} P(\boldsymbol{Y}_i = \boldsymbol{y}_i^*)} \\
&= \frac{\prod_{j=1}^{d_i(M+1)} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{1-y_{ij}} * I(\sum_j y_{ij} = d_i)}{\sum_{\{\boldsymbol{y}_i^*:\sum_j y_{ij}^* = d_i\}} \prod_{j=1}^{d_i(M+1)} \pi_{ij}^{y_{ij^*}} (1-\pi_{ij})^{1-y_{ij^*}}} \\
&= \frac{\prod_{j=1}^{d_i(M+1)} (\frac{e^{\beta_{0_i}+\boldsymbol{x_{ij}\beta}}}{1+e^{\beta_{0_i}+\boldsymbol{x_{ij}\beta}}})^{y_{ij}} (\frac{1}{1+e^{\beta_{0_i}+\boldsymbol{x_{ij}\beta}}})^{1-y_{ij}} * I(\sum_j y_{ij} = d_i)}{\sum_{\{\boldsymbol{y}_i^*:\sum_j y_{ij}^* = d_i\}} \prod_{j=1}^{d_i(M+1)} (\frac{e^{\beta_{0_i}+\boldsymbol{x_{ij}\beta}}}{1+e^{\beta_{0_i}+\boldsymbol{x_{ij}\beta}}})^{y_{ij}^*} (\frac{1}{1+e^{\beta_{0_i}+\boldsymbol{x_{ij}\beta}}})^{1-y_{ij}^*}} \\
&= \frac{\prod_{j=1}^{d_i(M+1)} (e^{\beta_{0_i}+\boldsymbol{x_{ij}\beta}})^{y_{ij}} * I(\sum_j y_{ij} = d_i)}{\sum_{\{\boldsymbol{y}_i^*:\sum_j y_{ij}^* = d_i\}} \prod_{j=1}^{t(M+1)} (e^{\beta_{0_i}+\boldsymbol{x_{ij}}\boldsymbol{\beta}})^{y_{ij}^*}} \\
&= \frac{e^{\sum_j (\beta_{0_i}+\boldsymbol{x_{ij}\beta}) y_{ij}} * I(\sum_j y_{ij} = d_i)}{\sum_{\{\boldsymbol{y}_i^*:\sum_j y_{ij}^* = d_i\}} e^{\sum_j (\beta_{0_i}+\boldsymbol{x_{ij}\beta}) y_{ij}^*}} \\
&= \begin{cases} \frac{\exp\{t\beta_{0i}+\sum_j \boldsymbol{x_{ij}\beta} y_{ij}\}}{\sum_{\{\boldsymbol{y}_i^*:\sum_j y_{ij}^* = d_i\}} \exp\{t\beta_{0i}+\sum_j \boldsymbol{x_{ij}\beta} y_{ij}^*\}} & \text{, if } \sum_j y_{ij} = d_i \\ 0 & \text{, otherwise} \end{cases} \\
&= \begin{cases} \frac{\exp\{\sum_j \boldsymbol{x_{ij}\beta} y_{ij}\}}{\sum_{\{\boldsymbol{y}_i^*:\sum_j y_{ij}^* = d_i\}} \exp\{\sum_j \boldsymbol{x_{ij}\beta} y_{ij}^*\}} & \text{, if } \sum_j y_{ij} = d_i \\ 0 & \text{, otherwise} \end{cases} . \qquad (3.1)
\end{aligned}
$$

To illustrate, let $d_i = 1$ and assume 1:1 matching ($M = 1$). Without loss of generality let the first observation, $\boldsymbol{x}_{i1}$ be the case exposure values and $\boldsymbol{x}_{i2}$ be the control exposure value. Then from Eq. (3.1), the likelihood contribution for the i^{th} subject reduces to

$$\frac{\exp\{\boldsymbol{x_{i1}\beta}\}}{\exp\{\boldsymbol{x_{i1}\beta}\} + \exp\{\boldsymbol{x_{i2}\beta}\}}. \tag{3.2}$$

3.2.1 Equivalence Between the CLR likelihood and Cox's Partial Likelihood

In Chapter 2 it was shown that with a single event per matched set, the conditional logistic likelihood is mathematically equivalent to the Cox's partial likelihood provided that the "survival times" for controls were specified to be greater than the specified "survival times" for cases. Here we consider Cox's partial likelihood in the setting of tied event times within a strata.

Let i denote the stratum, $d_{(i)}$ denote the number of events (all with tied times) in stratum i, $D_{(i)}$ be the set of events, and $(R_{(i)}; d_{(i)})$ denote all sets of size $d_{(i)}$ from $R_{(i)}$, the risk set for this stratum consisting of all observations (cases and controls). Under the proportional hazards model presented in Chapter 2, the partial likelihood contribution for this stratum is given by

$$\begin{aligned} L_{(i)} &= P\{\text{all } x_{il} \in D_{(i)} \text{ experience an event} |\, d_{(i)} \text{ subjects in } R_{(i)} \text{ experience an event}\} \\ &= \frac{P\{\text{all } x_{il} \in D_{(i)} \text{ experience an event}\}}{P\{d_{(i)} \text{ subjects in } R_{(i)} \text{ experience an event}\}} \\ &= \frac{\prod\limits_{l \in D_{(i)}} \exp\{\boldsymbol{x_{il}\beta}\}}{\sum\limits_{L \in (R_{(i)};d_{(i)})} \prod\limits_{l \in L} \exp\{\boldsymbol{x_{il}\beta}\}} \end{aligned} \tag{3.3}$$

It is easily seen that (3.1) and (3.3) are mathematically equivalent as long as all event times within strata are set to be equal and control times are set to be greater than the event times. With respect to a case-crossover design, where there is no time element, it can be viewed that the subscript (i) refers to the individual subject. Each subject i experiences d_i many events, with $D_{(i)}$ being the event set (the cases) and $R_{(i)}$ being the risk set for that subject (which contains all the events (cases) experienced by the subject and the accompanying controls matched for each case). The full data partial likelihood across all subjects is then taken as the product of these individual contributions across all i, $L = \prod_i L_{(i)}$.

In practice, maximization of the CLR likelihood is implemented by maximizing Cox's partial likelihood. The procedure involves manipulating the data to be used in the conditional logistic likelihood by creating a time variable, where all event (case) times are set to a certain fixed time and all control times are set to any times such that the event time is less than or equal to the control times. If event times are not set to be equal to each other, then this would imply there is some inherent ordering to the events, which is not generally the case in a matched retrospective design.

The more events a single subject experiences is analogous to more tied event times within each strata in the partial likelihood. The likelihood presented in (3.3) will combine all case and control exposures from a single subject to create the risk set. This amounts to breaking the bonds between each matched case-control pairing and is reasonable to do if there is no evidence of trends in the exposure covariates. If there are trends then breaking of the matched pair bonds could potentially lead to biased estimates as discussed in Section 3.1. All methods discussed in the following section will break the bonds between each individual matched case-control pair for subjects with numerous matched pairs. If a subject has only a single matched pair, then the bonds will be maintained.

3.2.2 Tied Event Times Under the Cox PH Setting

In the presence of numerous tied event times in a Cox PH model (numerous events within a subject in a CLR setting), several methods to obtain parameter estimates are available. The issue is that the denominator, which involves $(R_{(i)}; d_{(i)})$, for each likelihood contribution in (3.3) becomes computationally intensive to compute. For example in a 1:1 matching setting and having a subject with 10 events over the course of a study, there are 184,756 terms (number of possible ways to choose 10 events among 20 trials) required to compute for the denominator. This difficulty is analogous to conditioning on $\sum_{j=1}^{d_i(M+1)} y_{ij} = d_i > 1$ in the conditional logistic regression likelihood.

There are two widely used methods which attempt to approximate the partial likelihood with tied event times. These methods were proposed by Breslow (Breslow [1975]) and Efron (Efron [1977]). Both methods try to circumvent the need to calculate the computationally intensive denominator in (3.3).

The Breslow approximation assumes the risk set stays constant across the events. As such, denominator in the likelihood has the same exposures (event and non-event) across all the events so that

$$L_{(i)}^{Breslow} = \frac{\prod_{l \in D_{(i)}} \exp\{\boldsymbol{x}_{il}\boldsymbol{\beta}\}}{\left[\sum_{l \in R_{(i)}} \exp\{\boldsymbol{x}_{il}\boldsymbol{\beta}\}\right]^{d_{(i)}}}. \tag{3.4}$$

The Efron approximation attempts to account for the fact that observations associated with an observed event should not be in the risk set by giving them lower weights in the product across events. In this case, the resulting partial likelihood contribution is given by

$$L_{(i)}^{Efron} = \frac{\prod\limits_{l \in D_{(i)}} \exp\{\boldsymbol{x_{il}\beta}\}}{\prod\limits_{h=1}^{d_{(i)}} \left[\sum\limits_{l \in R_{(i)}} \exp\{\boldsymbol{x_{il}\beta}\} - \frac{h-1}{d_{(i)}} \sum\limits_{k \in D_{(i)}} \exp\{\boldsymbol{x_{ik}\beta}\} \right]}. \tag{3.5}$$

An approach provided by Kalbfleisch and Prentice (KP) assumes that time is truly continuous and that the probability of tied event times is 0 (Kalbfleisch and Prentice [1973], Kalbfleisch and Prentice [1980]). Under this setting, it is assumed tied event times are observed due to not being able to measure time on a finer scale. If the time of event was to be observed on a finer scale (eg. down to the millisecond), this method assumes that there is an inherent ordering to the event times. This approach then accounts for all possible ordering of event times and Kalbfleisch and Prentice consider the average partial likelihood contribution at time $t_{(j)}$ that arises from breaking the ties in all possible ways.

Let $Q_{(i)}$ denote the set of $d_{(i)}!$ permutations of the events in stratum i and $P = (p_1, \ldots, p_{d_{(i)}})$ be an element in $Q_{(i)}$. Let $R(i,P,r) = R_{(i)} - \{p_1, \ldots, p_{r-1}\}$. Setting $\boldsymbol{s_{(i)}} = \sum\limits_{l \in D_i} \boldsymbol{x_{il}}$ then yields the likelihood contribution:

$$L_{(i)}^{KP} = \frac{1}{d_{(i)}!} \exp\{\boldsymbol{s_{(i)}\beta}\} \sum_{P \in Q_{(i)}} \prod_{r=1}^{d_{(i)}} \left\{ \sum_{l \in R(i,P,r)} \exp\{\boldsymbol{x_{il}\beta}\} \right\}^{-1}. \tag{3.6}$$

Given the combinatorial complexity of the KP method, it is generally implemented in software by using an integral representation to the likelihood contribution for strata i, $L_{(i)}$ (DeLong et al. [1994]) such that

$$L_{(i)} = \int_0^\infty \prod_{k \in D_i} \{1 - \exp(-\lambda_{ik} t)\} \exp(-t) dt,$$

where $w_l = \exp(\boldsymbol{x_{il}\beta})$ and $\lambda_{ik} = \frac{w_k}{\sum\limits_{l \in R_i/D_i} w_l}$. Details of this approach are in Appendix 3.6.2.

The last considered approach for handling ties in the partial likellihood is due to Cox [1972] and does not assume there is an inherent ordering, and that the event times are truly discrete and tied. This method computes the likelihood under truly tied event times by using a recursive method to compute the computationally intensive combinatoric of the denominator in (3.3) (Gail et al. [1981]). Letting $r_{(j)}$ be the size of the risk set $R_{(j)}$, the main difficulty in computing the likelihood in (3.3) is the computation of $B(d_{(j)}, r_{(j)}) = \sum\limits_{I \in (R_{(j)};d_{(j)})} \prod\limits_{i \in I} \exp\{\boldsymbol{x_i\beta}\}$ which requires $\frac{r_{(j)}!}{(r_{(j)}-d_{(j)})!(d_{(j)}-1)!} - 1$ arithmetic operations, but reduces to $2d_{(j)}(r_{(j)} - d_{(j)} + 1)$ operations using recursion. This is accomplished by noting that $B(d_{(j)}, r_{(j)}) = B(d_{(j)}, r_{(j)} - 1) + \exp\{\boldsymbol{x_{r_j}\beta}\}B(d_{(j)} - 1, r_{(j)} - 1)$. Numerical differentiation is employed to obtain the score equations needed to obtain parameter estimates. Further details of this approach are provided in Appendix 3.6.1

All quantities above were for a single strata's contribution to the full likelihood. For stratified data with numerous independent strata, the partial likelihood over the full data is obtained by taking the product across all strata, yielding

$$L_p = \prod_{i=1}^{n} L_{(i)},$$

where $L_{(i)}$ is the likelihood contribution from strata/subject i, defined in equations (3.3), (3.4), (3.5), or (3.6) depending upon the method chosen for handling tied event times.

3.3 Simulation Study

In this section a simulation study is conducted to highlight the operating characteristics of the four methods discussed in the previous section. The goal is to show that as the true value of the parameter to be estimated deviates further from zero, there is an ordering in terms of bias among the four different computational methods.

3.3.1 Data Simulation

The focus of the data simulation procedure is to mimic data that would be obtained from an air exposure study similar to the motivating example presented in Chapter 1. In this case, the scientific objective is to make inference on the association between atmospheric exposures and a adverse health event, namely asthma-related hospital encounters. In this case, each subject has there own recorded exposure time series values, which can possibly be shared among other subjects living in close proximity. From those exposure values, a case exposure and control(s) exposure(s) are created.

To simulate the data, four strings of exposures were generated via an autoregressive time series of order 1, AR(1). For $k = 1, 2, 3, 4$ and $t = 1, \ldots, T$, where $T = 3,650$ to signify a 10 year study with daily exposures, an AR(1) process was generated for the exposure, x such that

$$x_{kt} = c + \varphi x_{kt-1} + \varepsilon_{kt} \text{ where } \varepsilon_{kt} \overset{iid}{\sim} N(0,1).$$

Each subject, $i = 1, \ldots, n$, was assigned to one of the four generated exposure time series with equal probability. For $t = 1, \ldots, 3,650$, the probability of an event on a given day was computed using the logistic link for a prospective probability of event. Specifically, it was assumed that $\pi_{k_i t} = \frac{exp(\beta_{0i}+\beta_1 x_{k_i t})}{1+exp(\beta_{0i}+\beta_1 x_{k_i t})}$, where k_i is the monitoring center assigned to subject i and $k \in \{1, 2, 3, 4\}$.

For each π_{it}, a Bernoulli random variable with $p = \pi_{it}$ was sampled. Given the pre-specified number of events per subject (v) to be investigated, v many events were randomly sampled from the success' generated from the T many Bernoulli trials. If a subject did not produce enough events via the Bernoulli simulation based on the π_{it}'s as they were set to have (events simulated$< v$), then all events that were produced for that subject were chosen.

Once the event days are selected, a control day was chosen for each event based on the suggestions of Navidi and Weinhandl [2002]. For a given event day, the control exposure values were taken to be the moving average of the 7 days between the 14-th and 20-th days prior to the event or the moving average of 7 days between the 8-th and 14-th days past the event, with each window being picked with equal probability. The event exposures are set to the be the average of the event day and the previous 6 days. This is done in air pollution studies to account for any lingering effect of the exposure on the risk of an adverse health outcome, and also to be able to incorporate the information from several days into the analysis. For events that had only a single available referent window to be selected, in cases where the event happens in the start or end of the exposure series or if another event is observed in one of the two available referent windows, an offset of log(2) was included in the parameter estimation methods (Janes et al. [2005a]). If for a specific event both available referent windows had events observed in them, then this event was dropped from the analysis (this occurred in less than 0.1% of the total events).

The time series used for generating exposures specified $c = 2$ and $\varphi = 0.5$, so that $\text{cov}(x_{kt}, x_{kt+r}) = \frac{1}{1-\varphi^2}\varphi^{|r|}$. The random intercept for each subject was generated according to $\beta_{0i} \overset{iid}{\sim} N(\mu_0, 1)$ where, depending on the scenario of the true value of β_1 and the number of events desired within each subject, μ_0 was varied from -7.5 to -5.5 in order to ensure a sufficient expected number of events occurred across the 3650 days. The expected number of events in a strata was computed using the Monte Carlo integral approximation given by

$$\begin{aligned} \mathrm{E}(v) &= \mathrm{E}_x[\mathrm{E}_v(v|x)] \\ &= \mathrm{E}_x\left[T\frac{\exp(\beta_{0i}+\beta_1 x)}{1{+}\exp(\beta_{0i}+\beta_1 x)}\right] \\ &= T\mathrm{E}_x\left[\frac{\exp(\beta_{0i}+\beta_1 x)}{1{+}\exp(\beta_{0i}+\beta_1 x)}\right] \\ &= T\int_x \frac{\exp(\beta_{0i}+\beta_1 x)}{1{+}\exp(\beta_{0i}+\beta_1 x)}dF(x). \end{aligned}$$

where $F(x)$ represents the normal cumulative distribution function with $\mu = 2$ and $\sigma^2 = \frac{1}{1-\varphi^2} = \frac{4}{3}$.

3.3.2 Simulation Output

The simulation studies presented here consider three different numbers of subjects per simulation: n=500, 1000, and 5000. For each sample size, four different true parameter values were used, $\beta_1 = \log(1){=}0$, log(1.3)=0.2624, log(1.5)=0.4055, and log(2.0)=0.6931, and three different number of events for each subject (3,5, and 10 events each) were considered. In each scenario, 5,000 simulated datasets were analyzed. The summarized output for each scenario of sample size, true β value, and number of events within subject are provided in Tables 3.1, 3.2, and 3.3.

Focusing on the scenarios with $\beta > 0$, for any number of events per subject there is a general ordering of bias among the methods. The ordering (from lowest to highest observed bias) is discrete, KP, Efron ,and then Breslow. The Breslow method provides estimates with approximately 45% bias, the Efron method about 30%, the KP method 23%, and the discrete method about 1%. The underestimation of the Breslow approximation method has been noted in previous work (Cox and Oakes [1984]). This approximation's estimates are

$\boldsymbol{\beta} = \log(\mathbf{1}) = \mathbf{0}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	0.0003	0.0634	0.0500	0.0023
Efron	0.0003	0.0640	0.0615	0.0037
Cont. (KP)	0.0003	0.0664	0.0663	0.0044
Discrete	0.0004	0.0800	0.0810	0.0060
	Events=5			
Breslow	0.0008	0.0472	0.0354	0.0012
Efron	0.0011	0.0475	0.0464	0.0021
Cont. (KP)	0.0011	0.0630	0.0638	0.0022
Discrete	0.0015	0.0500	0.0494	0.0041
	Events=10			
Breslow	0.0001	0.0324	0.0318	0.0010
Efron	0.0002	0.0326	0.0235	0.0005
Cont. (KP)	0.0003	0.0400	0.0440	0.0011
Discrete	0.0002	0.0334	0.0335	0.0020
$\boldsymbol{\beta} = \log(\mathbf{1.3}) = \mathbf{0.2624}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.1076	0.0635	0.0484	0.0139
Efron	-0.0675	0.0642	0.0610	0.0082
Cont. (KP)	-0.0497	0.0677	0.0681	0.0071
Discrete	0.0009	0.0840	0.0848	0.0070
	Events=5			
Breslow	-0.1188	0.0471	0.0339	0.0152
Efron	-0.0736	0.0476	0.0448	0.0074
Cont. (KP)	-0.0596	0.0497	0.06430	0.0059
Discrete	0.0010	0.0600	0.0492	0.0041
	Events=10			
Breslow	-0.1265	0.0324	0.0232	0.0165
Efron	-0.0782	0.0326	0.0316	0.0071
Cont. (KP)	-0.0671	0.0458	0.0341	0.0056
Discrete	0.0005	0.0337	0.0463	0.0021
$\boldsymbol{\beta} = \log(\mathbf{1.5}) = \mathbf{0.4055}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.1706	0.0637	0.0466	0.0312
Efron	-0.1095	0.0645	0.0589	0.0154
Cont. (KP)	-0.0771	0.0700	0.0694	0.0107
Discrete	0.0026	0.0870	0.0870	0.0075
	Events=5			
Breslow	-0.1884	0.0471	0.0332	0.0366
Efron	-0.1193	0.0477	0.0442	0.0160
Cont. (KP)	-0.0938	0.0505	0.0505	0.0113
Discrete	0.0017	0.0671	0.0671	0.0045
	Events=10			
Breslow	-0.2000	0.0323	0.0224	0.0405
Efron	-0.1261	0.0326	0.0310	0.0168
Cont. (KP)	-0.1061	0.0342	0.0346	0.0124
Discrete	0.0010	0.0473	0.0479	0.0023
$\boldsymbol{\beta} = \log(\mathbf{2}) = \mathbf{0.6931}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.3163	0.0643	0.0440	0.1019
Efron	-0.2177	0.0652	0.0550	0.0504
Cont. (KP)	-0.1362	0.0759	0.0764	0.0244
Discrete	0.0051	0.0960	0.0960	0.0093
	Events=5			
Breslow	-0.3463	0.0472	0.0320	0.1209
Efron	-0.2333	0.0479	0.0417	0.0561
Cont. (KP)	-0.1693	0.0543	0.0550	0.0317
Discrete	0.0022	0.0738	0.0744	0.0055
	Events=10			
Breslow	-0.365	0.0322	0.0210	0.1336
Efron	-0.2427	0.0326	0.2823	0.0597
Cont. (KP)	-0.1925	0.0361	0.0357	0.0383
Discrete	0.0013	0.0518	0.0511	0.0026

Table 3.1: Simulation output for methods to obtain estimates under tied event times in a Cox PH partial likelihood. Sample size $n = 500$.

$\boldsymbol{\beta} = \log(\mathbf{1}) = \mathbf{0}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	0.0001	0.0200	0.0153	0.0002
Efron	0.0001	0.0202	0.0193	0.0003
Cont. (KP)	0.0001	0.0210	0.0208	0.0004
Discrete	0.0001	0.0258	0.0256	0.0005
	Events=5			
Breslow	0.0002	0.0150	0.0111	0.0001
Efron	0.0003	0.0149	0.01459	0.0002
Cont. (KP)	0.0003	0.0154	0.0155	0.0002
Discrete	0.0004	0.0200	0.0200	0.0003
	Events=10			
Breslow	0.0000	0.0102	0.0080	0.0001
Efron	0.0000	0.0103	0.0102	0.0001
Cont. (KP)	0.0000	0.0105	0.0108	0.0001
Discrete	0.0000	0.0141	0.0144	0.0001
$\boldsymbol{\beta} = \log(\mathbf{1.3}) = \mathbf{0.2624}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.107	0.0200	0.0150	0.0117
Efron	-0.0670	0.0202	0.0190	0.0048
Cont. (KP)	-0.0500	0.0213	0.0210	0.0029
Discrete	0.0005	0.0265	0.0265	0.0006
	Events=5			
Breslow	-0.1188	0.0149	0.0110	0.0142
Efron	-0.0735	0.0150	0.0144	0.0056
Cont. (KP)	-0.0600	0.0157	0.0157	0.0038
Discrete	0.0004	0.0205	0.0206	0.0004
	Events=10			
Breslow	-0.1265	0.0102	0.0080	0.0160
Efron	-0.0783	0.0103	0.0100	0.0062
Cont. (KP)	-0.0674	0.0106	0.0106	0.0046
Discrete	0.0001	0.0144	0.0144	0.0002
$\boldsymbol{\beta} = \log(\mathbf{1.5}) = \mathbf{0.4055}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.1708	0.0201	0.0150	0.0293
Efron	-0.1097	0.0203	0.0185	0.0123
Cont. (KP)	-0.0786	0.0219	0.0216	0.0066
Discrete	0.0007	0.0274	0.0270	0.0007
	Events=5			
Breslow	-0.1885	0.0149	0.0110	0.0356
Efron	-0.1194	0.0150	0.0140	0.0144
Cont. (KP)	-0.0946	0.0160	0.0160	0.0092
Discrete	0.0005	0.0211	0.0210	0.0004
	Events=10			
Breslow	-0.2001	0.0102	0.0069	0.0401
Efron	-0.1262	0.0103	0.0095	0.0160
Cont. (KP)	-0.1065	0.0108	0.0106	0.0114
Discrete	0.0002	0.0149	0.0146	0.0002
$\boldsymbol{\beta} = \log(\mathbf{2}) = \mathbf{0.6931}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.3172	0.0203	0.0140	0.1008
Efron	-0.2186	0.0206	0.0178	0.0481
Cont. (KP)	-0.1397	0.0238	0.0242	0.0201
Discrete	0.0010	0.0303	0.0306	0.0009
	Events=5			
Breslow	-0.3463	0.0149	0.0100	0.1200
Efron	-0.2333	0.0151	0.01320	0.0546
Cont. (KP)	-0.1704	0.0171	0.01735	0.0294
Discrete	0.0006	0.0233	0.0230	0.0005
	Events=10			
Breslow	-0.3652	0.0101	0.0070	0.1334
Efron	-0.2431	0.0103	0.0090	0.0592
Cont. (KP)	-0.1935	0.0114	0.0114	0.0375
Discrete	0.0001	0.0163	0.0163	0.0002

Table 3.2: Simulation output for methods to obtain estimates under tied event times in a Cox PH partial likelihood. Sample size $n = 1000$.

$\boldsymbol{\beta} = \log(\mathbf{1}) = \mathbf{0}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.0001	0.0089	0.0070	0.0001
Efron	-0.0001	0.0090	0.0086	0.0001
Cont. (KP)	-0.0001	0.0093	0.009	0.0001
Discrete	-0.0001	0.0115	0.0114	0.0001
	Events=5			
Breslow	0.0000	0.0066	0.0049	0.0000
Efron	0.0000	0.0067	0.0064	0.0000
Cont. (KP)	0.0000	0.0069	0.0068	0.0000
Discrete	0.0000	0.0089	0.0088	0.0000
	Events=10			
Breslow	0.0000	0.0045	0.0034	0.0000
Efron	0.0000	0.0046	0.0045	0.0000
Cont. (KP)	0.0000	0.0047	0.0047	0.0000
Discrete	0.0000	0.0063	0.0063	0.0000
$\boldsymbol{\beta} = \log(\mathbf{1.3}) = \mathbf{0.2624}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.1073	0.0090	0.0070	0.0115
Efron	-0.0671	0.0090	0.0085	0.0045
Cont. (KP)	-0.0502	0.0095	0.0095	0.0026
Discrete	0.0002	0.0118	0.0118	0.0001
	Events=5			
Breslow	-0.1190	0.0066	0.0050	0.0142
Efron	-0.0739	0.0067	0.0064	0.0051
Cont. (KP)	-0.0603	0.0070	0.0071	0.0036
Discrete	0.0000	0.0091	0.0090	0.0001
	Events=10			
Breslow	-0.1267	0.0045	0.0033	0.0160
Efron	-0.0785	0.0046	0.0045	0.0061
Cont. (KP)	-0.0675	0.0047	0.0048	0.0045
Discrete	0.0000	0.0064	0.0065	0.0000
$\boldsymbol{\beta} = \log(\mathbf{1.5}) = \mathbf{0.4055}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.1711	0.0090	0.0082	0.0293
Efron	-0.1101	0.0091	0.0065	0.0121
Cont. (KP)	-0.0791	0.0097	0.0096	0.0063
Discrete	0.0001	0.0122	0.0120	0.0001
	Events=5			
Breslow	-0.1888	0.0067	0.0050	0.0350
Efron	-0.1199	0.0067	0.0063	0.0144
Cont. (KP)	-0.0952	0.0071	0.0072	0.0091
Discrete	-0.0001	0.0094	0.0096	0.0000
	Events=10			
Breslow	-0.2002	0.0045	0.0031	0.0401
Efron	-0.1264	0.0046	0.0043	0.0160
Cont. (KP)	-0.1067	0.0048	0.0048	0.0110
Discrete	0.0000	0.0066	0.0066	0.0000
$\boldsymbol{\beta} = \log(\mathbf{2}) = \mathbf{0.6931}$	**Events=3**			
	Bias	**Model SE**	**Empirical SE**	**MSE**
Breslow	-0.3175	0.0090	0.0063	0.1008
Efron	-0.2191	0.0092	0.0080	0.0480
Cont. (KP)	-0.1406	0.0106	0.0109	0.0198
Discrete	0.0000	0.0135	0.0137	0.0001
	Events=5			
Breslow	-0.3466	0.0067	0.0044	0.1201
Efron	-0.2337	0.0067	0.0059	0.0546
Cont. (KP)	-0.1711	0.0076	0.0077	0.0293
Discrete	-0.0001	0.0104	0.0104	0.0001
	Events=10			
Breslow	-0.3653	0.0045	0.0030	0.1330
Efron	-0.2433	0.0046	0.0041	0.0592
Cont. (KP)	-0.1938	0.0051	0.0051	0.0370
Discrete	-0.0001	0.0073	0.0074	0.0000

Table 3.3: Simulation output for methods to obtain estimates under tied event times in a Cox PH partial likelihood. Sample size $n = 5000$.

heavily attenuated towards the null ($\beta_1 = 0$), and for any given true parameter value greater than 0, the bias increases as the number of events increases. Previously it was noted that as the number of events per strata increases, the exposure values will be over represented in the effective risk set (the denominator of the likelihood contribution for a strata). The more events per strata leads to more tied event times in the Cox PH partial likelihood, which results in more event exposure values being overrepresented in the effective risk set of that strata. The KP method is also attenuated towards 0 but not as much as the Breslow and Efron approximation methods.

What can also be seen from the simulation results is that as the true β value deviates further from 0, the bias among the approximations and KP methods becomes more apparent. Looking at Table 3.2, where sample size $n = 1,000$ with 10 events within each subject, the bias among these methods when $\beta = 0.2624$ ranges from -0.1265 to -0.0674. When the true value is $\beta = 0.6934$ the bias range increases to -0.3652 to -0.1935. The bias associated with the discrete method is approximately 0 across these comparisons. When $\beta = 0$, there is virtually no difference among the methods in terms of bias, and each method has virtually 0 bias. Intuitively, this is because in the absence of an association between the exposure and the event, the exposure of all observation within a risk set are exchangeable within the risk set.

The simulations confirm that inference will depend on what method is chosen to analyze case-crossover data with repeated events when the true coefficient value is not 0. Most software used to analyze case-crossover data in the presence of numerous events among the subjects will automatically default to one of the approximation methods. In time-to-event analyses it is uncommon to have a high number of tied event times, with a large number of events occurring at each of these times. As a result, when only a few distinct times have a relatively few number of events occurring simultaneously, the approximation methods will obtain parameter estimates with relatively low bias. In the setting of case-crossover data with repeated events, relative to the strata size there are a large number of tied event times occurring in each strata. As a result, the approximation methods exhibit large bias when the true coefficient value is not 0.

The biases for each of the four methods are plotted against the true β parameter value are in Figures 3.2, Figure 3.3, and Figure 3.4. Mean squared error (MSE) plots are provided in Figures 3.5, 3.6, and 3.7. As can be seen across all figures, the further the true coefficient parameter value differs from zero, increases in bias occur at an almost linear rate among the approximation and KP methods, while the bias associated with the discrete method increases only slightly. As a result, the MSE plots show an increase in MSE among the approximation and KP methods as the true β value differs further from zero.

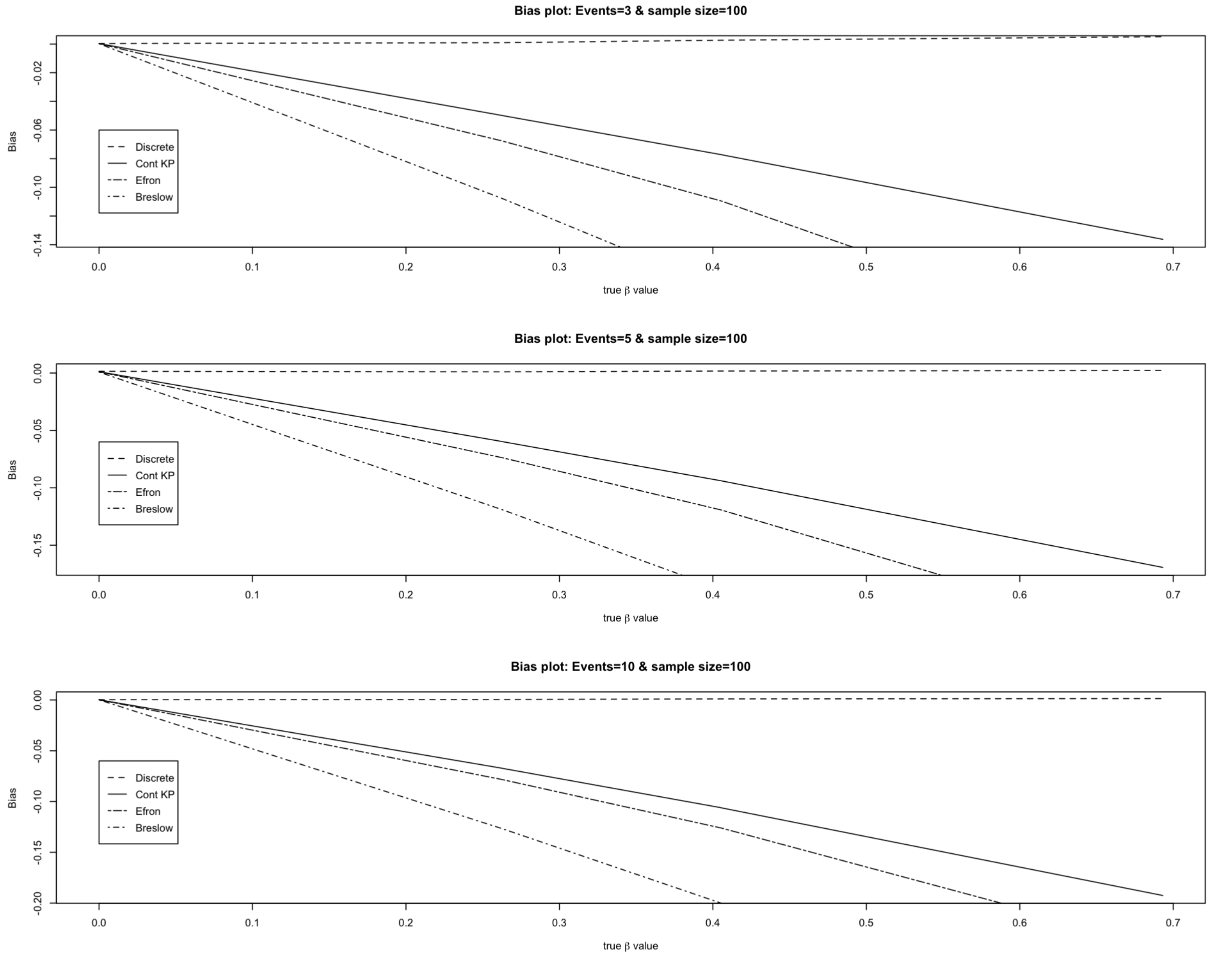

Figure 3.2: Plot of bias against the true β value for the Breslow, Efron, KP, and discrete estimation method. $n = 100$

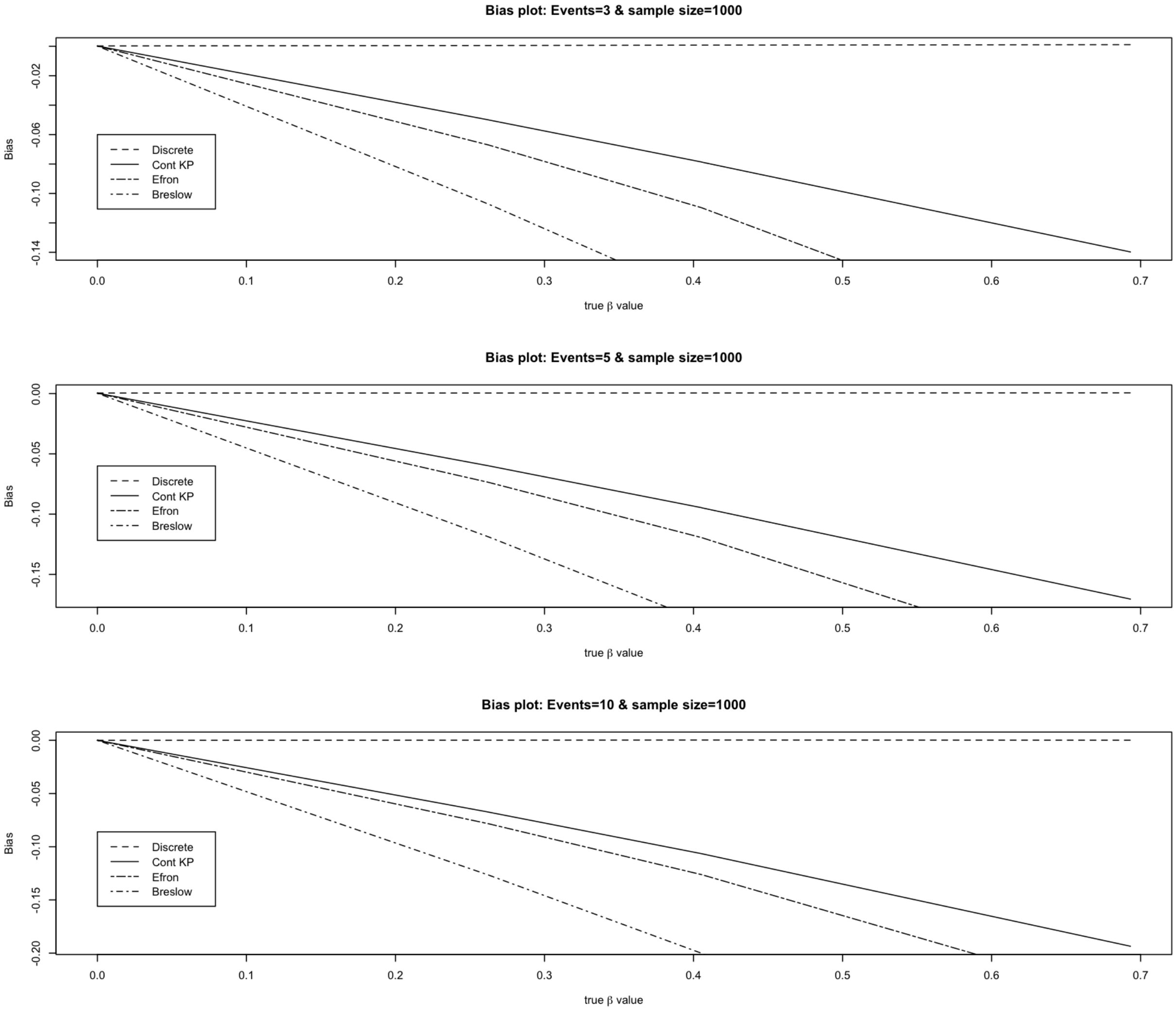

Figure 3.3: Plot of bias against the true β value for the Breslow, Efron, KP, and discrete estimation method. $n = 1000$

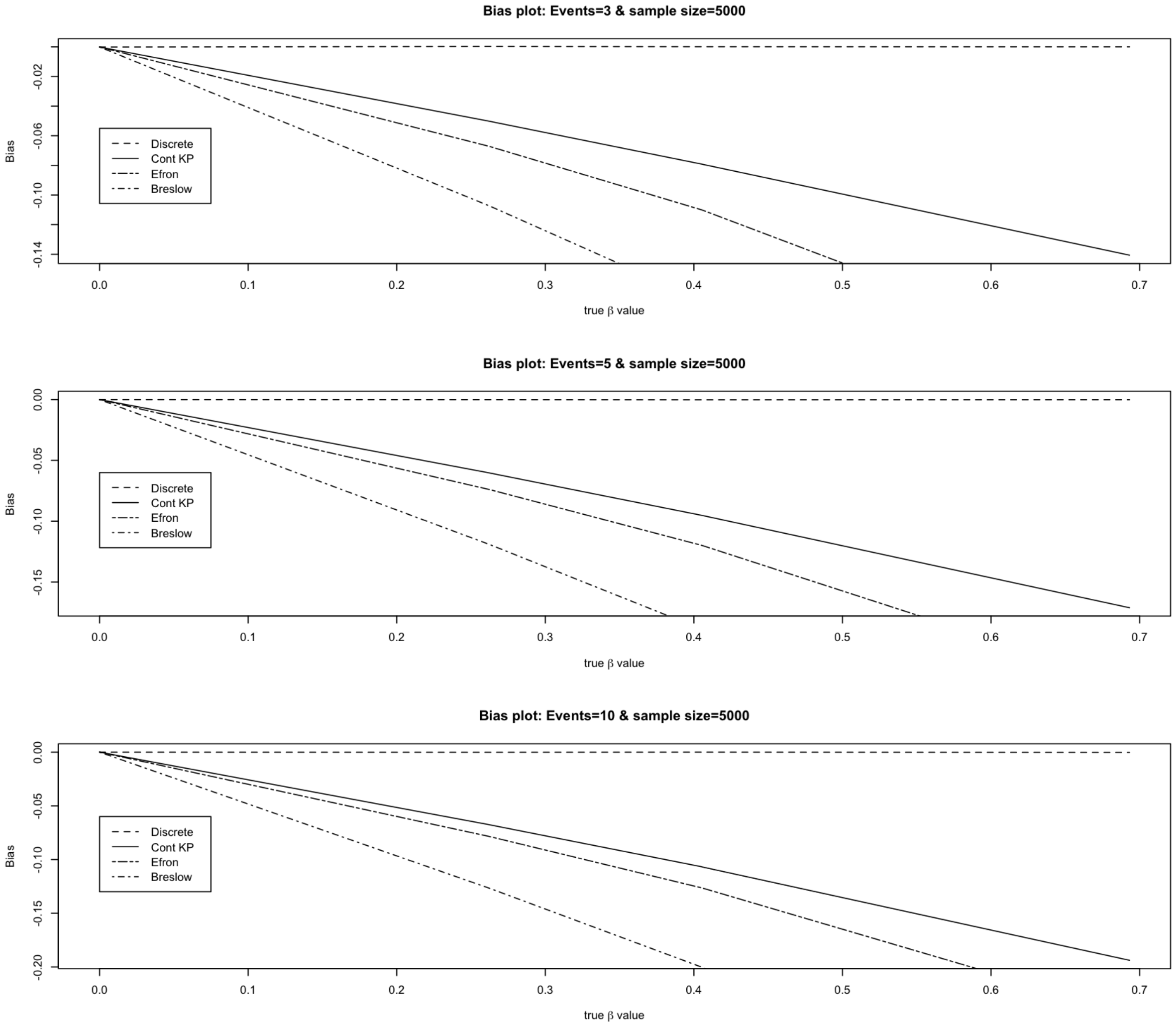

Figure 3.4: Plot of bias against the true β value for the Breslow, Efron, KP, and discrete estimation method. $n = 5000$

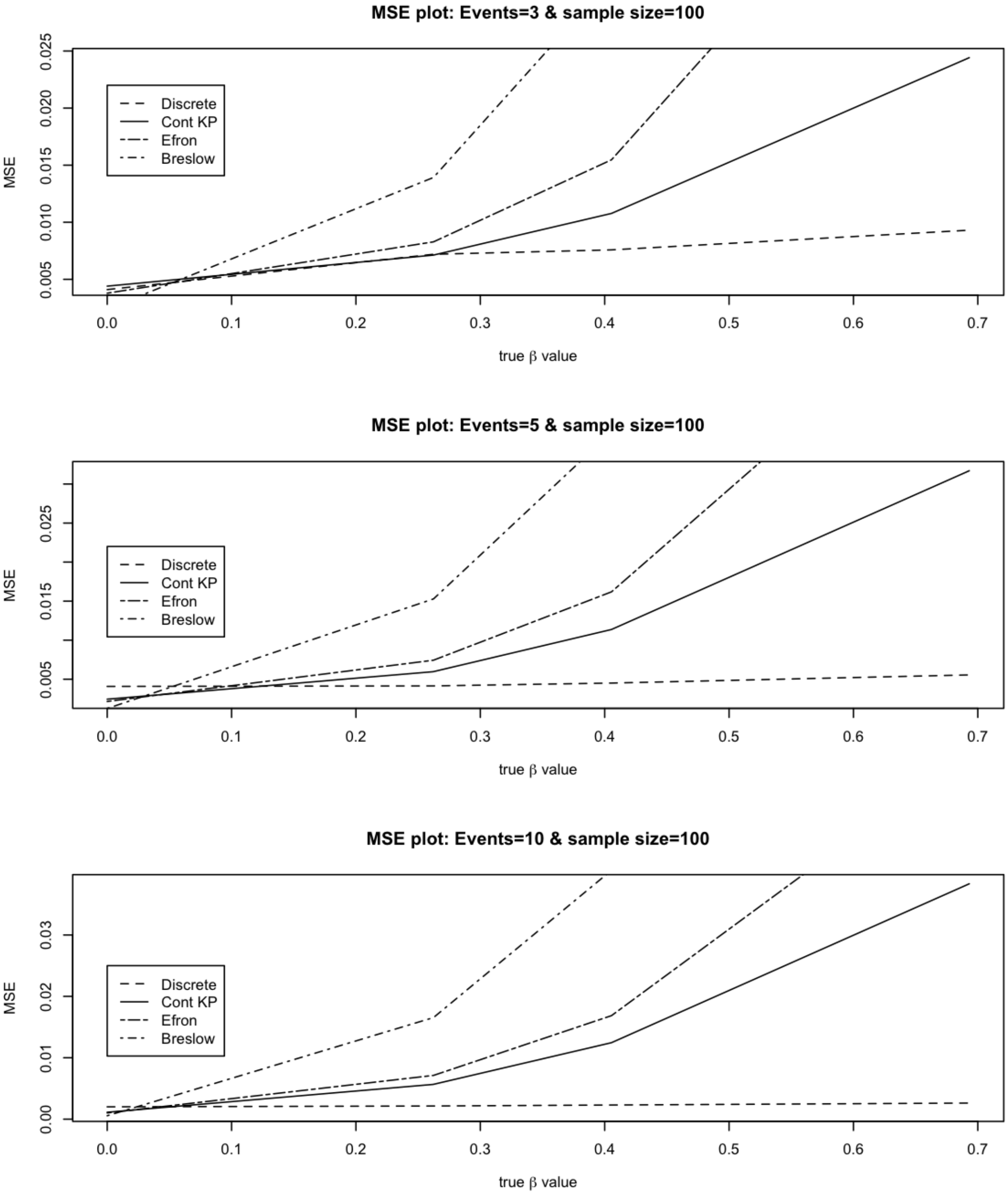

Figure 3.5: Plot of mean squared error against the true β value for the Breslow, Efron, KP, and discrete estimation method. $n = 100$

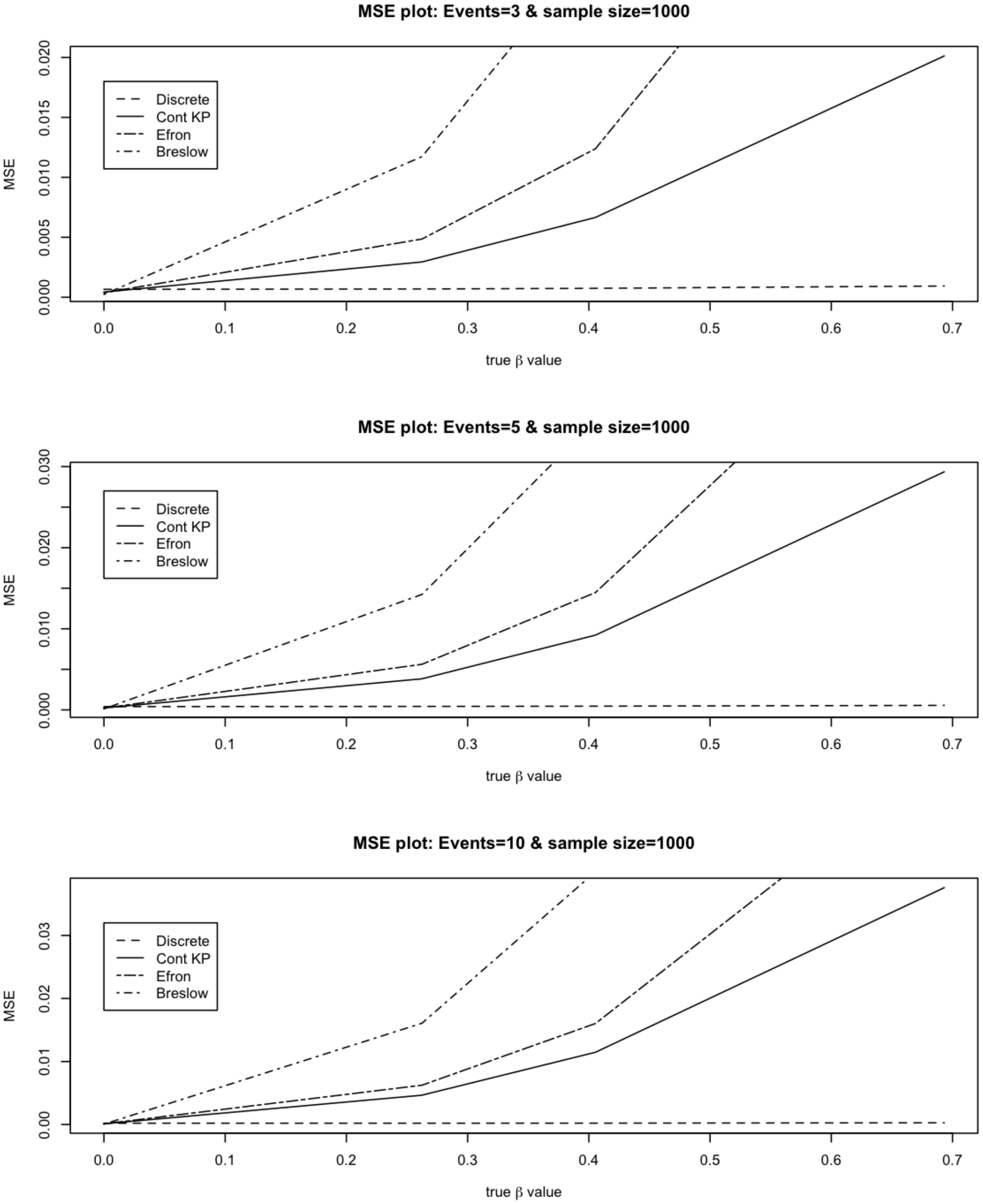

Figure 3.6: Plot of mean squared error against the true β value for the Breslow, Efron, KP, and discrete estimation method. $n = 1000$

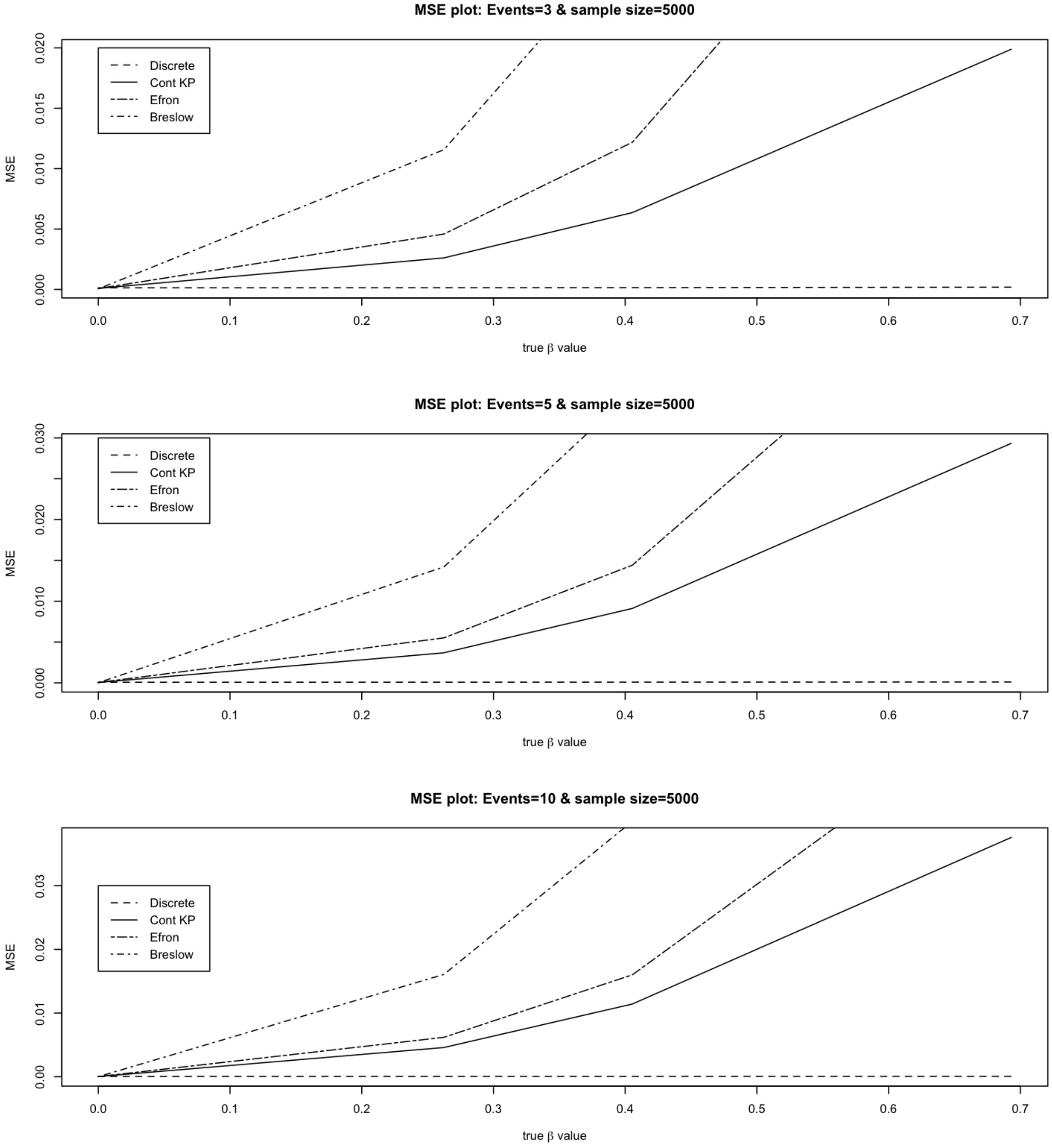

Figure 3.7: Plot of mean squared error against the true β value for the Breslow, Efron, KP, and discrete estimation method. $n = 5000$

3.4 Illustration: Air Pollution Study

Here the resulting inference when the four estimation methods considered in this chapter are applied to the air pollution study introduced in Chapter 1 are considered. As previously noted, the outcome event is exacerbated asthma requiring a hospital encounter, and the covariates of interest are ambient traffic-related air pollution exposures including carbon oxide (CO) and fine particulate matter ($\mathrm{PM}_{2.5}$).

Briefly, the data are comprised of n=7,751 children who made 11,394 visits to the hospital emergency room (Children's Hospital of Orange County or University of California, at Irvine's Medical Center) for asthma related issues between the start of the year 2000 and the end of 2008. 1,893 of these children experienced the event of interest at least twice, and the range of the number of events for a child varied from 1 to 17 events over the course of the study. The covariates of interest were recorded at four central site locations spread across Orange County. Exposure observations contain daily measurements of the environmental exposure factors of interest. Hospital admissions data were abstracted to obtain each patient's date of hospital admission for each visit and the patient's residency zipcode, along with their age, sex, insurance status and other socio-economic factors. Given a subject's date of event and their home zipcode, they were assigned to the nearest exposure monitoring station, which was then used to obtain exposures for their case and control values.

The referent selection scheme is precisely like the one mentioned in the simulation study. Once the control index day is chosen using the semi-symmetric bidirectional scheme, moving averages of 7 days (index day and previous 6 days) for both the case and control exposures were calculated. Additionally, the data were stratified based on whether the admission date was in the cold season (winter and fall, defined to between the months November and April) or the warm season (summer and spring, defined to be between the months May-October). Associations between environmental exposures and adverse health effects have the potential to vary depending on season (Chang et al. [2009]). Also, this is done with the hope that the effect of breaking the individual matched case-control pairing bonds will not be a significant issue. Additionally to alleviate any potential issues of breaking the bonds, each model is also adjusted for relative humidity and temperature.

Table 3.8 presents output using cold season data with the covariate of interest CO computed as a moving average of 7 days. The model includes adjustment covariates of relative humidity and outside temperature with the same moving average of the predictor of interest. The event odds ratio estimate is comparing the odds of an event for being in the top 90th percentile of the exposure to the odds of an event for being in the bottom 10th percentile of the exposure (a change of 1 unit of measurement for CO). Since most statistical software defaults to the Breslow method in the presence of numerous matched pairs within matched set (i.e. tied event times) and that the discrete method is the theoretically appropriate approach to handle case-crossover data with numerous events per subject, the focus will be on comparing the output from the Breslow method to the discrete method.

Method	**CO MA7 Est.**	**Std. Error**	**OR Est.**
Discrete	0.1308	0.0608	1.14
Cont. KP	0.1062	0.0533	1.11
Efron	0.0991	0.0518	1.10
Breslow	0.0876	0.0517	1.09

Figure 3.8: Event odds ratio estimates comparing a change in CO from the bottom 10% to top 90%. Using cold season data.

From Table 3.8, comparing the discrete to the Breslow approximation, it can be seen that the discrete method yields an estimate of the OR associated with CO that is 1.55 times that of the Breslow approxiation (a 9% increase compared to a 14% increase, which constitutes a 55% increase). This difference is substantial in the context of daily exposures to traffic related pollutants. The results suggests that choosing one of the approximation methods can result in substantial underestimation of the effect of the exposure on the risk of an event.

As was shown in Figure 3.2, Figure 3.3, and Figure 3.4, there is an ordering among the methods in terms of parameter estimates being biased towards the null value of 0 when the true parameter value differs from 0. Table 3.8 shows that parameter estimates are attenuated towards 0 as one goes from the discrete method to the KP method to the Efron method and to the Breslow method, with the Breslow method having the most attenuated parameter estimates.

Method	**$PM_{2.5}$ MA7 Est.**	**Std. Error**	**OR Est.**
Discrete	0.0077	0.0033	1.20
Cont. KP	0.0065	0.0030	1.16
Efron	0.0055	0.0026	1.13
Breslow	0.0045	0.0026	1.11

Figure 3.9: Event odds ratio estimates comparing a change in $PM_{2.5}$ from the bottom 10% to top 90%. Using cold season data and only subjects with more than 1 event.

To further highlight the issue of the approximation methods attenuating parameter estimates towards 0, an analysis was conducted using only subjects who had more than 1 event experienced. The data is stratified using cold season observations only. The predictor of interest is $PM_{2.5}$ summarized by a moving average of 7 days. The output for this analysis is provided in Table 3.9. Again, the model includes adjustment covariates of relative humidity and outside temperature with the same moving average of the predictor of interest and the odds ratio estimate compares the change in odds of an event going from the bottom 10% to the top 90% of exposure values (a change of 23 units of measurement for $PM_{2.5}$).

As can be seen from Table 3.9, the Breslow method yields an odds ratio estimate of 1.11 while the discrete estimate is 1.20. In the $PM_{2.5}$ case, comparing the discrete and Breslow, methods, the discrete method has a change in estimated OR that is 1.81 times that of the Breslow method (20% increase compared to a 11% increase, which constitutes an 81% increase).

Similar to the simulation output, the Breslow approximation method returns a parameter estimate that is roughly 45% below of what the discrete method estimate is, the Efron method returns a parameter estimate that is about 28% below the discrete method estimate, and the KP method returns a parameter estimate that is roughly 16% below the discrete method estimate. Considering that these are odds ratio increases for daily exposures, there is substantial difference between the methods both in the mathematical and scientific sense.

In summary, application to the asthma-related hospital encounter data demonstrates that the choice of method for handling ties in the partial likelihood can result in substantially biased inference on the effect of a covariate on the risk of an event of interest.

3.5 Discussion

Environmental exposure studies using case-crossover designs, such as air pollution studies, aim to study the association between an event outcome and an exposure by comparing the exposure distributions between cases and controls. The real world impact of such studies can be far reaching as policy making can potentially hinge on the inferences made in these studies, as the goal of such types of studies is to estimate the effect of an exposure on the risk of an adverse health event. It is crucial that the appropriate methods are utilized when analyzing data from a case-crossover design in order to ensure reliable parameter estimates. In this chapter, it was shown that the conditional logistic likelihood with numerous matched pairs per matched set (subject) is mathematically equivalent to the Cox PH partial likelihood with tied event times within strata. As a result, parameter estimates for case-crossover data with numerous events within subjects are obtained by fitting a Cox PH model to a transformation of the data. Four commonly used approaches for computing Cox's partial likelihood in the presence of tied event times were discussed and shown to result in varying degrees of bias in resulting parameter estimates.

In the presence of tied event times within strata (numerous events within subjects), most software will default to an approximation method to reduce the burden of computing the discrete Cox PH partial likelihood. This is done because in most time-to-event data it is rare to have a large number of tied event times where numerous events occur. As a result, these approximation methods will generally yield estimates with low bias. However, in the case-crossover design where subjects may experience multiple events, each strata will contain a large number of tied observations relative to the strata size.

It was demonstrated that the trade-off made by reducing computational complexity through the use of an approximation method comes at the cost of having parameter estimates that are attenuated towards zero. If the true parameter value is zero, then there is no issue and all methods behave the same way in terms of obtaining parameter estimates. However, if the true parameter value differs greatly from zero, then it was shown that the bias will increase from method to method, with the approximation methods having the most bias while the discrete partial likelihood yields parameter estimates with relatively no bias. Additionally, if each strata has only a single event, then all methods will produce identical results since no "ties" will be present in the partial likelihood in this case.

3.6 Appendix

3.6.1 Computation of the Discrete Likelihood

Letting $r_{(j)}$ be the size of the risk set $R_{(j)}$, the primary challenge in evaluating the partial likelihood in (3.3) lies in computing $B(d_{(j)}, r_{(j)}) = \sum\limits_{I \in (R_{(j)}; d_{(j)})} \prod\limits_{i \in I} exp\{\boldsymbol{x_i \beta}\}$ which requires $\frac{r_{(j)}!}{(r_{(j)}-d_{(j)})!(d_{(j)}-1)!} - 1$ arithmetic operations, that reduce to $2d_{(j)}(r_{(j)} - d_{(j)} + 1)$ operations using recursion. (Gail et al. [1981]) consider a recursive approach to obtain

$$B(d_{(j)}, r_{(j)}) = B(d_{(j)}, r_{(j)} - 1) + \exp\{\boldsymbol{x_{r_j} \beta}\} B(d_{(j)} - 1, r_{(j)} - 1).$$

Using $B(d_{(j)}, r_{(j)})$, each subject's likelihood contribution can be calculated, and numerical derivatives using symmetric differentiation can be utilized to get the score and the information matrix. Let $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and $h \approx 0$. Then the score equations are given by

$$U(\beta_1) = \frac{\partial L}{\partial \beta_1} = \frac{L(\beta_1 + h, \beta_2) - L(\beta_1 - h, \beta_2)}{2h} = 0$$

$$U(\beta_2) = \frac{\partial L}{\partial \beta_2} = \frac{L(\beta_1, \beta_2 + h) - L(\beta_1, \beta_2 - h)}{2h} = 0.$$

In addition, the second derivatives used to obtain the information matrix are:

$$\frac{\partial^2 L}{\partial \beta_1^2} = \frac{L(\beta_1 + h, \beta_2) + L(\beta_1 - h, \beta_2) - 2L(\beta_1, \beta_2)}{h^2}$$

$$\frac{\partial^2 L}{\partial \beta_2^2} = \frac{L(\beta_1 \beta_2 + h) + L(\beta_1, \beta_2 - h) - 2L(\beta_1, \beta_2)}{h^2}$$

$$\frac{\partial^2 L}{\partial \beta_1 \partial \beta_2} = \frac{L(\beta_1 + h, \beta_2) + L(\beta_1, \beta_2 - h) - L(\beta_1 + h, \beta_2 - h) - L(\beta_1, \beta_2)}{h^2}.$$

Note that as $h \to 0$ the definition of the first and second derivative are obtained.

3.6.2 Integral Representation of the Kalbfleisch-Prentice Likelihood

DeLong et al. [1994] present an integral representation to the KP likelihood shown in (3.6). This likelihood assumes there is a true ordering in the event times, and constructs the likelihood permuting all possible orderings of the tied event times.

Let $t_1 < t_2 < ... < t_k$ denote k ordered distinct event times. Let $w_l = \exp(\boldsymbol{x_{il}\beta})$, R_j be the risk set just before t_j, D_j be the set that fail at t_j and $R_j^* = R_j / D_j$.

Under the proportional hazard model, $S_l(t) = S(t)^{w_l}$, where $S(.)$ is the unknown baseline survival distribution function and S_l is the survival function of survival time T_l of the l^{th} unit for the population. Let $\lambda_{ik} = \frac{w_k}{\sum\limits_{l \in R_i^*} w_l}$.

The probability that all units in D_i fail before those in R_i^* is given by

$$P_i = p(\max_{l \in D_i} T_l < \min_{l \in R_i^*} T_l).$$

In addition, the cumulative distribution function of $\max\limits_{l \in D_j} T_l$ is given by

$$G(t) = \prod_{l \in D_i} \{1 - S_l(t)\} = \prod_{l \in D_i} \{1 - S(t)^{w_l}\},$$

and the cumulative distribution function of $\min\limits_{l \in R_j^*} T_l$ is given by

$$F(t) = 1 - \prod_{l \in R_j^*} S_l(t) = 1 - \prod_{l \in R_j^*} S(t)^{w_l}.$$

Using the convolution

$$P_i = \int\limits_0^\infty G(t)dF(t) = \int\limits_0^\infty \prod_{k \in D_i} \{1 - \exp(-\lambda_{ik}t)\}\exp(-t)dt,$$

Therefore

$$\frac{\partial^2 L_g}{\partial\lambda_{gm}\partial\lambda_{gn}} = \begin{cases} \int\limits_0^\infty \prod\limits_{k \in D_g, i \neq m, j \neq n} \{1 - \exp(-\lambda_{gk}t)\}\exp(-t)t^2 \exp(-(\lambda_{gm} + \lambda_{gn})t)dt & \text{if } n \neq m \\ -\int\limits_0^\infty \prod\limits_{k \in D_g,, i \neq m} \{1 - \exp(-\lambda_{gk}t)\}\exp(-t)t^2 \exp(-(\lambda_{gm}t)dt & \text{if } n = m. \end{cases}$$

Then the chain rule can be implemented to evaluate $\frac{\partial\lambda_{gm}}{\partial\beta_k}$ and $\frac{\partial^2\lambda gm}{\partial\beta_k\partial\beta_l}$ to obtain $\frac{\partial L_g}{\partial\beta_k}$ and $\frac{\partial^2 L_g}{\partial\beta_k\partial\beta_l}$ respectively.

Now, suppose subject i has d_i many events and risk set R_i. Thus $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_{d_i})$ and so $\frac{\partial\boldsymbol{\lambda}}{\partial\beta_k} = (\frac{\partial\lambda_1}{\partial\beta_k}, ..., \frac{\partial\lambda_{d_i}}{\partial\beta_k})'$ a $d_i \times 1$ vector. Also $\frac{\partial^2\boldsymbol{\lambda}}{\partial\beta_k\partial\beta_l} = (\frac{\partial^2\lambda_1}{\partial\beta_k\partial\beta_l}, ..., \frac{\partial^2\lambda_{d_i}}{\partial\beta_k\partial\beta_l})'$ another $d_i \times 1$ vector.

Now note that $\frac{\partial L}{\partial\beta_k} = \frac{\partial L}{\partial\boldsymbol{\lambda}}\frac{\partial\boldsymbol{\lambda}}{\partial\beta_k}$ and $\frac{\partial^2 L}{\partial\beta_k\partial\beta_l} = \frac{\partial}{\partial\beta_l}\frac{\partial L}{\partial\boldsymbol{\lambda}}\frac{\partial\boldsymbol{\lambda}}{\partial\beta_k} = \frac{\partial L}{\partial\boldsymbol{\lambda}}\frac{\partial^2\boldsymbol{\lambda}}{\partial\beta_k\partial\beta_l} + \frac{\partial\boldsymbol{\lambda}}{\partial\beta_j}\frac{\partial^2 L}{\partial\boldsymbol{\lambda}^2}\frac{\partial\boldsymbol{\lambda}}{\partial\beta_k}$ where $\frac{\partial^2 L}{\partial\boldsymbol{\lambda}^2}$ is a $d \times d$ square matrix.

Noting that $l = \log(L)$, $\frac{\partial l}{\partial\beta} = \frac{1}{L}\frac{\partial L}{\partial\beta}$, and $\frac{\partial^2 l}{\partial\beta_k\beta_l} = -\frac{1}{L^2}\frac{\partial l}{\partial\beta_k}\frac{\partial l}{\partial\beta_l} + \frac{1}{L}\frac{\partial^2 L}{\partial\beta_k\partial\beta_l}$, the score equations and information matrix can be computed.

Chapter 4

On Frequentist Parameter Estimation of Matched Case-Control Studies with Unbalanced Cluster Sizes

4.1 Introduction

In a wide range of epidemiologic studies, the aim is to investigate the effect of an exposure on the risk of a rare event. When the incidence of the event of interest is low, a prospective cohort design is infeasible, as only a relatively few cases will be observed in a given sample over the course of a study. To study a rare event, a more feasible study design is the case-control design. In this design, a pre-determined number of cases and controls are selected from the target population, and the exposure levels of the sample are retrospectively measured. Once the exposures for the cases and controls are determined, the distribution of the exposure is compared between the case sample and the control sample. In order to control for confounding covariates without having to explicitly include them in the model (thus avoiding potential misspecification of functional form reducing the number of parameters to be estimated), matching controls to cases based on a specified criteria is suggested (Pike and Morrow [1970] for binary exposures and Breslow et al. [1978b] for continuous exposures).

The difficulty with adjusting for confounders is to ensure no confounding factor is left unaccounted for, whether by matching controls to cases or by including them as factors in the model. If a confounder is left unaccounted for, either unmeasured by design or omitted from the model, then the estimation procedure runs the risk of yielding biased estimates for the association of interest. For example, if the case subject is a child admitted to the hospital for an adverse health outcome, and the control is a healthy child of similar age and gender, but not admitted to the hospital, confounding socio-economic factors such as insurance status will be left unaccounted for by design and should be included in the model presuming such data were collected. If it is believed controls are appropriately matched to cases but in fact mis-matching is occurring, these unaccounted confounding factors will not be included in the model, and therefore will not be adjusted for by the model nor by design.

The case-crossover design lends itself naturally to a matched case-control study, as the control subject is the same subject as the case. In this manner, time in-variant within subject confounders are controlled for by design. In a case-crossover design, a subject's control exposures are the exposures experienced by the subject in a time period where no event was experienced (Navidi [1998]). However, not all matched case-control studies can utilize the case-crossover design. The case-crossover design was developed to study the effects of transient, short-term exposures on the risk of acute events (Maclure [1991]). When it comes to environmental air exposure studies, it is believed that the effect of exposure to environmental factors, for example particulate matter 2.5 and ozone, on the risk of an adverse health outcome are transient and therefore the case-crossover design can be utilized (Chang et al. [2009]). The specifics of choosing controls properly given a case index day in order to ensure unbiased estimates were covered in Section 2.3.3 and Section 3.1.

4.2 Unbalanced Cluster Sizes in Matched Case-Control Studies

Studies that utilize the case-crossover design commonly have events of interest that can be experienced numerous times by a subject, such as gout attacks as a result of alcohol consumption (Zhang et al. [2006]) or falls in the elderly as a result of medication changes (Neutel et al. [2002]). Therefore it is likely that the number of events varies across subjects, with some subjects experiencing the event with low frequency and others experiencing the event with much higher frequency. It is reasonable to think that the effect of an exposure is not constant across subjects/clusters, and that the cluster size could possibly be informative of the effect modification. Let i denote the subject and j denote the index of the j^{th} observation, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_i$. If Y_{ij} is the j^{th} binary response for subject i, and $\boldsymbol{X_{ij}}$ are the covariate values for this observation, non-ignorable cluster size is defined to be any violation of the property $\mathrm{E}(Y_{ij}|m_i, \boldsymbol{X_{ij}}) = \mathrm{E}(Y_{ij}|\boldsymbol{X_{ij}})$ (Hoffman et al. [2001]).

When effect modification is present across clusters, subject-specific measures of effect can be included in the model (Laird and Ware [1982]). This approach requires the specification of the covariance matrix for within-cluster observations, and mis-specification can lead to invalid inference. Marginal approaches that utilize generalized estimating equations (GEE) can be used to obtain the marginal effect of an exposure on risk of an event (Zeger and Liang [1986], Liang and Zeger [1986]). These methods based on GEE implicitly presume the cluster size is unrelated to the parameter being estimated, which is to assume ignorable cluster sizes. Within-cluster and between-cluster covariate effects being equal is implied under this setting.

An illustrative applied example relates to a study of the effect of exposure to air pollution on the risk of experiencing exacerbated asthma which will be used as an example in a later section and was previously introduced in Chapter 1 (Delfino et al. [2014]). In this study, the number of events across subjects varies from 1 to 17. It can be postulated that the subjects that are experiencing a higher number of events are doing so because they are more susceptible to changes in the exposure, and therefore have a larger positive value for the coefficient of the exposure on the risk of an event than the subjects who have only a few events. Estimation methods based on GEE can account for the within cluster correlation. One possibility is to treat all observations as independent during parameter estimation, but then account for clustering in the estimation of model variances via the robust variance estimator. However, this estimation procedure will proportionally weight each cluster's contribution to the overall marginal parameter estimate by the precision within the cluster. Clusters that have a large number of observations, with relatively low correlation among the covariates, will have high precision compared to clusters with a low number of observations. As a result, the marginal parameter estimate will be attenuated towards that of the larger clusters estimates.

As an example, assume a case-crossover study with 1:1 matching of controls to cases and that there are 2 subjects in the study. Furthermore, assume the 1st subject contributes a single event (and a single control matched to it), and the 2nd subject contributes 9 events (with a single control matched to each of the events) to the dataset. In terms of contribution to the data to be analyzed, the 2nd subject contributes 90% of the observations and the 1st subject only 10%. Alternatively, one can weight each cluster equally, regardless of the size. If the goal is to make inference that addresses each subject equally, such as implementing policy to improve overall health of the populations represented by the sample, using estimation methods that account for within cluster correlation in the estimation procedure will result in inferences that are heavily biased towards the 2nd subject. If the goal is to make inference that addresses each observation equally, such as implementing policy to address the cost of treating events, using methods that weight each subject equally will not give enough weight to subjects that are a larger portion of the costs (by having more events than other subjects).

Neuhaus and Kalbfleisch [1998] suggest estimating both the within-cluster and between-cluster covariate effects. In this approach, the covariates $\boldsymbol{X_{ij}}$ have two components, one is the within-cluster component $\boldsymbol{X_{ij}} - \boldsymbol{\bar{X}_{i.}}$ and the between-cluster component $\sum_{j=1}^{n_i} \boldsymbol{X_{ij}} = \boldsymbol{\bar{X}_{i.}}$ Neuhaus and Kalbfleisch do not assume the coefficients of these terms are equal. As a result, both terms are included in the model as factors. Kim et al. [2011] develop a semi-parametric regression model for detecting effect modification across the matched sets (subjects) in a case-crossover study, but do so with respect to how the covariate coefficient changes across the matched sets as a result of the matching factors. Thus, in this approach, the matching factors that are believed to be influencing the differences in the effect of the exposure across subjects must be explicitly defined and measured. The approach then proceeds to include interactions between the exposure and the matching factor of interest into the exponentiated linear predictors of the likelihood contribution for each matched set.

The methods previously discussed in Chapter 3 combine all the matched pairs of case-controls from within a subject to create a single set of observations for that cluster, which will denote the risk set for that subject. This results in breaking the bond between each individually matched case-control pair for each event within a subject. Revisiting the example presented in Chapter 3, assume a subject has two events, one in the summer months and one in the winter months. Using the semi-symmetric bi-directional referent selection scheme discussed in Section 2.3.3 and Section 3.1, the summer event will have a control matched to it that will also be from summer and similarly for the winter event's control. When using methods that account for within-cluster correlation by analyzing the data at the cluster level, the two event exposures and the two control exposures will be combined to form a single risk set for that subject. The inherent bond between each matched case-control pair is broken, and it is no longer apparent which control exposure value was matched to each of the events.

In Chapter 2 Section 2.3 and Chapter 3 Section 3.2, it was discussed that the conditional logistic provides one commonly used approach for obtaining parameter estimates in a matched case-control design (Breslow and Day [1980]). Continuing the above example of a subject with two events, let $\boldsymbol{Y_i} = (Y_{i1}, ..., Y_{i4})$ denote the vector of binary outcomes, and $\boldsymbol{x_{ij}} = (x_{ij1}, \dots, x_{ijp})$ denote the p many covariate values for each of the j observations, $j = 1, 2, 3, 4$. The conditional logistic likelihood contribution ,L_i , for this subject is given by

$$L_i = \begin{cases} = \dfrac{\exp\{\sum_j \boldsymbol{\beta x_{ij}} y_{ij}\}}{\sum_{\{\vec{y}_i^*:\sum_j y_{ij}^*=2\}} \exp\{\sum_j \boldsymbol{\beta x_{ij}} y_{ij}^*\}} & \text{, if } \sum_j y_{ij} = 2 \\ = 0 & \text{, otherwise} \end{cases} . \tag{4.1}$$

Intuitively, (4.1) is comparing the covariate values for each of the events to the covariate values among the risk set, which include all the events and the controls matched to each event. When all events and their matched controls are combined to form a single risk set, an event's exposure from summer will be compared to the entire risk set, which will contain the matched pair from winter (and similarly for the winter event exposure being compared to the matched pair from summer). If there exists a seasonality trend in the exposure series, this could potentially lead to biased estimates being obtained and inaccurate inferences being made. If the bond is maintained, then the event exposure value will be compared to the risk set attributed to that event, which will only include the event itself and the matched control.

In the current chapter, three methods that can be readily implemented with standard software in order to analyze case-crossover data with unbalanced cluster sizes are explored. The first is what was termed the discrete method from Chapter 3, which uses the conditional logistic likelihood shown in (4.1) and was discussed in Section 3.2. In Chapter 3, it was shown that this method is the appropriate method for obtaining approximately unbiased parameter estimates in a case-crossover design with numerous events per subject, when the goal is to account for the correlation among the data in the estimation procedure and it is deemed reasonable to break the bond between matched cases and controls. The second is a working independence approach that assumes independence between clusters and among observations within cluster for parameter estimation, then corrects variance estimates post-hoc using a robust variance estimator. The third method is a within cluster resampling scheme (Hoffman et al. [2001]) which samples a single matched pair from each subject to create a subsampled dataset (of independent observations) that is used to obtain parameter estimates. Resampling and estimation is repeated multiple times and the resulting parameter estimates are then averaged to obtain a single marginal estimate.

The operating characteristics of each method are explored, and the advantages and disadvantages of each method are highlighted under the scenario of heavily unbalanced cluster sizes with and without effect modification across clusters. Simulations will show that the discrete and working independence methods will result in parameter estimates that are heavily attenuated towards the coefficient value of the clusters with larger number of events, and the within-cluster resampling method will result in parameter estimates that assigns equal weight to each cluster, regardless of its size. An illustrative example using hospital admission data from a study of air pollution on the risk of experiencing an event of exacerbated asthma (Delfino et al. [2014]) is used to show that the methods under investigation will return substantially different parameter estimates and inferences.

4.3 Methodology

As noted, the conditional logistic regression provides one commonly used approach for obtaining parameter estimates in a matched case-control design. Breslow and Day [1980] derived the conditional logistic likelihood under stratified binary outcome data. Assume a $1 : M$ matching scheme and suppose that d_i events are observed for subject i, $i = 1, \ldots, n$. Under the conditional logistic regression model the probability of an event occurring for j^{th}

observation within subject i, is given by

$$\pi_{ij} = \frac{e^{\beta_{0i} + \boldsymbol{\beta x_{ij}}}}{1 + e^{\beta_{0i} + \boldsymbol{\beta x_{ij}}}}, \ j = 1, ..., (M+1)d_i, \ i = 1, \ldots, n. \tag{4.2}$$

Conditioning on the number of events known to happen in each cluster/subject , the random intercepts β_{0i} in (4.2) are eliminated from the conditional likelihood, and as a result are treated as nuisance parameters. It was shown in Chapter 3 Section 3.2 and Section 3.2.2 that the conditional logistic likelihood is mathematically equivalent to the Cox's partial likelihood if all the event times within a subject are set to be equal, and all control times set to be greater than or equal to the event times.

4.3.1 Discrete likelihood method

The discrete method proceeds with parameter estimation by using Cox's partial likelihood under the assumption of tied event times. Let D_i be the set of the d_i many events for subject i $(i = 1, 2, \ldots, n)$, and R_i be this subjects risk set, which is all the events and their matched controls. Additionally let $(R_i, \, d_i)$ be all possible sets of R_i of size d_i. The conditional likelihood is of the form:

$$L_D = \prod_{i=1}^{n} \frac{\prod\limits_{j \in D_i} \exp\{\boldsymbol{\beta x_{ij}}\}}{\sum\limits_{J \in (R_i, d_i)} \prod\limits_{j \in J} \exp\{\boldsymbol{\beta x_{ij}}\}}. \tag{4.3}$$

As mentioned earlier, the inherent bond between each case-control matched pair is now broken, as all case exposure values and control values are combined to form a single risk set, R. This method is computationally intensive due the combinatoric in the denominator. Computation of the denominator is facilitated by the use of a recursion formula (Gail et al. [1981]). Parameter estimates are obtained by maximizing the likelihood (4.3) using numerical derivatives as shown in the appendix of Chapter 3. The discrete method will proportionately weight each cluster/subject's contribution to the estimation of a (marginal) parameter by the precision within the cluster. Clusters with a large number of observations, and low correlation among its covariates, will tend to contribute more weight to the overall marginal parameter estimate than clusters with few observations, or highly correlated covariate values. In Chapter 3 it was shown that the discrete method is the proper method to use when obtaining parameter estimates in a case-crossover study when the goal is to maintain the clustering of the data and account for the correlation in the data in the parameter estimation procedure.

4.3.2 Working Independence Likelihood Method

A method that does not require the intensive combinatorial computation required for computing the denominator in the likelihood (4.3) is considered. The approach proceeds with estimation by treating the data as independent across all matched pairs. Thus it ignores the clustering of observations based on subject, and treats each matched pair independently. Let d_i denote the number of events observed for subject i and assume the matching of M many controls to each event ($1 : M$ matching). Additionally, let $y_{ijl} = 0, 1$ be the indicator for an event for the l^{th} observation ($l = 1, 2, \ldots, M+1$) within the j-th matched pair ($j = 1, 2, \ldots, d_i$) for subject i, and $\boldsymbol{x_{ijl}}$ be the covariate values for the l-th observation in the j-th matched set for subject i. The working indepence likelihood is then given by

$$L_I = \prod_{i=1}^{n}\prod_{j=1}^{d_i} \frac{\exp(\sum\limits_{l=1}^{M+1} y_{ijl}\boldsymbol{\beta x_{ijl}})}{\sum\limits_{l=1}^{M+1}\exp(\boldsymbol{\beta x_{ijl}})}. \tag{4.4}$$

It can be seen that likelihood in (4.4) does not break the bond between the matched pairs of case-controls. Each event's covariate values are compared strictly to that of the event's risk set, which now only includes the event itself and the controls matched to it. However, the correlation among the clustered observations will not be accounted for in the parameter estimation procedure. Treating correlated outcomes as independent will tend to produce inconsistent standard errors estimates, resulting in invalid inference. To address this, a post-hoc sandwich variance estimator can be used to obtain robust standard errors (Huber [1967], Lin and Wei [1989]).

The sandwich variance estimator in the context of the current problem is introduced. First note that the likelihood in (4.4) can be written as

$$L_I = \prod_{i=1}^{n}\prod_{j=1}^{d_i}\prod_{l=1}^{M+1} y_{ijl}\left[\frac{\exp(\boldsymbol{x_{ijl}\beta})}{\sum\limits_{l=1}^{M+1}\exp(\boldsymbol{x_{ijl}\beta})}\right]. \tag{4.5}$$

Then from (4.5), the score equation is given by

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{d_i}\sum_{l=1}^{M+1} y_{ijl}\left[\boldsymbol{x_{ijl}} - \frac{S^1_{ijl}}{S^0_{ijl}}\right], \tag{4.6}$$

where $S^0_{ijl} = \sum\limits_{l=1}^{M+1}\exp(\boldsymbol{x_{ijl}\beta})$ and $S^1_{ijl} = \sum\limits_{l=1}^{M+1} x_{ijl}\exp(\boldsymbol{x_{ijl}\beta})$.

Setting $S^2_{ijl} = \sum\limits_{l=1}^{M+1}\boldsymbol{x_{ijl}}^{\otimes 2}\exp(\boldsymbol{x_{ijl}\beta})$,the observed information matrix is given by

$$I^*(\boldsymbol{\beta}) = \frac{\partial^2\log(L(\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}^2} = \sum_{i=1}^{n}\sum_{j=1}^{d_i}\sum_{l=1}^{M+1} y_{ijl}\left[\frac{S^2_{ijl}}{S^0_{ijl}} - \frac{(S^1_{ijl})^{\otimes 2}}{(S^0_{ijl})^{\otimes 2}}\right],$$

where $a^{\otimes 2} = aa'$.

Now, let

$$W_{ji}(\boldsymbol{\beta}) = \sum_{l=1}^{M+1} y_{ijl}\left[\boldsymbol{x_{ijl}} - \frac{S^1_{ijl}}{S^0_{ijl}}\right] - \sum_{r=1}^{n}\sum_{k=1}^{d_i}\sum_{l=1}^{M+1} y_{ijl}\frac{\exp(\boldsymbol{x_{rjk}\beta})}{S^0_{ijl}}\left[\boldsymbol{x_{ijl}} - \frac{S^1_{ijl}}{S^0_{ijl}}\right].$$

and construct the matrix B as follows:

$$B(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{d_i}\sum_{k=1}^{d_i} W_{ji}(\boldsymbol{\beta})W_{ki}(\boldsymbol{\beta})'. \tag{4.7}$$

Finally, using a first-order Taylor expansion of the score equation in (4.6) Lin and Wei [1989] show that

$$\hat{\boldsymbol{\beta}} \dot{\sim} N(\boldsymbol{\beta}, V),$$

where $V = I^*(\boldsymbol{\beta})^{-1}B(\boldsymbol{\beta})I^*(\boldsymbol{\beta})^{-1}$ and can be consistently estimated with $\hat{V} = I^*(\hat{\boldsymbol{\beta}})^{-1}B(\hat{\boldsymbol{\beta}})I^*(\hat{\boldsymbol{\beta}})^{-1}$. Under correct model specification (not ignoring the clustering of observations) $I^*(\boldsymbol{\beta})^{-1} = B(\boldsymbol{\beta})^{-1}$, and as a result under correct model specification $V = I^*(\boldsymbol{\beta})^{-1}$. Hence, relying on the result of Lin and Wei [1989], the within cluster correlation can be for when estimating the variance of the parameter estimate in a post-hoc fashion by using (4.7) .

4.3.3 Within-cluster resampling

A within-cluster resampling (WCR) scheme for the analysis of clustered data was proposed by Hoffman et al. [2001] when the cluster sizes vary and it is believed that the covariate effects are not constant across clusters. The notion that covariate effects differ across clusters as a result of the differing sizes of the clusters is termed non-ignorable cluster size. This framework, adapted here to a case-crossover design, proceeds by sampling one matched pair (case and matched control) from each cluster with replacement, with equal probability. The resulting subsample of the full data set is then analyzed using any independent method as no clustering within the subsampled data exists. In the case-crossover setting, the independent method used to obtain parameter estimates is the one that maximizes the likelihood (4.3) or (4.4) with $d_i = 1$ for all i. Since only a single matched pair is within each subject, (4.3) and (4.4) are equivalent. This procedure is repeated many times, and an overall parameter estimate, and it's estimated variance is computed from the numerous model fits from the sub-samples. This procedure was termed "multiple outputation" by Follman et al. [2003] as it leaves parts of the data out of the estimation procedure across each iteration. This is essentially the converse of "multiple imputation", where parts of the missing data are repeatedly filled in (Rubin [1996]).

The WCR method is implemented as follows:

1. For each unique cluster (subject), randomly sample one matched pair (case event and its matched control), giving each matched pair equal probability of being selected.
2. Estimate $\hat{\boldsymbol{\beta}}_{(q)}$ with this reduced dataset using the conditional logistic likelihood. Note there is only 1 case and it's matched controls for each subject.
3. Repeat steps (1.) and (2.) Q many times ($q = 1, 2, ..., Q$), storing each $\hat{\boldsymbol{\beta}}_{\boldsymbol{(q)}}$ and $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}_{(q)})$, the estimated model variance of the parameter estimate from fitting the model in step (2.).
4. Compute $\hat{\boldsymbol{\beta}}_{WCR} = \frac{1}{Q}\sum\limits_{q=1}^{Q}\hat{\boldsymbol{\beta}}_{(q)}$ and
   $\hat{V}_{WCR} = \widehat{\text{var}}(\hat{\boldsymbol{\beta}}_{WCR}) = \frac{1}{Q}\sum\limits_{q=1}^{Q}\text{var}(\hat{\boldsymbol{\beta}}_{(q)}) - \frac{1}{Q}\sum\limits_{q=1}^{Q}(\hat{\boldsymbol{\beta}}_{(q)} - \hat{\boldsymbol{\beta}}_{WCR})(\hat{\boldsymbol{\beta}}_{(q)} - \hat{\boldsymbol{\beta}}_{WCR})'$

Next some of the more important properties of the WCR estimator are considered. Let the randomly chosen index for subject i be $J(i)$. Let $\boldsymbol{J} = (J(1), ..., J(n))$ and $\boldsymbol{J}_q$ be the q^{th} outputation. Note that $\boldsymbol{J}$ has $d_1 \times .... \times d_n$ many support points, each equally likely.

Assume 1:1 matching for ease of notation and let
$\underline{\boldsymbol{X}} = [\boldsymbol{X}_{111}, \boldsymbol{X}_{112}, \boldsymbol{X}_{121}, \boldsymbol{X}_{122}, \ldots, \boldsymbol{X}_{1d_11}, \boldsymbol{X}_{1d_i2}, \ldots, \boldsymbol{X}_{n11}, \boldsymbol{X}_{n12} \ldots, \boldsymbol{X}_{nd_n1}, \boldsymbol{X}_{nd_n2}]$
denote the entire data set for all clusters, where $\boldsymbol{X}_{ijl}$ is the covariate vector for the l^{th} observation for the i^{th} subject's j^{th} matched pair ($l = 1, 2$ and $j = 1, \ldots, d_i$). Additionally set $\mathbf{X}(\mathbf{J})$ to be the n data points for an outputation, and let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{X}(\mathbf{J}))$ be the parameter

estimate based on this outputation.

If each subject has d_i multiple matched pairs, define $\hat{\boldsymbol{\beta}}^{\infty}_{WCR} = \frac{1}{\prod\limits_i d_i}\sum\limits_{\mathbf{J}}\hat{\boldsymbol{\beta}}(\mathbf{J})$, the average of the parameter estimates obtained across all possible outputations. Note that since conditional on $\underline{\boldsymbol{X}}$ the $\hat{\boldsymbol{\beta}}(\mathbf{X}(\mathbf{J}))$'s are independent and identically distributed, $\lim\limits_{Q\to\infty}\hat{\boldsymbol{\beta}}_{WCR} = \hat{\boldsymbol{\beta}}^{\infty}_{WCR}$ and $\hat{\boldsymbol{\beta}}^{\infty}_{WCR} = \mathrm{E}_{\mathbf{J}}\left[\hat{\boldsymbol{\beta}}(\mathbf{X}(\mathbf{J}))|\underline{\boldsymbol{X}}\right]$.

Hoffman et al. [2001] use the variance decomposition formula to obtain the variance of $\hat{\boldsymbol{\beta}}^{\infty}_{WCR}$ as

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}(\mathbf{X}(\mathbf{J}))) &= \mathrm{E}(\mathrm{var}[\hat{\beta}(\mathbf{X}(\mathbf{J}))|\underline{\boldsymbol{X}}) + \mathrm{var}(\mathrm{E}[\hat{\boldsymbol{\beta}}(\mathbf{X}(\mathbf{J}))|\underline{\boldsymbol{X}}) \\ \sigma^2 &= \sigma_c^2 + \mathrm{var}(\hat{\boldsymbol{\beta}}^{\infty}_{WCR}),\end{aligned}$$

where σ_c^2 is the variance conditional on upon the outputation set. An estimate of the variance of $\hat{\boldsymbol{\beta}}_{WCR}$ is then given by

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_{WCR}) = \frac{1}{Q}\sum_{q=1}^{Q}\mathrm{var}(\hat{\boldsymbol{\beta}}_{(q)}) - \frac{1}{Q}\sum_{q=1}^{Q}(\hat{\boldsymbol{\beta}}_{(q)} - \hat{\boldsymbol{\beta}}_{WCR})(\hat{\boldsymbol{\beta}}_{(q)} - \hat{\boldsymbol{\beta}}_{WCR})'. \tag{4.8}$$

Hoffman et al. [2001] proved the consistency of this estimate under standard regularity conditions. Further, they showed that for large Q:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{WCR} - \boldsymbol{\beta}) \overset{D}{\to} N(\mathbf{0}, \Sigma),$$

where Σ is consistently estimated by the variance shown in (4.8).

It is worth noting that Williamson et al. [2003] presented the WCR method in a weighted estimating equations (WEE) framework. They observed that since the sampling scheme within subject is essentially a discrete uniform with probability mass of $1/d_i$ for each matched pair, then one can derive a WEE method which assigns weights of $1/d_i$ for each matched pair within subject i's estimating equations. The details of this approach are in the chapter appendix.

The inspiration for drawing the relationship between multiple outputation estimator and the WEE estimator was because of the intensive computation sometimes necessary to yield the multiple outputation estimator, as Q needs to be large. Processing power has increased greatly since and as a result the WCR method only takes marginal more time to compute than the WEE. Since both methods are shown to give similar results, the focus is on the WCR method in terms of simulation studies and applied data illustration.

4.3.4 Method Comparisons

As noted, the working independence approach treats each matched pair as an individual cluster, constructs a likelihood based on the independent observations, and then proceeds to maximize the likelihood with respect to the parameters to obtain estimates. As such, this method inherently assigns higher weight to larger clusters as the clusters with a higher frequency of observations will contribute a larger percentage of the data set used to obtain the parameter estimate.

Return to the hypothetical example mentioned earlier where 2 subjects are in the study and the first one experiences a single event and the second one experiences 9 events. In terms of cluster representation, each subject represents 50% of the clusters, but in terms of the data that the working independence approach will use to obtain estimates, the second subject will account for 90% of the observations. Using the working independence approach will return parameter estimates that are heavily attenuated towards the true parameter value of the second subject. However, this method does not break the inherent bond between each matched pair within subjects, as each matched pair is treated as its own cluster. For a given subjects event, the likelihood contribution for this event will compare the covariate value of the event to only the risk set for that event, which includes the event itself and the control matched to it.

The discrete method assumes independence across clusters, but not within cluster. The product in the likelihood (4.3) is therefore across clusters only. Within each cluster, the event exposures are compared to the exposures of the risk set for that cluster, which will comprise of all events and the controls matched to those events. Within the risk set of a given subject, it is no longer apparent which control was matched to which event and as a result the inherent bond between each matched case-control pair is broken. If there are seasonality trends in the exposure time series, this could lead to potentially biased results.

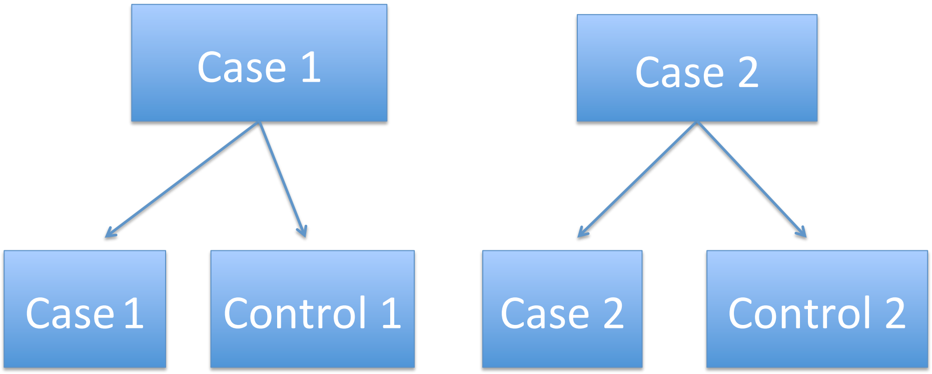

Figure 4.1: Matched case-control pair bonds are maintained.

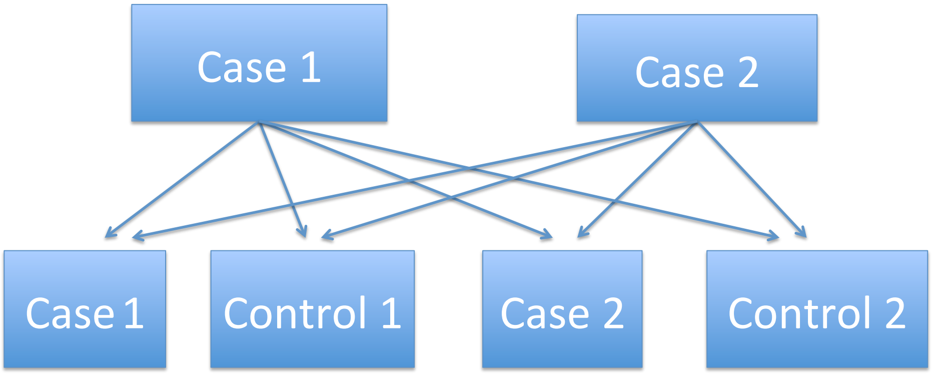

Figure 4.2: Matched case-control pair bonds are not maintained.

Consider the example mentioned earlier of a subject with two events, one in summer and one in winter. The summer event will have a control matched to it that will be from summer, and similarly for the winter event. Cold seasons tend to have higher levels of environmental exposures such as NO_2 and NO_x. By breaking the matched case-control bonds, the summer event exposure will be compared to both the summer and winter matched pair exposures, and likewise for the winter event. Figure 4.1 shows when the bonds are maintained. Each case exposure value is compared to its own risk set (i.e. matched pair). In Figure 4.2, the bonds are no longer maintained, and each case exposure is compared to the risk set of the subject (i.e. both matched pairs).

To try and alleviate the issue of the bonds being broken, the air pollution data is stratified into seasons, cold and warm, and additionally all models are adjusted for temperature and relative humidity. The hope is that breaking the bonds will not be a substantial issue as each events risk set within a subject will contain controls from only the same season as the event itself.

The discrete method will tend to compute parameters that are attenuated towards the coefficient value of the larger clusters similar to the working independence approach, but the contribution of each cluster to the overall marginal parameter estimate is mitigated by the fact that each clusters contribution to the marginal parameter estimate is weighted by the precision of the estimate from within the cluster. If a cluster has highly correlated covariates, then it will contribute less information as compared to a similar cluster with covariates with little to no correlation among the matched pairs.

Finally, the WCR estimator attempts to neutralize the disproportionate attenuation of the marginal parameter estimate to the coefficient parameter of the larger clusters by only sampling a single matched case-control pair from each cluster. At each iteration of the multiple outputation procedure, a dataset is created that contains only a single matched pair from each cluster, thus alleviating the issue of varying cluster sizes since the parameter estimated from one of the multiple outputation iterations will weight each subject equally (as all clusters are the same size). This method does not break the bonds since each matched case-control pair is sampled together. Within each cluster in the subsampled data, there will be only a single event, whose exposure values will be compared to only its risk set, which will be the event itself and its matched control. This method will result in parameter estimates that will weight each subject equally, regardless of the number of observations within each subject. Note that if each subject has only a single event, then the working independence method. the discrete method, and the WCR method are identical.

In the hypothetical example of 2 subjects, the WCR method will tend to compute an estimate that is the equally weighted average of each subjects parameter value. At the same time, since at each iteration of this procedure a single matched pair is randomly sampled within clusters that have numerous pair, all observations will still be used in the estimation procedure (assuming the number of iterations is large compared to the number of pairs within clusters). The precision in the overall marginal parameter gained by using all pairs from clusters is evident in the variance estimate in (4.8).

The estimated variance of WCR parameter estimate is the mean of the model based variance estimates over the Q many outputations, but with the subtraction of a correction term. How to view this correction term is it measures how the parameter estimates varies over the Q many outputations. The less correlated the observations are within a cluster, the more variability the parameter estimate $\hat{\boldsymbol{\beta}}_{(q)}$ will exhibit across the outputations. Thus, the precision of the overall marginal parameter $\hat{\boldsymbol{\beta}}_{WCR}$ will increase due to the information gained by sampling different matched pairs from the clusters with several pairs. The higher the correlation among the observations in a cluster, the smaller the correction term will be as the estimated parameter, $\hat{\boldsymbol{\beta}}_{(q)}$, will vary only slightly across the outputations.

	No modification	Yes modification
No trend	Discrete=Ind.=WCR	Discrete $\approx$ Ind. $\neq$ WCR
Yes trend	Discrete $\neq$ Ind. $=$ WCR	Discrete $\neq$ Ind. $\neq$ WCR

Table 4.1: Comparison of WCR, discrete, and Working Independence (Ind.).

Table 4.1 summarizes the three methods depending upon whether or not seasonality trends are evident in the exposure series (yes/no) and whether or not effect modification across clusters exists (yes/no). If there is a seasonality trend in the exposures, then breaking the bonds between each matched pairs will result in the discrete method computing parameter estimates that will differ than both the multiple outputation and the working independence method. Within the scenario of seasonality trend in exposures, if there is effect modification across clusters, then the working independence method assigns higher weight to the larger clusters, which will have a different parameter estimate than the smaller clusters due to the effect modification. Since the WCR method will assign equal weight to all subjects whereas the working independence method will assign equal weight to the individual matched pairs, the WCR method and the working independence method will obtain different results. If there is no seasonality trend and no effect modification across clusters, all methods will be equivalent. If there is no seasonality trend in the exposures, but there is effect modification, then the discrete method will yield estimates similar to that of the working independence model. The multiple outputation procedure will yield a parameter estimate that will differ from the discrete and working independence methods since it will weight each subject equally, which will prevent the attenuation of the overall parameter estimate towards the parameter value of the larger clusters, unlike the working independence and discrete methods.

4.4 Simulation study

Luo and Sorock [2008] conducted a simulation study in which they investigated the operating characteristics of the methods mentioned in this chapter, but in their simulations cluster sizes varied only slightly (average of two events per cluster), and all clusters had the same true parameter value. As a result, all methods produced similar results. As described in the simulation process in the following section, cluster sizes can vary greatly, as they do in practice with the applied hospital admission data presented in Chapter 1. Here, a variety of parameter values are used to induce effect modification across the clusters (informative cluster sizes) and also scenarios where no effect modification are evident (non informative cluster sizes). It will be shown that the working independence method will obtain parameter estimates that are attenuated heavily towards the coefficient value of the high frequency clusters, β_h. The discrete method will return estimates that are near that of the working independence method, but not as attenuated towards β_h because in this method the high frequency clusters will have their contributions towards the overall marginal parameters reduced due the correlation in the covariates within the cluster. The WCR method will not take cluster size into account, and will have an estimand that assigns equal weight to clusters/subjects regardless of number of observations in the cluster. Parameter estimates will be attenuated towards the true value of the subjects that make up a majority of the subject sample size.

4.4.1 Data simulation

The data generation scheme used for this simulation study is similar to the one described in Section 3.3.1, but with some modifications. For a given sample size of n many subjects, 90% of subjects are set to represent clusters with low frequency observations, and 10% are set to represent clusters with high frequency of observations. To represent a fixed number of monitoring stations, 4 strings of exposures were generated via an auto regressive time series of order 1, AR(1). For $k = 1, 2, 3, 4$ and $t = 1, 2, \ldots, T$ where $T = 3,650$ to signify a 10 year study with daily exposures, an AR(1) process was generated for the exposures, x:

$$x_{kt} = c + \varphi x_{kt-1} + \varepsilon_{kt} \text{ where } \varepsilon_{kt} \overset{iid}{\sim} N(0,1).$$

Each subject, i ($i = 1, 2, \ldots, n$) was assigned be either a low or high frequency cluster with probability of 0.9 and 0.1, respectively and also was assigned to one of the four generated exposure time series with equal probability. If assigned to low frequency, the number of events, v_i, for this subject was sampled from the set $\{1, 2, 3\}$ with probability $p = (0.9, 0.05, 0.05)$.

If assigned to high frequency, number of events for this subject was sampled from a uniform{5,6,...,17}.

For $t = 1, 2, ..., 3650$, the probability of an event on a given day using the logistic link for a prospective probability of event, $\pi_{k_i t} = \frac{exp(\beta_{0i}+\beta_1 x_{k_i t})}{1+exp(\beta_{0i}+\beta_1 x_{k_i t})}$, where k_i is the monitoring center assigned to subject i (1,2,3, or 4) was computed. For the low frequency clusters, $\beta_1 \equiv \beta_l$, and for the high frequency clusters $\beta_1 \equiv \beta_h$. Values used for β_l and β_h will be stated when simulation results are presented. For each π_{it}, a Bernoulli random variable with $p = \pi_{it}$ was sampled. Given the pre-specified number of events for each subject (v_i), v_i many events were randomly sampled from the successes generated from the T many Bernoulli trials. If a subject did not produce enough events via the Bernoulli simulation based on the π_{it}'s as they were set to have (events simulated$< v_i$), then all events that were produced for that subject were chosen.

Once the event days are selected, a control day was chosen for each event based on the suggestions of Navidi and Weinhandl [2002]. For a given event day, the control exposure values were the moving average of the 7 days between the 14-th and 20-th days prior to the event or the moving average of 7 days between the 8-th and 14-th days past the event, with each window being picked with equal probability. The event exposures are set to the be the average of the event day and the previous 6 days. This is done in air pollution studies to account for any lingering effect of the exposure on the risk of an adverse health outcome, and also to be able to incorporate the information from several days into the analysis. For events that had only a single available referent window to be selected, in cases where the event happens in the start or end of the exposure series or if another event is observed in one of the two available referent windows, an offset of log(2) was used (Janes et al. [2005a]). If for a specific event both available referent windows had events observed in them, then this event is dropped from the analysis (this occurred in less than 0.1% of the total events).

The parameters in the time series were specified as $c = 2$ and $\varphi = 0.5$,leading to

$\text{cov}(x_{kt}, x_{kt+r}) = \frac{1}{1-\varphi^2}\varphi^{|r|}$. A variety of different scenarios were simulated. Let β_l denote the true parameter value for the subjects in the low frequency clusters and β_h be the true parameter value for the high frequency clusters. Six different pairs of (β_l, β_h) values were investigated, those respectively being (log(1),log(1)) ,(log(1),log(1.5)) ,(log(1),log(2)), (log(1.2),log(1.2)), (log(1.2),log(1.5)), and (log(1.2),log(2)). The random intercept for each subject was generated according to $\beta_{0i} \overset{iid}{\sim} N(\mu_0, 1)$ where depending on the scenario of (β_l, β_h), μ_0 varied from -7.5 to -6, in order to ensure a sufficient expected number of events occurring across the 3,650 days. Sample sizes of $n = 500, 1,000,$ and $5,000$ were considered.

4.4.2 Simulation results

The means of the parameter estimates, mean of model standard errors, and empirical standard errors of parameter estimates over the simulation runs are given in Table 4.2, Table 4.3, and Table 4.4. 10,000 data sets were simulated for each scenario and $Q = 250$ outputions were used for the multiple outputation algorithm.

According to the 90-10 breakdown of low frequency to high frequency clusters in the simulation explained in the prior section, it is expected that the WCR method will compute estimates close to a subject weighted parameter, θ_s, where

$$\begin{aligned} \theta_s &= P(\text{low}) \times \beta_l + P(\text{high}) \times \beta_h \\ &= 0.9 \times \beta_l + 0.1 \times \beta_h, \end{aligned}$$

and where $P(\text{low})$ and $P(\text{high})$ are the probability of being a low and high frequency cluster, respectively. Additionally let $\text{E}[v_l]$ and $\text{E}[v_h]$ be the expected number of events in the low and high frequency clusters, respectively. The working independence method will estimate a parameter that is weighted on the events, θ_e. Noting that based on the simulation procedure the expected number of events in the low frequency group is 1.15 and for the high frequency it is 11, this parameter amounts to

$$\begin{aligned}
\theta_e &= \frac{\mathrm{E}[v_l] \times P(\text{low}) \times n}{\mathrm{E}[v_l] \times P(\text{low}) \times n + \mathrm{E}[v_h] \times P(\text{high}) \times n}\beta_l \\
&\quad + \frac{\mathrm{E}[v_l] \times P(\text{low}) \times n}{\mathrm{E}[v_h] \times P(\text{high}) \times n) \times n + \mathrm{E}[v_h] \times P(\text{high}) \times n}\beta_h \\
&= \frac{1.15 \times 0.9 \times n}{1.15 \times 0.9 \times n + 11 \times 0.1 \times n}\beta_l + \frac{11 \times 0.1 \times n}{1.15 \times 0.9 \times n + 11 \times 0.1 \times n}\beta_h
\end{aligned}$$

where n denotes the number of subjects.

As mentioned earlier, the parameter estimates obtained from the working independence method will tend to be attenuated towards the coefficient value of the subjects who make up a majority of the observations. This can be seen in tables across all tables of simulation outputs. The discrete method is close to the parameter estimate of the working independence method, but since each clusters contribution to the estimation procedure is weighted proportionally by the precision within that cluster, the parameter estimate from this method is attenuated away from the working independence method towards the coefficient value of the smaller clusters due to the correlation among the observations within clusters. The WCR method will compute a parameter estimate that is attenuated towards the coefficient value of the subjects who make a majority of the subject cohort.

As can be seen, when the coefficient for clusters with high frequency of events and the coefficient for clusters with low frequency of events differs, the working independence method obtains parameter estimates that are attenuated to the high frequency clusters coefficient value and the discrete method's estimate is close to the working independence method but its parameter estimate is attenuated away from the working independence method towards the coefficient value of the low frequency clusters. The WCR method's estimate is always attenuated towards the low frequency clusters (as these clusters make up a majority of the cluster sample size). This is highlighted in the scenario where $\beta_l = 0$ and $\beta_h = 0.693$. The

$\boldsymbol{\beta_l} = \boldsymbol{\beta_h} = \log(\mathbf{1}) = \mathbf{0}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0-0.0004	0.0587	0.0585
Discrete	0.0002	0.0351	0.0337
Working Ind.	-0.0002	0.0431	0.0425
$\boldsymbol{\beta_l} = \log(\mathbf{1}) = \mathbf{0}\,, \boldsymbol{\beta_h} = \log(\mathbf{1.5}) = \mathbf{0.4055}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.0400	0.0586	0.0581
Discrete	0.1792	0.0351	0.0338
Working Ind.	0.2025	0.0461	0.0435
$\boldsymbol{\beta_l} = \log(\mathbf{1}) = \mathbf{0}\,, \boldsymbol{\beta_h} = \log(\mathbf{2}) = \mathbf{0.6931}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.0664	0.0586	0.0581
Discrete	0.2952	0.0351	0.0340
Working Ind.	0.3313	0.0506	0.0452
$\boldsymbol{\beta_l} = \boldsymbol{\beta_h} = \log(\mathbf{1.2}) = \mathbf{0.1823}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.1823	0.0597	0.0608
Discrete	0.1822	0.0352	0.0443
Working Ind.	0.1800	0.0352	0.0344
$\boldsymbol{\beta_l} = \log(\mathbf{1.2}) = \mathbf{0.1823}\,, \boldsymbol{\beta_h} = \log(\mathbf{1.5}) = \mathbf{0.4055}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.2045	0.0602	0.0611
Discrete	0.2394	0.0353	0.0452
Working Ind.	0.2943	0.0458	0.0452
$\boldsymbol{\beta_l} = \log(\mathbf{1.2}) = \mathbf{0.1823}\,, \boldsymbol{\beta_h} = \log(\mathbf{2}) = \mathbf{0.6931}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.2302	0.0607	0.017
Discrete	0.3551	0.03535	0.0338
Working Ind.	0.4281	0.0501	0.0481

Table 4.2: Simulation summaries for within cluster resampling (WCR), discrete, and working independence methods under heavily unbalanced cluster sizes with and without effect modification across clusters. Sample size n=500 subjects

$\boldsymbol{\beta_l = \beta_h} = \log(\mathbf{1}) = \mathbf{0}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.0002	0.0185	0.0187
Discrete	0.0002	0.0111	0.0109
Working Ind.	0.0002	0.0136	0.0137
$\boldsymbol{\beta_l} = \log(\mathbf{1}) = \mathbf{0}\,, \boldsymbol{\beta_h} = \log(\mathbf{1.5}) = \mathbf{0.4055}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.0405	0.0185	0.0182
Discrete	0.1808	0.0111	0.0106
Working Ind.	0.2041	0.0145	0.0136
$\boldsymbol{\beta_l} = \log(\mathbf{1}) = \mathbf{0}\,, \boldsymbol{\beta_h} = \log(\mathbf{2}) = \mathbf{0.6931}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.0664	0.0185	0.0182
Discrete	0.2965	0.0111	0.0104
Working Ind.	0.3325	0.0160	0.0139
$\boldsymbol{\beta_l = \beta_h} = \log(\mathbf{1.2}) = \mathbf{0.1823}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.1823	0.0188	0.0188
Discrete	0.1828	0.0111	0.0107
Working Ind.	0.1823	0.0139	0.0138
$\boldsymbol{\beta_l} = \log(\mathbf{1.2}) = \mathbf{0.1823}\,, \boldsymbol{\beta_h} = \log(\mathbf{1.5}) = \mathbf{0.4055}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.2040	0.0190	0.0190
Discrete	0.2404	0.0111	0.0107
Working Ind.	0.2950	0.0144	0.0143
$\boldsymbol{\beta_l} = \log(\mathbf{1.2}) = \mathbf{0.1823}\,, \boldsymbol{\beta_h} = \log(\mathbf{2}) = \mathbf{0.6931}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.2293	0.0191	0.0188
Discrete	0.3557	0.0112	0.0104
Working Ind.	0.4283	0.0158	0.0146

Table 4.3: Simulation summaries for within cluster resampling (WCR), discrete, and working independence methods under heavily unbalanced cluster sizes with and without effect modification across clusters. Sample size n=1000 subjects

$\boldsymbol{\beta_l = \beta_h = \log(1) = 0}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0-0.0004	0.0131	0.0129
Discrete	-0.0002	0.0078	0.0072
Working Ind.	-0.0002	0.0096	0.0096
$\boldsymbol{\beta_l = \log(1) = 0\,, \beta_h = \log(1.5) = 0.4055}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.0400	0.0131	0.0129
Discrete	0.1805	0.0078	0.0075
Working Ind.	0.2037	0.0103	0.0097
$\boldsymbol{\beta_l = \log(1) = 0\,, \beta_h = \log(2) = 0.6931}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.0656	0.0131	0.0128
Discrete	0.2963	0.0078	0.0073
Working Ind.	0.3321	0.0113	0.0098
$\boldsymbol{\beta_l = \beta_h = \log(1.2) = 0.1823}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.1824	0.0133	0.0132
Discrete	0.1822	0.0078	0.0076
Working Ind.	0.1828	0.0098	0.0098
$\boldsymbol{\beta_l = \log(1.2) = 0.1823\,, \beta_h = \log(1.5) = 0.4055}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.2043	0.0134	0.0132
Discrete	0.2405	0.0078	0.0075
Working Ind.	0.2951	0.0102	0.0100
$\boldsymbol{\beta_l = \log(1.2) = 0.1823\,, \beta_h = \log(2) = 0.6931}$			
	Mean $\hat{\beta}$	**Model SE**	**Emp. SE**
WCR	0.2297	0.0135	0.0131
Discrete	0.4286	0.0112	0.0102
Working Ind.	0.3559	0.0078	0.0072

Table 4.4: Simulation summaries for within cluster resampling (WCR), discrete, and working independence methods under heavily unbalanced cluster sizes with and without effect modification across clusters. Sample size n=5000 subjects

working independence and discrete method have mean parameter value of roughly 0.30 while the WCR method is only about 0.06. When both low and high frequency clusters have the same true coefficient value, as is the case when $\beta_l = \beta_h = 0$ and $\beta_l = \beta_h = 0.1823$, there is virtually no difference among the methods in terms of parameter estimate. The WCR method does have higher standard errors because in each outputation the data set used to compute a parameter estimate is substantially smaller than the whole data of all observations from all subjects and since each clusters does have correlated covariate values, the correction term in (4.8) is not large.

In terms of scientific significance of this simulation study, the estimand of interest needs to be defined prior to conducting the analysis. If there is no effect modification across clusters $((\beta_l = \beta_h)$, then all methods will obtain similar results, given no evidence of seasonality trend in the exposure series. But when cluster sizes vary and effect modification is present $(\beta_l \neq \beta_h))$, the estimands of the different methods discussed earlier in this section will result in substantially different estimates.

To illustrate the difference in estimands, take for example the air pollution study that has been under focus throughout the dissertation, where the predictor of interest is $\mathrm{PM}_{2.5}$ and the response variable is admission to the hospital for exacerbated asthma. If the aim of the analysis is to obtain inference in order to create public policy that will improve the overall health of the target population (by reducing risk of an event with respect to the exposure), then it reasonable to think that the estimand of interest should be the one that weights each subject equally. If it is believed each subject equally represents the target population, then subjects with more events observed should not receive higher weight and have the inference biased towards there effects. If the goal of the study is to obtain inference in order to create policy that will address the cost of treating such events, then it is reasonable that the subjects with higher number of events receive higher weight. Subjects with a large number of events account for a higher proportion of the costs of treatment (compared to subjects with a low number of events), and as a result should have higher weight when obtaining parameter estimates. In this scenario, it would be more appropriate to choose the discrete or working independence method to obtain estimates as one would want the inference procedure to be biased towards subjects with more events observed. Additionally, if seasonality trends are evident in the exposure series, then the discrete method will obtain biased estimates as a result of breaking the matched case-control bonds for subjects with numerous events.

In the simulation study, the exposure series was exchangeable, and as a result the discrete method obtained parameter estimates similar to the working independence method. It will be seen in the following section that in the applied illustrative example that this is not the case. Even though the data set was stratified based on season, and each model was adjusted for relative humidity and temperature, the discrete method obtains estimates that are substantially different from the working independence method. This could be because of the issue of breaking the bonds for subjects with numerous events or that subjects with numerous events have highly correlated observations.

4.5 Illustration: Air Pollution Study

The methods discussed up to now are applied to the asthma-related hospital encounters study discussed in Chapter 1. During the study, $n = 7,751$ children made a total of 11,394 visits to the hospital. The number of events across subjects varied from 1 to 17 with $1,893$ children experiencing the event of interest at least twice. The covariates of interest were recorded at four locations spread across Orange County. Exposure observations contained daily measurements of the environmental exposure factors of interest. The event of interest was a asthma-related hospital encounter. Hospital admissions data contain each subjects date of hospital admission for each visit and the home address zip code, along with subjects age, sex, insurance status and other socio-economic factors. Given a subject's date of event and their home zip code, they are assigned to the nearest exposure monitoring station, which is then used to obtain exposures for their case and control values. For each event day, a referent was selected using the adjusted semi-symmetric bidirectional referent sampling scheme mentioned in the simulation study.

Vines and Farrington [2001] explored the potential for biased estimates from analyzing case-crossover data using the conditional logistic likelihood under a binary exposure. They suggest that the conditional logistic likelihood has the potential to obtain biased estimates unless there is global exchangeability in the exposure series. This is referring to in part to the discussion earlier about the issue of breaking the bonds between the matched pairs within a subject, which is what the discrete method is doing but is not an issue with the other methods. In the simulation study, the exposure series was exchangeable. With the applied data, the observations are stratified into the cold and warm seasons. Along with stratifying the data in terms of season, each model mentioned in this section also controls for relative humidity and temperature of the same moving average as the predictor of interest. The goal in doing so is to obtain global exchangeability of exposures within each stratified dataset. Additionally, it is worth noting that previous environmental exposure studies suggests that the effect of an environmental exposure can vary across season (Delfino et al. [2014]). The cold season is defined as containing events that occurred between the months of November to April, and the warm season is defined as containing events that occurred between the months of May to October.

In the following illustrative examples, all odds ratios are computed for an interquartile change in exposure (comparing the odds of event going from the bottom 25^{th} percentile of an exposure to the top 75^{th} percentile). Additionally the odds ratio confidence intervals are for interquartile changes. All estimates are adjusted for temperature and relative humidity of the same moving average time as the exposure of interest.

$PM_{2.5}$ MA7	**Est.**	**S.E.**	**Odds Ratio 95% C.I.**
WCR	0.0088	0.0030	(1.0416 , 1.2425)
Discrete	0.0083	0.0023	(1.0586 , 1.2118)
Working Independence	0.0123	0.0027	(1.1105 , 1.3015)

Table 4.5: Parameter estimates using cold season data and adjusting for relative humidity and temperature of the same averaging time as the predictor of interest. $PM_{2.5}$ IQR of 15.

O_3 MA7	**Est.**	**S.E.**	**Odds Ratio 95% C.I.**
WCR	0.0143	0.0050	(1.0674 , 1.2306)
Discrete	0.0124	0.0037	(1.0775 , 1.3297)
Working Independence	0.0161	0.0045	(1.1113 , 1.4352)

Table 4.6: Parameter estimates using warm season data and adjusting for relative humidity and temperature of the same averaging time as the predictor of interest. O_3 IQR of 14.5.

$PM_{2.5}$ MA7	**Est.**	**S.E.**	**Odds Ratio 95% C.I.**
WCR	0.0043	0.0040	(0.9483 , 1.2000)
Discrete	0.0056	0.0030	(0.9223 , 1.2823)
Working Independence	0.0091	0.0036	(1.0312 , 1.2742)

Table 4.7: Parameter estimates using cold season data and only subjects living in zipcodes with population levels below the median level for Orange county and adjusting for relative humidity and temperature of the same averaging time as the predictor of interest. $PM_{2.5}$ IQR of 15.

The results presented in Table 4.5, Table 4.6, and Table 4.7 show that the parameter estimate and inference of the effect of an exposure on the risk of an event of experiencing exacerbated asthma differs among the methods mentioned in this chapter. The data stratification used is mentioned in each table. Focusing on Table 4.5, the estimated odds ratio for an IQR change in $PM_{2.5}$ in the cold season for the WCR method is 1.14 while for the working independence method it is 1.20. This signifies that the working independence method's estimated increase in odds for an IQR change in $PM_{2.5}$ is 43% higher than the estimated increase in odds for the WCR method (comparing a %14 increase to a %20 increase). Looking at Table 4.6, the discrete method's estimate is the smallest of all methods. The discrete method's estimated odds ratio for an IQR increase if O_3 is 1.19 while for the working independent method it is 1.26. This implies the working independent method's estimated increase in odds for an

IQR change is 26% higher than the increase estimated with the discrete method. Table 4.7 uses data that is further stratified to only subjects living in areas with population levels below the median level for Orange County (rural areas). Output shows that the working independence method is the only one that returns a 95% confidence interval for the odds ratio for an IQR change in $PM_{2.5}$ in the cold season and among subjects that live below the median population level that does not contain 1 (thus the other methods would fail to reject the null hypothesis of equal odds based on a significance level of 5%).

As can be seen, the discrete method yields estimates that are sometimes heavily attenuated away from the working independence method's estimates, such as in Table 4.7. This could be a result of bias caused by breaking the bonds between matched pairs for subjects with numerous pairs or that observations among subjects with numerous events are highly correlated. As a result, this next example will only use the working independence and WCR method. Since both these methods maintain the matched case-control bond, there is no need to stratify the data based on season.

$PM_{2.5}$ MA7	**Est.**	**S.E.**	**Odds Ratio 95% C.I.**
WCR	0.0045	0.0033	(0.9710,1.1788)
Working Independence	0.0065	0.0030	(1.0079,1.2023)

Table 4.8: Parameter estimates only subjects living in zipcodes below the median income level adjusting for relative humidity and temperature of the same averaging time as the predictor of interest. pm2.5 IQR of 15.

Table 4.8 shows output comparing only the working independence method and the WCR method. The data set represents areas of lower socio-economic status. The inference obtained will vary substantially between the methods. The WCR estimate is about 30% below that of the working independence method, and also is not significant at a 5% level.

As mentioned in the simulation section, the scientific significance of these findings is that careful consideration needs to be taken prior to conducting the analysis on what the estimand of interest is, and if it is acceptable to break the matched case-control bonds for subjects (matched sets) with numerous matched pairs. Since the applied focus is an air exposure study, and air exposures tend to have a high degree of seasonality trend (even when the data is stratified based on season, as each season is defined to a 6 month period), the discrete method can tend to give inaccurate estimates due to the lack of exchangeability of the exposure series.

Focusing on the working independence method and the within cluster resampling method (multiple outputation), the researcher must consider what the goal of the study is, and whether they should weight each subject equally or weight each observation equally. If the goal of the study is to obtain inference that addresses each subject equally, as would be the case of wanting to implement policy to reduce risk of an event for the target population, then it is reasonable to weight each cluster (subject) equally in the estimation procedure. This would result in using the WCR method. If the goal of the study is to obtain inference that addresses each event equally, as would be the case of wanting to implement policy to reduce cost of treatment of events, then it is reasonable to weight each event equally in the estimation procedure (clusters are weighted proportional to their size). This would result in using the working independence method.

4.6 Discussion

Many studies that use the case-crossover design have events of interest that can be experienced numerous times. As a result, the number of events across subjects can vary greatly. Informative cluster size is when the size of a cluster is associated to the risk for the outcome of interest. Using a study of the effect of air pollution on the risk of asthma related issues as an example, it could be that the subjects that experience a high number of events have a higher coefficient value for the effect of the exposure on the risk of an event or have a higher baseline risk of event. Three existing methods were considered in this chapter that can be used to obtain parameter estimates in a case-crossover designed study with heavily unbalanced cluster size, and characteristics of each method were highlighted using simulations and an applied illustrative dataset.

In Chapter 2 it was shown that the appropriate method to obtain estimates in a case-crossover design with numerous events per subject when one is willing to break the bond between matched case-control pairs and also account for the correlation among the observations in the estimation procedure was the discrete method. The discrete method assumes independence across the subjects, but not within subject. As a result, this method will combine all matched pairs within subjects to create a single risk set, and proceeds to compare the exposure values for the events to the entire risk set. When subjects have numerous events, the risk set will no longer maintain the bond between each matched pair of case-control. By stratifying the data by season, and adjusting for temperature and relative humidity, the hope is that the breaking of the bonds of matched pairs within the seasonally stratified data will not result in a significant bias. The working independence method assumes independence across all subjects and within subjects, which essentially treats each matched pair as its own cluster. As a result, matched case-control bonds are maintained. This method will have parameter estimates that are attenuated substantially to the parameter of the larger clusters, since these clusters will represent a greater portion of the data being used to obtain estimates. A within cluster resampling scheme termed multiple outputation aims to alleviate the issue of informative cluster size by giving each cluster equal weight in the overall marginal parameter estimate. The details of this method (and how it maintains the bonds between matched case-control pairs was discussed in Section 4.3.3).

The inference obtained from the study at hand will differ depending on the method used for analysis. As shown in the simulation study, both the working independence and discrete methods resulted in marginal parameter estimate that were heavily attenuated towards the larger clusters true parameter value, with the discrete method having less attenuation than the working independence method due to correlated covariates across matched pairs within the large clusters. The WCR method results in a parameter estimate that was the subject weighted average of the small clusters coefficient value and the larger clusters coefficient value, since in each outputation each subject only contributes a single events information to the sub sampled data being used to obtain parameter estimates. In the illustration, it was shown that the magnitude of the parameter estimates can vary across methods, with some methods showing no significance of the odds ratio differing from 1. Since the discrete method was not always close to the working independence method's estimate, the belief is that some subjects had highly correlated observations or that stratifying on season and adjusting for humidity and temperature did not completely alleviate the issue of breaking the bonds.

Importantly, it is critical to focus on the overall scientific goal of the analysis, which will dictate whether to obtain a parameter estimate that weights the clusters equally, or to obtain a parameter estimate that weights the individual observations equally. In epidemiological studies, many times the aim of the study is to implement new policies that will improve the overall health of the target population. If that is the case, and if each subject in the study is believed to equally represent the target population, then one might want to consider using a subject weighted estimand. If it is thought that the subjects with more events experienced have a greater representation of the target population than subjects with only a few events experienced, then an observation weighted estimand is more appropriate. The researchers must consider, a priori to conducting the analysis, what estimand is of greatest scientific importance and why.

4.7 Appendix

We briefly review the result of Williamson et al. [2003] which presents the WCR method in a WEE framework. First consider the usual generalized estimating equation given by

$$U_k(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \beta_k} \Sigma_i^{-1} \{\boldsymbol{Y_i} - \boldsymbol{\mu}_i(\boldsymbol{\beta})\}$$

where $\boldsymbol{\mu}_i$ is the vector of means for cluster i, $\boldsymbol{Y}_i$ is the vector of outcomes variable for cluster i and its specified covariance structure is Σ_i. Estimates for β_k , $k = 1, 2, ..., p$ are obtained by setting these equations to 0 and solving for β_k.

Assume 1:1 matching for each event within a subject. In each round of the multiple outputation procedure, $\boldsymbol{\beta_{(q)}}$ solves the following equation:

$$S_{(q)}(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} U_{ij}(\beta) * I[(i,j) \in r_q] = 0$$

where r_q being the set of indices (i,j) sampled in the q-th resampling. In setting of case-crossover data with 1:1 matching, and without loss of generality assuming the first exposure in the matched pair is the event exposure and the second be the control exposure:

$$\boldsymbol{U}_{ij}(\boldsymbol{\beta}) = \boldsymbol{x}_{1ij} - \frac{\sum_{l=1}^{2} \boldsymbol{x}_{lij} \exp(\boldsymbol{\beta} \boldsymbol{x}_{lij})}{\sum_{l=1}^{2} \exp(\boldsymbol{\beta} \boldsymbol{x}_{lij})}$$

Since the resampling distribution is a discrete uniform distribution with a probability mass of $\frac{1}{m_i}$ on each observation within cluster/subject i $(P(I[(i,j) \in r_q]) = 1/m_i)$, the WEE becomes:

$$S_{WEE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{m_i} \left[\sum_{j=1}^{m_i} \boldsymbol{x}_{1ij} - \frac{\sum\limits_{l=1}^{2} \boldsymbol{x}_{lij} \exp(\boldsymbol{\beta} \boldsymbol{x}_{lij})}{\sum\limits_{l=1}^{2} \exp(\boldsymbol{\beta} \boldsymbol{x}_{lij})} \right]$$

As noted earlier, this is the same as weighting the estimating equations for each subject by the inverse of the number of events they have. If a subject has 5 events, then each event contributes information with one-fifth the weight of that of a subject who has one event. Williamson et al. [2003] show that $\boldsymbol{\beta}_{WCR}^{\infty}$ and the WEE estimator are asymptotically equivalent, and that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{WEE} - \boldsymbol{\beta}) \overset{D}{\rightarrow} N(\mathbf{0}, \hat{H}^{-1} \hat{V} \hat{H}^{-1})$$

where $\hat{H} = \frac{1}{n} \sum\limits_{i=1}^{n} \frac{1}{m_i} \sum\limits_{j=1}^{m_i} \frac{\partial \boldsymbol{U}_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \mid_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}}$,

and $\hat{V} = \frac{1}{n} \sum\limits_{i=1}^{n} \{ \frac{1}{m_i} \sum\limits_{j=1}^{m_i} \boldsymbol{U_{ij}}(\hat{\boldsymbol{\beta}}) \} \{ \frac{1}{m_i} \sum\limits_{j=1}^{m_i} \boldsymbol{U_{ij}}(\hat{\boldsymbol{\beta}}) \}'$.

Chapter 5

A Semi-Parametric Bayesian Hierarchical Model for Analyzing Matched Case-Control Studies with Unbalanced Cluster Sizes

5.1 Introduction

The research presented in this chapter will build on the foundation set forth by Sinha et al. [2004] and Sinha et al. [2005], and present a conclusion to the discussion from the previous chapters. The discussion up to this point has focused on matched case-control studies, with specific focus on the case-crossover design. A recap of the discussion thus far follows. In Chapter 1, the applied study under focus was presented: An air pollution study aimed at making inference on the effect of environmental exposures on the risk of experiencing exacerbated asthma requiring a hospital encounter. Chapter 1 Section 1.1.1 presented statistical issues in addressing the goal of the applied study under focus. In Chapter 2, Section 2.3.2, the methodology used to obtain parameter estimates in a general matched case-control study was presented. It was shown the conditional logistic likelihood forms the basis of a proper method to obtain reliable parameter estimates in a matched case-control study. This method expands a simple logistic model by conditioning on the number of events known to happen in each matched set (subjects in a case-crossover design). By doing so, the stratum specific intercept coefficients are factored out of the likelihood. This method proceeds to obtain estimates by maximizing the conditional likelihood. Additionally, in Chapter 2, Section 2.3.3, it was shown that a case-crossover design is an appropriate study design to investigate the applied study presented in Chapter 1.

The case-crossover design was discussed at length in Chapter 2, Section 2.3.3, and again reviewed in Chapters 3 and 4. In a case-crossover study, a case subject is also the control subject. Control exposures are the exposures experienced by the subject a sufficient amount of time away from the event time, either in the past or future. Details about the validity of this method depend on the study under focus. How to appropriately choose the control times in order to obtain unbiased estimating equations was discussed. Chapter 2, Section 2.4 showed that maximization of the conditional logistic likelihood is accomplished by maximizing Cox's partial likelihood on a transformed dataset.

In matched case-control studies in general, it is likely the matched sets contain several matched pairs. In a case-crossover design, the event of interest can be experienced numerous times. Chapter 3, Section 3.2.1 showed that numerous pairs in the matched sets in the conditional logistic likelihood is analogous to tied event times within strata in the Cox proportional hazards partial likelihood. With the foundation laid for the estimation of parameters in a matched case-control study using the Cox proportional hazards partial likelihood, Chapter 3 continued to discuss the different methods that can be used to obtain parameter estimates under the scenario of tied event times in the partial likelihood. Issues were highlighted with approximation methods and the Kalbfleisch-Prentice method, which assumes a true ordering of tied event times. It was determined that the discrete method presented by Cox (1972) is the appropriate method to obtain unbiased parameter estimates when one wants to account for the correlation among observations in the parameter estimation procedure (i.e. maintain the clustering of the data). All methods discussed in Chapter 3 assume the willingness to break the individual matched case-control bonds within each matched set.

Chapter 4 expanded on the scenario presented in Chapter 3 by allowing for varying sizes of matched sets, and more importantly allowing for effect modification across the clusters. Keeping the discrete method presented in Chapter 3 as the appropriate method to obtain parameter estimates while accounting for the correlation within the data in the estimation procedure, two additional procedures were introduced, the independent method and within cluster resampling method (termed as the multiple outputation method). The goal of these methods is to obtain a (marginal) parameter estimate in the scenario of varying cluster sizes and effect modification across clusters. The issue of the choice of estimand was highlighted as well as the issue of breaking the bonds between each matched pair of case-controls when maintaining the clustering of the data. In the discrete method, each subjects matched pairs will be combined to form a single risk set. If a matched set has numerous matched pairs, then the risk set will contain numerous cases and their respectively matched controls, but it will no longer be apparent which control was matched to which case. It was discussed in Chapter 3, Section 3.2.2 that the method that will account for the clustering of the data in the estimation procedure will compare the individual event exposures to all exposures in the risk set. This could lead to biased results if there are trends in the exposure series, as an event exposure will be compared all exposure values in the risk set, which potentially can contain exposures from different seasons that were not matched to the event. In air exposure studies, the scenario of comparing event exposures to exposures from a different season will arise.

When the parameter of interest varies across the matched sets, one must consider what the most scientifically relevant target of inference is. From a case-crossover viewpoint, the matched sets represent the individual subjects. Within each subject, there is potential to have numerous matched pairs if the event of interest can be experienced numerous times. If the estimand weights the individual matched pairs equally across all subjects, then subjects with a higher frequency of events (and thus more matched pairs) will attenuate the (marginal) parameter estimate towards the value of their coefficient. If the estimand weights individual subjects equally, the estimation procedure will result in a (marginal) parameter estimate that will not give higher weight to the subjects with more events even though they represent a larger proportion of the set of observations. There is no absolutely correct estimand, but one must decide which estimand best addresses the study's goal and inference aims prior to conducting the analysis.

In this chapter, a solution is proposed to alleviate the issues posed and highlighted in the previous chapters. By implementing a semi-parametric hierarchical Bayesian model, one will be able to compute robust estimates across and within the matched sets. In a single model fit, it can be determined if effect modification is apparent across the subjects. The model also extends to estimating subject specific effects on the distribution of the exposure, as well parameters that define the relationship between exposure of interest and other covariates in the model. These are parameters and effects that are not estimated in the conditional logistic regression method.

A Dirichlet process (DP) prior is used on the mean of the prior of the stratum specific effects, as this allows for a flexible class of distributions to be used as the prior for the stratum specific parameters. Specifically, the specified prior of the subject specific effects will be Gaussian, and the mean of the prior distribution will have a Dirichlet process prior. This is called a DP location mixture of normals. DP mixtures (DPM) are countable mixtures with an infinite number of components and a specific prior on the weights and the component specific parameter (Ferguson [1983], Escobar [1994]).

For an appropriate choice of the kernel for the subject specific effects, the DPM model has support on a large class of distributions (Lo [1984]). This will avoid misspecification of single parametric form for the prior of the stratum specific effects. If there is effect modification across stratums, it is reasonable to think the distribution of the stratum specific effects is a mixture (with several modes). The Dirichlet process prior will also cluster subjects based on their similar stratum specific effects. This will allow for borrowing of information across subjects when sampling a mean value for the prior of that cluster. Gibbs sampling will be implemented to obtain draws from the full conditional distributions of the subject specific effects (Escobar and West [1995], MacEachern and Muller [1998]).

5.2 Model and Notation

5.2.1 Preliminaries

In this section, notation and relevant quantities needed to derive the likelihood are presented. Let $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m_i$, and $l = 1, 2, \ldots, M + 1$ (1:M matching), where n is the number of subjects (or matched sets), m_i is the number events (i.e. matched pairs) for this subject, and l is the index of the l^{th} observation in this subjects matched pair. Let X_{ijl} be a single exposure of interest with possibly missing values, D_{ijl} represent the event outcome (0 or 1), and set $\boldsymbol{Z}_{ijl}$ to be a 1 by p vector of covariates not of primary interest ($\boldsymbol{Z}_{ijl} = (Z_{ijl1}, \ldots, Z_{ijlp})$). Set $\boldsymbol{S}_i$ to be the collection of measured and unmeasured stratification variables for the stratum i. Additionally let δ_{ijl} be an indicator equal to 0 if X_{ijl} is missing and 1 otherwise. Without loss of generality, let $l = 1$ represent the index of the event observation in each matched pair and $l = 2, \ldots, M + 1$ be the indices of the controls.

First, consider the exposure distribution among the controls belonging to an exponential family. Namely, consider a model of the form

$$p(X_{ijl}|D_{ijl}=0,\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) = \exp[\xi_{ijl}\{\theta_{ijl}X_{ijl}-b(\theta_{ijl})\}+c(X_{ijl},\xi_{ijl})]\,, \tag{5.1}$$

where $\theta_{ijl} = \gamma_{0i} + \boldsymbol{\gamma_1}\boldsymbol{Z_{ijl}}$ and $\boldsymbol{\gamma_1} = (\gamma_{11},\ldots,\gamma_{1p})^T$. (5.1) represents the conditional distribution of the exposure given $D_{ijl}=0$, $\boldsymbol{Z}_{ijl}$, and $\boldsymbol{S}_i$, written in a canonical exponential family form where θ_{ijl} represents the natural parameter. The varying intercept, γ_{0i}, will capture the stratum effect on the natural parameter θ_{ij}. As a result it will capture the stratum effect on the exposure distribution.

Next, assume a prospective probability of an event of the form of a logistic link:

$$\begin{aligned} P(D_{ijl}=1|X_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) &= H(\beta_{0i}+\boldsymbol{\beta_1}\boldsymbol{Z_{ijl}}+\beta_{2i}X_{ijl}) \\ &= \frac{\exp(\beta_{0i}+\boldsymbol{\beta_1}\boldsymbol{Z_{ijl}}+\beta_{2i}X_{ijl})}{1+\exp(\beta_{0i}+\boldsymbol{\beta_1}\boldsymbol{Z_{ijl}}+\beta_{2i}X_{ijl})}. \end{aligned} \tag{5.2}$$

The stratum varying coefficients are β_{0i} and β_{2i}. With the preceding specifications (5.1) and (5.2), the quantities required to construct the joint likelihood can be derived. The model structure will be of the form

$$\begin{aligned} p(D_{ijl},X_{ijl},\delta_{ijl}|\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) &= p(X_{ijl}|D_{ijl},\delta_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) \\ &\quad \times p(\delta_{ijl}|D_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) \times p(D_{ijl}|\boldsymbol{Z_{ijl}},\boldsymbol{S_i}), \end{aligned}$$

where $p(\delta_{ijl}|D_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i})$ does not depend on any parameters of interest. Additionally, the missing exposures are assumed to be missing at random, i.e. $p(X_{ijl}|D_{ijl},\delta_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) = p(X_{ijl}|D_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i})$.

Now to derive the quantities needed to construct the likelihood. The probability of an event marginalized over X is given by

$$\begin{aligned}
p(D_{ijl}=1|\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) &= \int P(D_{ij}=1|X_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i})p(X_{ij}|\boldsymbol{Z_{ij}},\boldsymbol{S_i})dX_{ij} \\
&= \int \exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}+\beta_{2i}X_{ijl}]p(D_{ijl}=0|X_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) \\
&\quad \times p(X_{ijl}|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})dX_{ijl} \\
&= \exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ij}}]\int \exp[\beta_{2i}X_{ijl}]p(D_{ijl}=0,X_{ijl}|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})dX_{ij} \\
&= \exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}] \\
&\quad \times \int \exp[\beta_{2i}X_{ijl}]p(X_{ijl}|D_{ij}=0,\boldsymbol{Z_{ijl}},\boldsymbol{S_i})p(D_{ijl}=0|\boldsymbol{Z_{ij}},\boldsymbol{S_i})dX_{ijl}.
\end{aligned} \tag{5.3}$$

Using (5.3), it then follows that the marginal prospective odds of an event is given by

$$\begin{aligned}
\frac{p(D_{ijl}=1|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}{p(D_{ijl}=0|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})} &= \exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}]\int \exp[\beta_{2i}X_{ijl}]p(X_{ijl}|D_{ijl}=0,\boldsymbol{Z_{ijl}},\boldsymbol{S_i})dX_{ijl} \\
&= \exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}] \\
&\quad \times \int \exp[\beta_{2i}X_{ij}]\exp[\xi_{ij}\{\theta_{ij}X_{ij}-b(\theta_{ij})\}+c(X_{ij},\xi_{ij})]dX_{ij} \\
&= \exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}]\exp[-\xi_{ijl}b(\theta_{ijl})] \\
&\quad \times \int \exp[\beta_{2i}\xi_{ijl}\xi_{ijl}^{-1}X_{ijl}+\xi_{ijl}\theta_{ijl}X_{ijl}+c(X_{ijl},\xi_{ijl})]dX_{ijl}.
\end{aligned} \tag{5.4}$$

Letting $\theta^*_{ijl}=\theta_{ijl}+\xi_{ijl}^{-1}\beta_{2i}$, (5.4) factors into

$$\begin{aligned}
\frac{p(D_{ijl}=1|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}{p(D_{ijl}=0|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})} &= \exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}]\exp[-\xi_{ijl}b(\theta_{ijl})]\exp[\xi_{ijl}b(\theta^*_{ijl})] \\
&\quad \times \int \exp[\xi_{ijl}\{X_{ijl}\theta^*_{ijl}-b(\theta^*_{ijl})\}+c(X_{ijl},\xi_{ijl})]dX_{ijl} \\
&= \exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}+\xi_{ijl}\{b(\theta^*_{ijl})-b(\theta_{ijl})\}].
\end{aligned} \tag{5.5}$$

Observe that $\int \exp[\xi_{ijl}\{X_{ijl}\theta^*_{ijl}-b(\theta^*_{ijl})\}+c(X_{ijl},\xi_{ijl})]dX_{ijl}=1$, since it is a valid probability density function.

Now it is necessary to derive another needed quantity for the construction of the likelihood, the exposure distribution among cases, $p(X_{ijl}|D_{ijl}=1,\boldsymbol{Z_{ijl}},\boldsymbol{S_i})$. This is done as follows:

$$\begin{aligned}
p(X_{ijl}|D_{ijl}=1,\boldsymbol{Z_{ijl}},\boldsymbol{S_i}) &= \frac{p(X_{ijl},D_{ijl}=1|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}{p(D_{ijl}=1|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}\\
&= \frac{p(D_{ijl}=1|X_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i})p(X_{ijl}|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}{p(D_{ijl}=1|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}\\
&= \frac{\exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}+\beta_{2i}X_{ijl}]p(D_{ijl}=0|X_{ijl},\boldsymbol{Z_{ijl}},\boldsymbol{S_i})p(X_{ijl}|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}{\exp[\beta_{0i}+\boldsymbol{\beta_1 Z_{ijl}}+\xi_{ijl}\{b(\theta^*_{ijl})-b(\theta_{ijl})\}]p(D_{ijl}=0|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}\\
&= \frac{\exp[\beta_{2i}X_{ijl}]p(D_{ijl}=0,X_{ijl}|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}{\exp[\xi_{ijl}\{b(\theta^*_{ijl})-b(\theta_{ijl})\}]p(D_{ijl}=0|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}\\
&= \frac{\exp[\beta_{2i}X_{ijl}]p(D_{ijl}=0|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})p(X_{ijl}|D_{ijl}=0,,\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}{\exp[\xi_{ijl}\{b(\theta^*_{ijl})-b(\theta_{ijl})\}]p(D_{ijl}=0|\boldsymbol{Z_{ijl}},\boldsymbol{S_i})}\\
&= \frac{\exp[\beta_{2i}X_{ijl}]\exp[\xi_{ijl}\{\theta_{ijl}X_{ijl}-b(\theta_{ijl})\}+c(X_{ijl},\xi_{ijl})]}{\exp[\xi_{ijl}\{b(\theta^*_{ijl})-b(\theta_{ijl})\}]}\\
&= \exp[\beta_{2i}X_{ijl}+\xi_{ijl}\theta_{ijl}X_{ijl}-\xi_{ijl}b(\theta^*_{ijl})+c(X_{ijl},\xi_{ijl})]\\
&= \exp[\xi_{ijl}\{\theta^*_{ijl}X_{ijl}-b(\theta^*_{ijl})\}+c(X_{ijl},\xi_{ijl})],
\end{aligned} \tag{5.6}$$

where again $\theta^*_{ijl}=\theta_{ijl}+\xi^{-1}_{ijl}\beta_{2i}$. With the quantities specified in (5.1), (5.2), (5.5) and (5.6), the likelihood can be constructed.

5.2.2 Likelihood

The likelihood presented follows the form of Satten and Carroll [2000]. A joint likelihood of event status and exposure, $p(D,X,\delta|\boldsymbol{Z},\boldsymbol{S})$ is derived. This can lead to the needed estimates since $p(D,X,\delta|\boldsymbol{Z},\boldsymbol{S})\propto p(X|D,\boldsymbol{Z},\boldsymbol{S})p(D|\boldsymbol{Z},\boldsymbol{S})$, where the right hand side pieces were derived in the previous section.

The full likelihood across all subjects and matched pairs will be constructed as:

$$L_c = \prod_{i=1}^{n}\prod_{j=1}^{m_i} L_{c_{ij}},$$

where $L_{c_{ij}}$ is the i^{th} stratums j^{th} matched pair's contribution to the likelihood (to be shown shortly). This can be done because conditional on the matching factors and the stratum specific effects, one can treat the matched pairs within stratums as independent. Therefore, need only to derive the likelihood contribution for the i^{th} stratum's j^{th} matched pair. This is done as follows:

$$\begin{aligned}
L_{c_{ij}} &= p(\boldsymbol{D}_{ij.}, \boldsymbol{X_{ij.}}, \boldsymbol{\delta_{ij.}} | \boldsymbol{Z}_{ij.}, \boldsymbol{S}_i, \sum_{l=1}^{M+1} D_{ijl} = 1) \\
&\propto p(\{X_{ijl}\}_{l=1}^{M+1} | \boldsymbol{Z}_{ij.}, \boldsymbol{S}_{ij}, \boldsymbol{D}_{ij.}, \boldsymbol{\delta_{ij.}}) p(\boldsymbol{D}_{ij.} | \boldsymbol{Z}_{ij.}, \boldsymbol{S}_i, \sum_{l=1}^{M+1} D_{ijl} = 1) \\
&= p(X_{ij1} | \boldsymbol{Z_{ij.}}, \boldsymbol{S_{ij}}, D_{ij1} = 1, \delta_{ij1}) \prod_{l=2}^{M+1} p(X_{ijl} | \boldsymbol{Z}_{ijl}, \boldsymbol{S}_{ij}, D_{ijl} = 0, \delta_{ijl}) \\
&\quad \times \frac{p(D_{ij1} = 1 | \boldsymbol{S_i, Z_{ij1}}) \prod_{l=2}^{M+1} p(D_{ijl} = 0 | \boldsymbol{S_i, Z_{ijl}})}{\sum_{l=1}^{M+1} p(D_{ijl} = 1 | \boldsymbol{S_{ij}, Z_{ijl}}) \prod_{k \neq l}^{M+1} p(D_{ijk} = 0 | \boldsymbol{S_{ij}, Z_{ijk}})} \\
&= p(X_{ij1} | \boldsymbol{Z_{ij.}}, \boldsymbol{S_{ij}}, D_{ij1} = 1, \delta_{ij1}) \prod_{l=2}^{M+1} p(X_{ijl} | \boldsymbol{Z}_{ijl}, \boldsymbol{S}_{ij}, D_{ijl} = 0, \delta_{ijl}) \\
&\quad \times \frac{p(D_{ij1} = 1 | \boldsymbol{S_i, Z_{ij1}}) / p(D_{ij1} = 0 | \boldsymbol{S_i, Z_{ijl}})}{\sum_{l=1}^{M+1} p(D_{ijl} = 1 | \boldsymbol{S_{ij}, Z_{ijl}}) / p(D_{ijl} = 0 | \boldsymbol{S_{ij}, Z_{ijk}})}.
\end{aligned}$$

Using the quantities derived in Section 5.2.1, the likelihood contribution for strata i, $L_{c_{ij}}(\boldsymbol{\beta_1}, \beta_{2i}, \boldsymbol{\gamma_1}, \gamma_{0i})$, is:

$$\begin{aligned}
&\exp[\delta_{ij1}\xi_{ij1}\{\theta^*_{ij1}X_{ij1} - b(\theta^*_{ij1})\} + \delta_{ij1}c(X_{ij1}, \xi_{ij1})] \\
&\times \prod_{l=2}^{M+1} \exp[\delta_{ijl}\xi_{ijl}\{\theta_{ijl}X_{ijl} - b(\theta_{ijl})\} + \delta_{ijl}c(X_{ijl}, \xi_{ijl})] \\
&\times(1 + \sum_{l=2}^{M+1} \exp[\boldsymbol{\beta_1(Z_{ijl} - Z_{ij1})} + \xi_{ijl}\{b(\theta^*_{ijl}) - b(\theta_{ijl})\} - \xi_{ij1}\{b(\theta^*_{ij1}) - b(\theta_{ij1})\})^{-1}.
\end{aligned} \tag{5.7}$$

The subject specific intercepts β_{0i} have been factored out of the likelihood contributions in (5.7), but subject specific coefficients β_{2i} and subject specific nuisance parameters γ_{0i} still remain.

Two specific scenarios are investigated for the distribution of the exposure. The focus is on a normal distribution for a continuous exposure and a Bernoulli distribution for a binary exposure. To incorporate different exposure distributions (from an exponential family of distributions) into the likelihood, one would need to match the probability density (or mass) function of the exposure to the exponential family form given in (5.1) for the controls and (5.6) for the cases. Then one would obtain the needed quantities θ, $b(\theta)$, ξ and $c(X, \xi)$.

In the binary exposure case, where $X \sim \text{Bernoulli}(\log(\frac{p}{1-p}) = \theta)$, the probability mass function is given by

$$f(X_{ijl} = x_{ijl} | D_{ijl} = 0, \boldsymbol{Z_{ijl}, S_i}) = p_{ijl}^{x_{ijl}}(1 - p_{ijl})^{1-x_{ijl}}.$$

The individual components for the binary exposure in the exponential family form are:

$$\begin{aligned}
\theta_{ijl} &= \log\left(\frac{p_{ijl}}{1-p_{ijl}}\right) = \gamma_{0i} + \boldsymbol{\gamma_1 Z_{ijl}} \\
b(\theta_{ijl}) &= \ln(1+\exp(\theta_{ijl})) \\
\xi_{ijl} &= 1 \\
c(x_{ijl}, \xi_{ijl}) &= 0 \\
\theta^*_{ijl} &= \theta_{ijl} + \beta_{2i}.
\end{aligned}$$

In the continuous exposure case, where $X \sim \mathrm{N}(\theta, \sigma^2)$, the probability density function of the exposure is:

$$f(X_{ijl} = x_{ijl} | D_{ijl} = 0, \boldsymbol{Z_{ijl}, S_i}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}(x_{ijl} - \theta_{ijl})^2]$$

The individual components for the continuous exposure, written in exponential family form, are

$$\begin{aligned}
\theta_{ijl} &= \gamma_{0i} + \boldsymbol{\gamma_1 Z_{ijl}} \\
b(\theta_{ijl}) &= \frac{(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{ijl}})^2}{2} \\
\xi_{ijl} &= \frac{1}{\sigma^2} \\
c(x_{ijl}, \xi_{ijl}) &= \log(\sqrt{\frac{\xi_{ijl}}{2\pi}}) - \frac{x^2_{ijl}\xi_{ijl}}{2} \\
\theta^*_{ijl} &= \theta_{ijl} + \sigma^2\beta_{2i}.
\end{aligned}$$

In this section, the model specifications were presented. Conditional on the subject specific effects and parameters, the likelihood was derived. The subject specific effects (random effects) are γ_{0i} and β_{2i}, and the fixed parameters are $\boldsymbol{\beta_1}, \boldsymbol{\gamma_1}$ and in the case of a continuous exposure there is the additional parameter of σ^2. The following section will present a Bayesian semi-parametric model that will be used to obtain estimates.

5.3 Bayesian Hierarchical Model

A hierarchical model using a Bayesian paradigm is employed to obtain parameter estimates. Similar to a random effects model, the frequentist approach to obtain estimates would be to obtain the marginal likelihood to compute marginal parameter estimates and random effects distribution parameters (Diggle et al. [1994]). This would require specifying a fully parametric distribution for the random effects. The marginal likelihood would be obtained by integrating the full augmented data likelihood with respect to this distribution, in order to factor out the random effects. Estimates would be obtained by maximization of this marginal likelihood with respect to the parameters. Misspecification of the random effects distribution can lead to non robust estimation. This is particularly an issue when the random effects do not arise from a single distribution, but a mixture. Additionally, the integration can be intensive within non-standard likelihoods (Heagerty [1999]).

The proposed approach to circumvent these issues is to model the random effects as arising from a non-parametric distribution. This allows for more flexibility than parametric assumptions and is accomplished by assuming a parametric prior distribution (Gaussian) for the subject specific random effects and a Dirichlet process prior on the parameters of this distribution. In doing so, a flexible class of distributions is assumed for the subject specific coefficients by the fact that the Dirichlet process mixtures allows for a countable mixture. This approach uses a Dirichlet process as the random mixing distribution for the Gaussian parameters with an infinite number of components. This was first introduced by Antoniak [1974] and formalized by Ferguson [1983], Lo [1984] and Escobar [1994]. The idea is demonstrated and implemented using a Gibbs sampler in Escobar [1994] and Escobar and West [1995]. Additionally, this approach will allow for the borrowing of information across subjects by clustering subjects together. The hierarchical model presented in Chapter 2.4.2 highlights the nature of clusters generated by ties among the draws from the discrete probability measure G.

The approach proposed here expands on that set forth by Sinha et al. [2005]. The focus of this work was to create a semi-parametric Bayesian model for the analysis of matched case-control studies with missing exposures. By modeling the association of the completely observed covariate Z to the exposure of X (with possible missing values), matched pairs with missing values for X can still be included in the analysis. This is not the focus of the research presented here, but is implemented to handle missing values of the exposure in the illustrated applied examples presented in a later section.

In previous work by Sinha et al. [2005], a fixed value for β_2 is assumed across all matched sets. Additionally, only a single event per matched set is assumed and only a binary exposure case is investigated. The prior structures in the approach in this chapter also differ from that of Sinha et al. [2005] in which the Dirichlet process is placed directly on the γ_{0i} subject specific effects, which is to say $\gamma_{0i}|G \overset{iid}{\sim} G,\ G \sim \mathrm{DP}(\alpha, G_0)$. This implies equality of the γ_{0i}'s across some of the matched sets. In this approach, as mentioned earlier, the Dirichlet process prior is assumed for the parameters of the γ_{0i} prior distribution's parameter. This will potentially allow each matched set to have its own unique exposure distribution, as would be the case in a case-crossover study where each subject has their own personal exposure monitor (i.e. no shared exposures across any of the matched sets).

5.3.1 Fixed Effect Parameters

In this section, the prior specifications for the fixed parameters are presented. The full conditional distributions for the parameters $\boldsymbol{\beta_1}$, $\boldsymbol{\gamma_1}$ and in the continuous exposure case σ^2 are presented in the appendix. All fixed effect parameters are assumed to be independent of each other in the prior. The following priors are specified for the fixed effects of the model:

$$\boldsymbol{\beta}_1 \sim \mathrm{N}(\boldsymbol{\mu}_{\beta_1}, \Sigma_{\beta_1}),$$

$$\boldsymbol{\gamma}_1 \sim \mathrm{N}(\boldsymbol{\mu}_{\gamma_1}, \Sigma_{\gamma_1}).$$

Finally, for the continuous exposure, an inverse gamma prior is specified for σ^2: $\sigma^2 \sim \mathrm{IG}(a, b)$.

5.3.2 Random Effect Parameters

The subject specific effects in the model are γ_{0i} and β_{2i}, for $i = 1, 2, \ldots, n$. Independent DPs for the γ_{0i}'s and the β_{2i}'s are specified. To avoid presenting the approach twice, the model specification for β_{2i} will be presented, noting that the specifications for γ_{0i} is similar. The hierarchical set up is as follows:

$$\begin{aligned}
\beta_{2i}|\mu_{\beta i}, \sigma^2_\beta &\sim \mathrm{N}(\mu_{\beta i}, \sigma^2_\beta) \\
\mu_{\beta i}|G &\sim G \\
G &\sim \mathrm{DP}(\alpha, G_0) \\
G_0 &\equiv \mathrm{N}(\mu_{\beta 0}, \sigma^2_{\beta 0}) \\
\sigma^2_\beta &\sim \mathrm{IG}(a_{\beta 0}, b_{\beta 0}) \\
\mu_{\beta 0} &\sim \mathrm{N}(\mu_{\beta 00}, \sigma^2_{\beta 00}) \\
\sigma^2_{\beta 0} &\sim \mathrm{IG}(a_{\beta 00}, b_{\beta 00}) \\
\alpha_\beta &\sim \mathrm{Gamma}(a_{\beta\alpha}, b_{\beta\alpha}).
\end{aligned} \tag{5.8}$$

The random probability measure G in (5.8) is be constructed via a stick breaking process (Sethuraman [1994]). In this paper, Sethuraman proved that the stick breaking procedure he introduces is equivalent to the Dirichlet process set forth by Ferguson [1973] that satisfies the Kolmogorov consistency definitions. These points are discussed in Chapter 2, Section 2.4.2. To repeat, G will be constructed as

$$G \approx \sum_{l=1}^{K} \pi_l \delta_{\mu_l}(\cdot)$$

$$\mu_l \overset{iid}{\sim} G_0$$

$$\pi_l = v_l \prod_{j=1}^{l-1} (1 - v_j) \text{ for } l = 1, 2, \ldots, K-1 \text{ and } \pi_K = 1 - \sum_{l=1}^{K-1} \pi_l$$

$$v_j \overset{iid}{\sim} \text{Beta}(1, \alpha) \text{ for } j = 1, 2, \ldots, K-1.$$

The full conditional distributions for the π's and v's from the stick breaking process, as well as the full conditional distributions of the parameters $\sigma^2_{\beta i}, \mu_{0\beta}, \sigma^2_{0\beta}$, and α_β are in the chapter appendix. The Metropolis-Hastings algorithm used to implement the sampling is also provided in the chapter appendix.

With respect to the applied study under focus throughout the dissertation, it is reasonable to believe that not all subjects exhibit the same effect of the exposure on the risk of an event (or have the same exposure distribution). Additionally, it is reasonable to think that not all subjects are different than one another. As a result, it is possible the subject specific effects arise from a mixture of distributions (with several modes) as opposed to a single distribution. Specifying a single parametric prior distribution for the subject specific effects will be an inefficient approach, as it would have to be a fairly diffuse prior distribution.

By specifying a Dirchlet process prior for the means of the prior distributions of the subject specific effects, a flexible class of distributions is implied for the prior of the subject specific effects. Additionally, this approach will allow for the borrowing of information across subjects by clustering subjects together. The hierarchical model presented below highlights the nature of clusters generated by ties among the draws from the discrete probability measure G.

Subjects will be clustered together based on the similarity of their subject specific effect values at any given iterate. Subjects that belong to same cluster (i.e. have equal values for their $\mu_{\beta i}$) will pool their information together (i.e. their β_{2i}'s or γ_{0i}'s)) to update the mean of the prior for this cluster ($\mu_{\beta i}$ or $\mu_{\gamma i}$). This can be seen in Appendix 5.7.2 with the derivations of the full conditional distributions of individual subject labels, c_i and the cluster means, $\mu_{\beta i}$.

Observe the stick breaking procedure described is called the truncated DP. In Sethuraman's proof of the equivalence of the stick breaking construction to the definition set forth by Ferguson $K = \infty$, but in practice $K < \infty$. How to determine if the value set for K is too restrictive is as follows. Conduct an initial analysis and record the number of occupied clusters (number of unique $\mu_{\beta i}$'s) across the MCMC iterates. If the number of occupied clusters is K across most iterates, then the value for K is too restrictive and would need to be increased. Note that at most K can only be as large as n.

As an aside, a bivariate approach was also implemented where the stratum specific effects were specified a bivariate prior. The mean vector comprising of the prior means of γ_{0i} and β_{2i} was a given a single DP prior (and G_0 was a bivariate kernel). This approach gave similar output with respect to the simulation study and illustrated examples to be presented in the following sections. Details of this approach are in the Appendix 5.7.5.

5.4 Simulation Study

5.4.1 Data Simulation

The aim of the data simulation is to generate data in which subjects exhibit effect modification with respect to the exposure. Additionally, there will be clusters of subject who exhibit similar effects. This is implemented by having the subject specific effects of β_{2i} coming about from a mixture of three normal distributions. This creates three subpopulations of subjects, with respect to the effect of an exposure on the risk of an event, within the dataset. The subject specific effects of γ_{0i} will also be generated from a mixture of 3 normals. This is done to have 3 subpopulations of subjects with respect to their exposure distributions. The data generation explained below will induce correlation between a given subject's γ_{0i} and β_{2i} values. Specifically, the larger the γ_{0i} (indicating a higher mean level of exposures) the more likely the β_{2i} value will be large.

For each subject, a label c_γ was generated from a uniform{1,2,3}. The label indicated which normal distribution the value for γ_{0i} was drawn from. Based on the label of c_γ for a certain subject, the probability of what the label c_β would be (1,2, or 3) was generated using a multinomial regression. The label c_β determined which normal the β_{2i} was from. This was done in order to induce correlation between the γ_{0i} and β_{2i} values.

Given the label for c_γ, the probabilities of c_β being 1,2, or 3 are stated as follows:

$$\pi_{i1} = P(c_\beta = 1|c_\gamma) = \frac{1}{1 + \sum\limits_{k=2}^{3} \exp(\eta_{ik})},$$

$$\pi_{i2} = P(c_\beta = 2|c_\gamma) = \frac{\exp(\eta_{i2})}{1 + \sum\limits_{k=2}^{3} \exp(\eta_{ik})},$$

$$\pi_{i3} = P(c_\beta = 3|c_\gamma) = \frac{\exp(\eta_{i3})}{1 + \sum\limits_{k=2}^{3} \exp(\eta_{ik})},$$

where for $k = 2, 3$:

$$\eta_{ik} = \log\left(\frac{P(c_\beta = k|c_\gamma)}{P(c_\beta = 1|c_\gamma)}\right) = \theta_{0k} + \theta_{1k}\mathrm{I}(c_\gamma = 1) + \theta_{2k}\mathrm{I}(c_\gamma = 2) + \theta_{3k}\mathrm{I}(c_{\gamma 0} = 3).$$

The following values were used for the coefficients in the data generating model:

$\theta_{02} = -1.75$, $\theta_{12} = 1$, $\theta_{22} = 3$, $\theta_{32} = 1.5$

$\theta_{03} = -2$, $\theta_{13} = 1$, $\theta_{23} = 1.5$, $\theta_{32} = 2.75$

Based on the formulation just described, the simulation of γ_{0i} and β_{2i} were from their respective mixture of normals as follows:

$$\gamma_{0i} \sim \frac{1}{3}N(\mu_{\gamma 1}, 0.4^2) + \frac{1}{3}N(\mu_{\gamma 2}, 0.4^2) + \frac{1}{3}N(\mu_{\gamma 2}, 0.4^2),$$

$$\beta_{2i} \sim \pi_{i1}N(\mu_{\beta 1}, 0.4^2) + \pi_{i2}N(\mu_{\beta 2}, 0.4^2) + \pi_{i3}N(\mu_{\beta 3}, 0.4^2),$$

where the π's were given earlier. For γ_{0i}, the means of the normals used for the mixture were $\mu_{\gamma 1} = -2$, $\mu_{\gamma 2} = -1$, and $\mu_{\gamma 3} = 0$. For the exposure coefficient, two sets of means were used. One set was $\mu_{\beta 1} = 0$, $\mu_{\beta 2} = 1$, and $\mu_{\beta 3} = 2$ and the other set was $\mu_{\beta 1} = 0$, $\mu_{\beta 2} = 0.3$, and $\mu_{\beta 3} = 0.7$. In the simulation study, a single covariate Z was used, however the process can be easily extended to thee scenario of a multivariate $\boldsymbol{Z}$. The fixed parameters were set to be $\beta_1 = 1$ and two values were used for γ_1, 0.05 and 0.3. In the case of a continuous exposure, $\sigma^2 = 1$.

The generation scheme described and the values used for the coefficients in the multinomial probabilities resulted in an induced correlation between the labels c_γ and c_β. Given the label for c_γ, the probability that the label c_β will be the same value was higher than the probability that it would be some other value. Observe that from Section 5.2.1, γ_{0i} can viewed as the mean exposure level of a subject given covariate $Z = 0$. This implied that subjects with lower values for their mean exposure level were more likely to have a lower value of the effect coefficient (and similarly for higher values of the mean exposure level with having higher values for the effect coefficient).

The expected values of the subject specific effects are as follows. Since $c_\gamma \sim$ uniform$\{1,2,3\}$ and $\mu_\gamma \in \{-2,-1,0\}$, the following is obtained

$$\begin{aligned} \mathrm{E}[\gamma_{0i}] &= \frac{1}{3}\mu_{\gamma 1} + \frac{1}{3}\mu_{\gamma 2} + \frac{1}{3}\mu_{\gamma 3} \\ &= -1 \end{aligned} \tag{5.9}$$

For the mean of μ_β, iterated expectations are used to obtain

$$\begin{aligned} \mathrm{E}[\beta_{2i}] &= \mathrm{E}_{c_\gamma}\left\{\mathrm{E}[\beta_{2i}|c_\gamma]\right\} \\ &= \mathrm{E}_{c_\gamma}\left[\pi_1\mu_{\beta 1} + \pi_2\mu_{\beta 2} + \pi_3\mu_{\beta 3}\right] \\ &= \sum_{h=1}^{3}\mu_{\beta h}\mathrm{E}[\pi_h] \\ &= \sum_{h=1}^{3}\left[\mu_{\beta h}\sum_{j=1}^{3}\frac{1}{3}\left\{\pi_h|c_\gamma = j\right\}\right] \end{aligned} \tag{5.10}$$

In the case where $\mu_\beta \in \{0, 0.3, 0.7\}$, (5.9) is equal to 0.3153. When $\mu_\beta \in \{0,1,2\}$, then (5.10) is equal to 0.9553. These will be used later to compare methods in the simulation study.

Once a subject's true parameter values are determined, the simulation of the data for this subject could begin. For the j^{th} event from subject i, the case event exposures and control event exposures are simulated as follows. Given $\boldsymbol{S_i}$, the observed and unobserved matching factors that define strata i, and the strata specific effects γ_{0i} and β_{2i}, the events within a subject are independent of one another. Let m_i denote the number of events for a given subject. To simulate m_i events with controls for a subject, a single matched case-control pair is simulated and this process is repeated m_i times. Thus for a subject with 5 events, an event and controls is simulated, and this is repeated 5 times. Therefore, the j subscript is suppressed for the derivation of the simulation procedure.

For 1:1 matching corresponding to each of subject i's events, first simulate two $\boldsymbol{Z}$ values in order, $\boldsymbol{Z}_{i1}$ and $\boldsymbol{Z}_{i2}$. It will be determined which of these covariates will be for the case and which will be for the control. Compute the probability that the first $\boldsymbol{Z}$ value ($\boldsymbol{Z}_{i1}$) will generate an event by

$$
\begin{aligned}
&= p(D_{i1} = 1 | D_{i1} + D_{i2} = 1, \boldsymbol{Z_{i1}, Z_{i2}, S_i}) \\
&= \frac{p(D_{i1} = 1, D_{i2} = 0 | \boldsymbol{Z_{i1}, Z_{i2}, S_i})}{p(D_{i1} + D_{i2} = 1 | \boldsymbol{Z_{i1}, Z_{i2}, S_i})} \\
&= \frac{p(D_{i1} = 1 | \boldsymbol{Z_{i1}, Z_{i2}, S_i}) p(D_{i2} = 0 | \boldsymbol{Z_{i1}, Z_{i2}, S_i})}{p(D_{i1} = 1 | \boldsymbol{Z_{i1}, Z_{i2}, S_i}) p(D_{i2} = 0 | \boldsymbol{Z_{i1}, Z_{i2}, S_i}) + p(D_{i1} = 0 | \boldsymbol{Z_{i1}, Z_{i2}, S_i}) p(D_{i2} = 1 | \boldsymbol{Z_{i1}, Z_{i2}, S_i})} \\
&= \frac{p(D_{i1} = 1 | \boldsymbol{Z_{i1}, S_i}) p(D_{i2} = 0 | \boldsymbol{Z_{i2}, S_i})}{p(D_{i1} = 1 | \boldsymbol{Z_{i1}, S_i}) p(D_{i2} = 0 | \boldsymbol{Z_{i2}, S_i}) + p(D_{i1} = 0 | \boldsymbol{Z_{i1}, S_i}) p(D_{i2} = 1 | \boldsymbol{Z_{i2}, S_i})} \\
&= (1 + \exp[\boldsymbol{\beta_1 (Z_{i2} - Z_{i1})} + \xi_{i2}\{b(\theta^*_{i2}) - b(\theta_{i2})\} - \xi_{i1}\{b(\theta^*_{i1}) - b(\theta_{i1})\}])^{-1}.
\end{aligned}
$$

One can then simulate a Bernoulli random variable with the foregoing specified probability. If the simulated value is a 1, then $D_{i1} = 1$ and $\boldsymbol{Z}_{i1}$ is the $\boldsymbol{Z}$ covariate value for the case and $D_{i2} = 0$ with $\boldsymbol{Z}_{i2}$ the control covariate value. If D_{i1} is 0, then $D_{i2} = 1$ and $\boldsymbol{Z}_{i2}$ is the $\boldsymbol{Z}$ covariate for the case and $\boldsymbol{Z}_{i1}$ is the covariate value for the control.

Once D_{i1} and D_{i2} are simulated, the corresponding exposures will be simulated. In the scenario of a binary exposure, given the case covariate values, simulate the exposure for the case from a Bernoulli(p) distribution where $p = \frac{\exp(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}} + \beta_{2i})}{1 + \exp(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}} + \beta_{2i})}$ and for the control $p = \frac{\exp(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}})}{1 + \exp(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}})}$. That is to say if $D_{i1} = 1$, then $X_{i1} \sim$ Bernoulli($p = \frac{\exp(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}} + \beta_{2i})}{1 + \exp(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}} + \beta_{2i})}$) and if $D_{i1} = 0$, then $X_{i1} \sim$ Bernoulli($p = \frac{\exp(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}})}{1 + \exp(\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}})}$). Likewise for X_{i2} with the value for D_{i2}.

In the scenario of a normally distributed exposure, for the case's X simulate the exposure from a N($\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}} + \sigma^2 \beta_{2i}$, σ^2) and for the control from N($\gamma_{0i} + \boldsymbol{\gamma_1 Z_{il}}$, σ^2). Details of the simulation procedure for $1 : M$ matching, where $M > 1$, are provided in the chapter appendix.

Two generation schemes for the number of events per subject were considered depending on whether the exposure was continuous or binary. For the continuous exposure, the number of events per subject was sampled from a uniform{5,6,7}, resulting in a mean of 6 events per subject. For the binary exposure case, the number of events per subject was sampled from uniform{7,8,9}, resulting in a mean of 8 events per subject. Two different data generating scenarios were used for the Z covariate. A continuous scenario where $Z \sim \mathrm{N}(0, 1)$ and a binary case $Z \sim \text{Bernoulli}(p = 0.3)$. This was done to mimic a standardized continuous covariate and a categorical yes/no covariate, respectively. Within each of these scenarios, both a continuous exposure and a binary exposure were investigated.

The specifics of each simulation scenario are listed in the respective tables that follow. Each scenario had a sample size of $n = 500$ subjects. 500 data sets were simulated and the posterior means in the BSP case and the means of the estimates of the CLR across the 500 simulations are listed. The true parameter values are in the footnote of each table in the following section.

5.4.2 Simulation Results

The following pages contain tables of output summaries for the simulation study. The details of each simulation study are described in the table captions.

		Estimation Method					
		BSP			CLR		
Parameter	True Value[1]	Mean[2]	MSE[3]	95% PI[2]	Mean	MSE[3]	95% CI
β_1	1.0000	0.9896	0.0020	(0.9000.1.0850)	0.9832	0.0028	(0.8838,1.0825)
$\widetilde{\beta}_2$	0.9553	0.9740	0.0020	—	0.7000	0.0686	(0.6317,0.7682)
$\widetilde{\gamma}_0$	0.3153	-1.0010	0.0012	—	**	**	**
γ_1	0.3000	0.3000	0.0003	(0.2686,0.3240)	**	**	**
σ^2	1.0000	0.9806	0.0008	(0.9435,1.0200)	**	**	**
β_1	1.0000	0.9915	0.0007	(0.9091,1.0772)	0.9978	0.0017	(0.9066,1.0889)
$\widetilde{\beta}_2$	0.9553	0.9746	0.0025	—	0.7221	0.0559	(0.6554,0.7887)
$\widetilde{\gamma}_0$	0.3153	-1.0026	0.0012	—	**	**	**
γ_1	0.0500	0.0492	0.0001	(0.0222,0.0764)	**	**	**
σ^2	1.000	0.9785	0.0004	(0.9414,1.0155)	**	**	**

Table 5.1: Simulation summaries for continuous exposure. $\mu_\beta \in \{0,1,2\}$. Covariate Z simulated as $Z \sim \mathrm{N}(0,1)$. — for PSB method means no suitable estimate, as a PI is computed for each subject's specific effect. ** for CLR method means this parameter is not estimated in this method.
The first and second set of values differ in the γ_1 parameter (the coefficient of Z in the mean of X).

[1] True values for $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ defined to be $\mathrm{E}(\beta_{2i})$ and $\mathrm{E}(\gamma_{0i})$ respectively, as derived in (5.9) and (5.10).

[2] Probability intervals for the BSP method taken as the mean of the individual PI's across the 500 simulations. BSP estimates for the subject specific effects for a given dataset are taken as the mean of the individual posterior means across all subjects. The overall estimate is taken as the mean of these values across all simulations.

[3] MSE for subject specific effects computed with respect to the truth being that which is described in footnote 1.

		Estimation Method					
		BSP			CLR		
Parameter	True Value[1]	Mean[2]	MSE[3]	95% PI[2]	Mean	MSE[3]	95% CI
β_1	1.0000	0.9905	0.0053	(0.8489,1.1323)	0.9913	0.0061	(0.8431,1.1394)
$\widetilde{\beta}_2$	0.9553	0.9536	0.0021	–	0.7000	0.0665	(0.6388,0.7599)
$\widetilde{\gamma}_0$	-1.0000	-1.0003	0.0011	–	**	**	**
γ_1	0.3000	0.3017	0.0001	(0.2451,0.3566)	**	**	**
σ^2	1.0000	0.9892	0.0009	(0.9524,1.0273)	**	**	**
β_1	1.0000	0.9942	0.0068	(0.8602,1.1311)	1.0020	0.0070	(0.8589,1.1451)
$\widetilde{\beta}_2$	0.9553	0.9513	0.0022	–	0.7118	0.0604	(0.6516,0.7717)
$\widetilde{\gamma}_0$	-1.0000	-1.0004	0.0012	–	**	**	**
γ_1	0.0500	0.0542	0.0011	(-0.0007,0.1091)	**	**	**
σ^2	1.0000	0.9903	0.0008	(0.9528,1.0295)	**	**	**

Table 5.2: Simulation summaries for continuous exposures. $\mu_\beta \in \{0,1,2\}$. Covariate Z simulated as $Z \sim \text{Bernoulli}(p = 0.3)$. — for PSB method means no suitable estimate, as a PI is computed for each subject;s specific effect. ** for CLR method means this parameter is not estimated in this method.
The first and second set of values differ in the γ_1 parameter (the coefficient of Z in the mean of X).

[1] True values for $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ defined to be $\text{E}(\beta_{2i})$ and $\text{E}(\gamma_{0i})$ respectively, as derived in (5.9) and (5.10).

[2] Probability intervals for the BSP method taken as the mean of the individual PI's across the 500 simulations. BSP estimates for the subject specific effects for a given dataset are taken as the mean of the individual posterior means across all subjects. The overall estimate is taken as the mean of these values across all simulations.

[3] MSE for subject specific effects computed with respect to the truth being that which is described in footnote 1.

		Estimation Method					
		BSP			CLR		
Parameter	True Value[1]	Mean[2]	MSE[3]	95% PI[2]	Mean	MSE[3]	95% CI
β_1	1.0000	0.9907	0.0023	(0.9080,1.0780)	0.9870	0.0023	(0.90154,1.0724)
$\widetilde{\beta}_2$	0.3153	0.3296	0.0009	–	0.3092	0.0013	(0.2533,0.3653)
$\widetilde{\gamma}_0$	-1.000	-1.0026	0.0012	–	**	**	**
γ_1	0.3000	0.2973	0.0003	(0.2691,0.3250)	**	**	**
σ^2	1.0000	0.9672	0.0015	(0.9319,1.0040)	**	**	**
β_1	1.0000	0.9944	0.0020	(0.9145,1.0779)	0.9960	0.0020	(0.9150,1.0781)
$\widetilde{\beta}_2$	0.3153	0.3280	0.0009	–	0.3003	0.0012	(0.2557,0.3700)
$\widetilde{\gamma}_0$	-1.000	-1.0017	0.0011	–	**	**	**
γ_1	0.0500	0.0498	0.0003	(0.0217,0.0768)	**	**	**
σ^2	1.0000	0.9667	0.0015	(0.9312,1.0030)	**	**	**

Table 5.3: Simulation summaries for continuous exposures. $\mu_\beta \in \{0, 0.3, 0.7\}$. Covariate Z simulated as $Z \sim \mathrm{N}(0,1)$. — for PSB method means no suitable estimate, as a PI is computed for each subject's specific effect. ** for CLR method means this parameter is not estimated in this method. The first and second set of values differ in the γ_1 parameter (the coefficient of Z in the mean of X).

[1] True values for $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ defined to be $\mathrm{E}(\beta_{2i})$ and $\mathrm{E}(\gamma_{0i})$ respectively, as derived in (5.9) and (5.10).

[2] Probability intervals for the BSP method taken as the mean of the individual PI's across the 500 simulations. BSP estimates for the subject specific effects for a given dataset are taken as the mean of the individual posterior means across all subjects. The overall estimate is taken as the mean of these values across all simulations.

[3] MSE for subject specific effects computed with respect to the truth being that which is described in footnote 1.

		Estimation Method					
		BSP			CLR		
Parameter	True Value[1]	Mean[2]	MSE[3]	95% PI[2]	Mean	MSE[3]	95% CI
β_1	1.0000	0.9934	0.0037	(0.8661,1.1213)	0.9354	0.0040	(0.8168,1.0540)
$\widetilde{\beta}_2$	0.3153	0.3151	0.0006	–	0.3000	0.0013	(0.2456,0.3510)
$\widetilde{\gamma}_0$	-1.000	-0.9999	0.0011	–	**	**	**
γ_1	0.3000	0.3014	0.0010	(0.2445,0.3571)	**	**	**
σ^2	1.0000	0.9767	0.0011	(0.9408,1.0135)	**	**	**

Table 5.4: Simulation summaries for continuous exposures. $\mu_\beta \in \{0, 0.3, 0.7\}$. Covariate Z simulated as $Z \sim$ Bernoulli($p = 0.3$). — for PSB method means no suitable estimate, as a PI is computed for each subject's specific effect. ** for CLR method means this parameter is not estimated in this method.

[1] True values for $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ defined to be E(β_{2i}) and E(γ_{0i}) respectively, as derived in (5.9) and (5.10).

[2] Probability intervals for the BSP method taken as the mean of the individual PI's across the 500 simulations. BSP estimates for the subject specific effects for a given dataset are taken as the mean of the individual posterior means across all subjects. The overall estimate is taken as the mean of these values across all simulations.

[3] MSE for subject specific effects computed with respect to the truth being that which is described in footnote 1.

		Estimation Method					
		BSP			CLR		
Parameter	True Value[1]	Mean	MSE[2]	95% PI	Mean	MSE[2]	95% CI
β_1	1.0000	1.0048	0.0003	(0.9890,1.1073)	1.0944	0.0094	(1.0189,1.1698)
$\widetilde{\beta}_2$	0.9553	0.9583	0.0015	–	0.9480	0.0040	(0.819,1.0774)
$\widetilde{\gamma}_0$	-1.000	-1.0241	0.0010	–	**	**	**
γ_1	0.3000	0.3029	0.0006	(0.2552,0.3509)	**	**	**
β_1	1.0000	1.0047	0.0007	(0.9504,1.0599)	0.9858	0.0017	(0.9168,1.0547)
$\widetilde{\beta}_2$	0.9553	0.9741	0.0022	–	0.9613	0.0038	(0.8368,1.0857)
$\widetilde{\gamma}_0$	-1.000	-1.0330	0.0012	–	**	**	**
γ_1	0.0500	0.0495	0.0006	(0.0040,0.0955)	**	**	**

Table 5.5: Simulation summaries for binary exposure. $\mu_\beta \in \{0, 1, 2\}$. Covariate Z simulated as $Z \sim \mathrm{N}(0, 1)$. — for PSB method means no suitable estimate, as a PI is computed for each subject's specific effect. ** for CLR method means this parameter is not estimated in this method.
The first and second set of values differ in the γ_1 parameter (the coefficient of Z in the mean of X).

[1] True values for $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ defined to be $\mathrm{E}(\beta_{2i})$ and $\mathrm{E}(\gamma_{0i})$ respectively, as derived in (5.9) and (5.10).

[2] Probability intervals for the BSP method taken as the mean of the individual PI's across the 500 simulations. BSP estimates for the subject specific effects for a given dataset are taken as the mean of the individual posterior means across all subjects. The overall estimate is taken as the mean of these values across all simulations.

[3] MSE for subject specific effects computed with respect to the truth being that which is described in footnote 1.

		Estimation Method					
		BSP			CLR		
Parameter	True Value[1]	Mean	MSE[2]	95% PI	Mean	MSE[2]	95% CI
β_1	1.0000	1.0088	0.0005	(0.9990,1.1468)	1.0354	0.0090	(0.9840,1.0900)
$\widetilde{\beta}_2$	0.3153	0.2982	0.0025	–	0.3326	0.0030	(0.2370,0.4290)
$\widetilde{\gamma}_0$	-1.000	-1.0190	0.0010	–	**	**	**
γ_1	0.3000	0.3068	0.0006		**	**	**
β_1	1.0000	1.0137	0.0007	(0.9596,1.0701)	0.9359	0.0011	(0.8904,0.9874)
$\widetilde{\beta}_2$	0.3153	0.3000	0.0030	–	0.3300	0.0050	(0.2260,0.4152)
$\widetilde{\gamma}_o$	-1.000	-1.0180	0.0010	–	**	**	**
γ_1	0.0500	0.0527	0.0006	(0.0078,0.0980)	**	**	**

Table 5.6: Simulation summaries for binary exposure. $\mu_\beta \in \{0, 0.3, 0.7\}$. Covariate Z simulated as $Z \sim \mathrm{N}(0,1)$. — for PSB method means no suitable estimate, as a PI is computed for each subject's specific effect. ** for CLR method means this parameter is not estimated in this method.
The first and second set of values differ in the γ_1 parameter (the coefficient of Z in the mean of X).

[1] True values for $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ defined to be $\mathrm{E}(\beta_{2i})$ and $\mathrm{E}(\gamma_{0i})$ respectively, as derived in (5.9) and (5.10).

[2] Probability intervals for the BSP method taken as the mean of the individual PI's across the 500 simulations. BSP estimates for the subject specific effects for a given dataset are taken as the mean of the individual posterior means across all subjects. The overall estimate is taken as the mean of these values across all simulations.

[3] MSE for subject specific effects computed with respect to the truth being that which is described in footnote 1.

		Estimation Method					
		BSP			CLR		
Parameter	True Value[1]	Mean	MSE[2]	95% PI	Mean	MSE[2]	95% CI
β_1	1.0000	1.0397	0.0030	(0.9512,1.1294)	1.1668	0.0310	(1.0872,1.2460)
$\widetilde{\beta}_2$	0.9553	0.9877	0.0022	–	0.9403	0.0041	(0.8587,1.0280)
$\widetilde{\gamma}_0$	-1.000	-1.0419	0.0011	–	**	**	**
γ_1	0.3000	0.3132	0.0017	(0.2213,0.4043)	**	**	**

Table 5.7: Simulation summaries for binary exposure. $\mu_\beta \in \{0, 1, 2\}$. Covariate Z simulated as $Z \sim$ Bernoulli($p = 0.3$). — for PSB method means no suitable estimate, as a PI is computed for each subject's specific effect. ** for CLR method means this parameter is not estimated in this method.

[1] True values for $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ defined to be E(β_{2i}) and E(γ_{0i}) respectively, as derived in (5.9) and (5.10).

[2] Probability intervals for the BSP method taken as the mean of the individual PI's across the 500 simulations. BSP estimates for the subject specific effects for a given dataset are taken as the mean of the individual posterior means across all subjects. The overall estimate is taken as the mean of these values across all simulations.

[3] MSE for subject specific effects computed with respect to the truth being that which is described in footnote 1.

		Estimation Method					
		BSP			CLR		
Parameter	True Value[1]	Mean	MSE[2]	95% PI	Mean	MSE[2]	95% CI
β_1	1.0000	1.0685	0.0059	(0.9830,1.1566)	1.1684	0.0300	(1.0090,1.1246)
$\widetilde{\beta}_2$	0.3153	0.3000	0.0020	–	0.3263	0.0033	(0.2423,0.4110)
$\widetilde{\gamma}_0$	-1.000	-1.0200	0.0011	–	**	**	**
γ_1	0.3000	0.3129	0.0019	(0.2220,0.4055)	**	**	**

Table 5.8: Simulation summaries for binary exposure. $\mu_\beta \in \{0, 0.3, 0.7\}$. Covariate Z simulated as $Z \sim \text{Bernoulli}(p = 0.3)$. — for PSB method means no suitable estimate, as a PI is computed for each subject's specific effect. ** for CLR method means this parameter is not estimated in this method.

[1] True values for $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ defined to be $\text{E}(\beta_{2i})$ and $\text{E}(\gamma_{0i})$ respectively, as derived in (5.9) and (5.10).

[2] Probability intervals for the BSP method taken as the mean of the individual PI's across the 500 simulations. BSP estimates for the subject specific effects for a given dataset are taken as the mean of the individual posterior means across all subjects. The overall estimate is taken as the mean of these values across all simulations.

[3] MSE for subject specific effects computed with respect to the truth being that which is described in footnote 1.

Table 5.1 through Table 5.4 contain simulation summaries for the continuous exposure case and Table 5.5 through Table 5.8 contain summaries for the binary exposure. The frequentist CLR method implemented in this chapter uses the independent likelihood discussed in Chapter 4.3.2. Across all scenarios, the BSP method performs better than the CLR method in terms of mean squared error (MSE) on all parameters when comparable (the CLR method will not obtain estimates for γ_1, σ^2 or the γ_{0i}'s). The parameter β_1 is the coefficient on the covariate Z in the prospective probability model of an event. Using a probability model of the logistic form, it is interpreted as the log odds ratio of comparing the odds of an event with covariate value $Z+1$ to odds of an event with covariate value Z. β_{2i} is the subject specific coefficient on the exposure in the prospective probability model of an event. Additionally γ_1 and γ_{0i} (and in the continuous exposure case also σ^2) are factors that model the exposure distribution within each subject.

The CLR method will not produce estimates of the individual subject effects (β_{2i}) and will only estimate a single marginal parameter. In order to compare the BSP method to the CLR method, the true marginal parameter value, $\widetilde{\beta}_2$, is defined as the mean of the mixture of normals the subject specific effects are generated from. The BSP method's estimate of this marginal parameter is taken to be the mean of all the subject specific effect's posterior means. With respect to obtaining a single marginal parameter, $\widetilde{\beta}_2$, the BSP method outperforms the CLR method across all scenarios.

Although β_1 is not the parameter of interest, the BSP method has a lower MSE for it across all simulation scenarios. Additionally, the BSP method performs fairly well for estimating γ_1 and the γ_{0i}'s (and σ^2 in the continuous exposure scenario). The parameter of interest is $\widetilde{\beta}_2$ and that is what will be discussed. In the continuous exposure simulations, when $\mu_\beta \in \{0, 1, 2\}$ the MSE of the BSP method is about 0.0020 while the CLR method's MSE is 0.0600 (across both scenarios of a Gaussian and binary covariate Z). When the exposure is continuous and $\mu_\beta \in \{0, 0.3, 0.7\}$, the MSE of BSP is about 0.0008 and for the CLR method it is about 0.0013 (across both scenarios of a Gaussian and binary covariate Z). Results are similar when the exposure is binary. Across both scenarios of a Gaussian and binary covariate Z, the BSP method results in an MSE of about 0.0020 and the CLR method results in an MSE of about 0.0040 when $\mu_\beta \in \{0, 1, 2\}$. When $\mu_\beta \in \{0, 0.3, 0.7\}$, the BSP method has an MSE of roughly 0.0025 while the CLR method has an MSE of roughly 0.0035.

As can be seen the BSP method has lower MSE across all simulation scenarios of exposure and covariate type. If the goal of the scientific study is to obtain a consistent and unbiased marginal parameter estimate of the effect of an exposure on the risk of an event, while weighting each subject equally, then the BSP method is the preferred choice over the CLR approach. The main advantage of the BSP method is that it generates estimates of the subject specific effects for each subject. This will allow researchers to investigate the potential presence of effect modification across subjects while also being able to obtain a single marginal estimate. The mean squared prediction error (MSPE) of the BSP method with respect to the subject specific effects β_{2i} is computed as the mean of the squared differences between the true effect value and the posterior mean of the effect produced by the BSP method, across all subjects. That is to say MSPE=$\frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_{2i} - \beta_{2i})^2$, where n is the number of subjects.

When the mixture of normals that the β_{2i}'s are being generated from do not have considerable overlap, as is the case when $\mu_\beta \in \{0, 1, 2\}$), the BSP method will have an MSPE of about 0.17 (across the different Z generation and γ_1 values) for the continuous exposure scenario. In the binary exposure case, the MSPE is about 0.30 across the different scenarios of Z and γ_1. In the binary exposure case, there is less information about β_{2i} being contributed to the likelihood among the individual subjects, as is the case when a model estimates the effect of a single binary exposure on the risk of an event. Figure 5.1 has a representative graph that plots the posterior means of the β_{2i}'s against the true β_{2i} value for the continuous exposure scenario and Figure 5.3 for the binary exposure. The line represents the 45-degree line.

Plotted points are centered about the 45-degree line, indicating a fair degree of accuracy for estimating the subject specific parameters. Additionally, the distinct grouping moving along the x-axis shows clustering. The different symbols (circle, square, or triangle) signify which of the 3 normals the subject specific effect was generated from. Average number of occupied clusters across iterates was 6 for both the continuous and binary exposure simulations with a majority of the subjects belonging to 3 clusters.

Observe that the standard deviations of the Gaussian distributions comprising the mixture from which the subject specific effects are generated from are 0.4. As a result when $\mu_\beta \in \{0, 0.3, 0.7\}$, there is considerable overlap in the densities. In this scenario, the BSP method will have a MSPE roughly equal to 0.06 when the exposure is continuous and 0.10 when the exposure is binary. Figure 5.2 has a representative graph that plots the posterior means of the β_{2i}'s against the true β_{2i} value for the continuous exposure scenario and Figure 5.4 for the binary exposure case. The average number of occupied clusters across iterates was 4.5, with a majority of subjects occupying a single cluster. Clustering is not as evident in these plots due the fact that the normals used to create the mixture that generates the β_{2i}'s have considerable overlapping densities. The points are still concentrated about the 45-degree line. A single parametric prior with hyper-priors for the parameters would probably suffice in this scenario. But prior to conducting an analysis, one does not know if the true subject specific effect generating distribution is a separable mixture or not. Additionally, Figure 5.5 and Figure 5.6 present graphs which plot the true value of the γ_{oi}'s against the posterior means. These figures indicate a fair degree of accuracy in estimating the individual subject specific effects of γ_{0i} across all subjects.

When a single parametric prior for the subject specific effects would have sufficed, the Dirichlet process mixture prior for the subject specific effects will only be slightly less efficient. All matched sets will tend to create a single cluster, but the probability that a new cluster will be created is non-zero, and thus throughout MCMC iterations, new clusters will be formed causing its members to deviate from the center of the true data generating distribution.

As the number of expected events per matched set (subject) increases, the MSPE will decrease. In the continuous case when $\mu_\beta \in \{0, 1, 2\}$, increasing the number of expected events from 6 to 10 (number of events for each subject sampled according to a uniform{9,10,11}) decreased MPSE from an about 0.17 to 0.11 and in the binary case, going from an expected number of events per subject of 8 to 14 (sampled from a uniform{13,14,15}) decreased MPSE from an about 0.30 to 0.25. When $\mu_\beta \in \{0, 0.3, 0.7\}$, increasing the number events per subject from 6 to 10 in the continuous case resulted in MSPE decreasing from 0.06 to 0.04, and in the binary exposure case going from 8 to 16 events per subject decreased MSPE from 0.10 to 0.08.

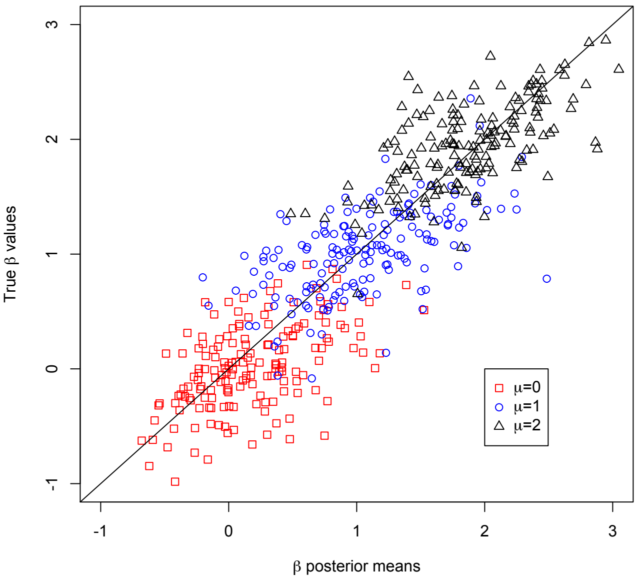

Figure 5.1: Continuous exposure simulation. $\mu_{\beta_{2i}} = \{0, 1, 2\}$. Plot of true β_{2i} against posterior means. The symbol of the plotted point designates which of the Gaussian distributions the true value was generated from.

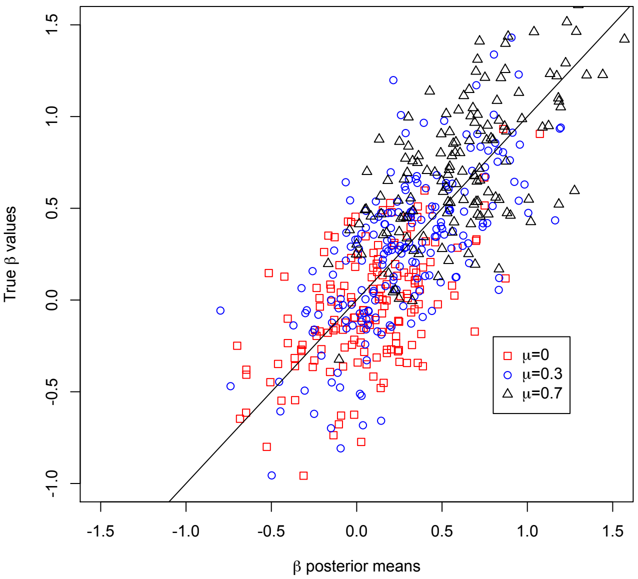

Figure 5.2: Continuous exposure simulation. $\mu_{\beta_{2i}} = \{0, 0.3, 0.7\}$. Plot of true β_{2i} against posterior means. The symbol of the plotted point designates which of the Gaussian distributions the true value was generated from.

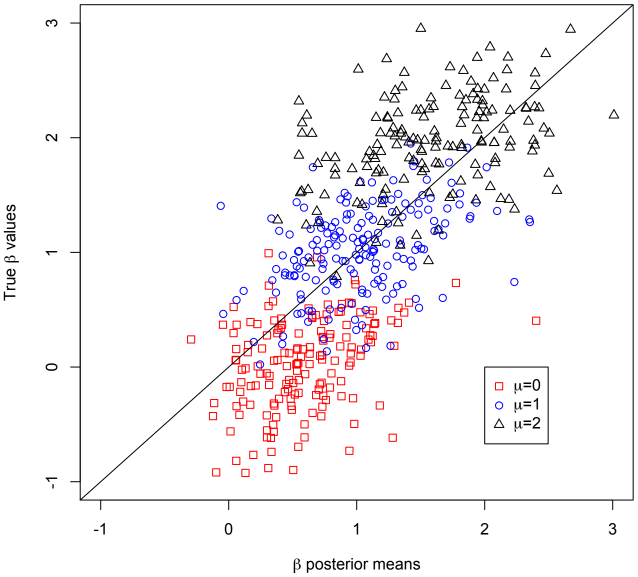

Figure 5.3: Binary exposure simulation. $\mu_{\beta_{2i}} = \{0, 1, 2\}$. Plot of true β_{2i} against posterior means. The symbol of the plotted point designates which of the Gaussian distributions the true value was generated from.

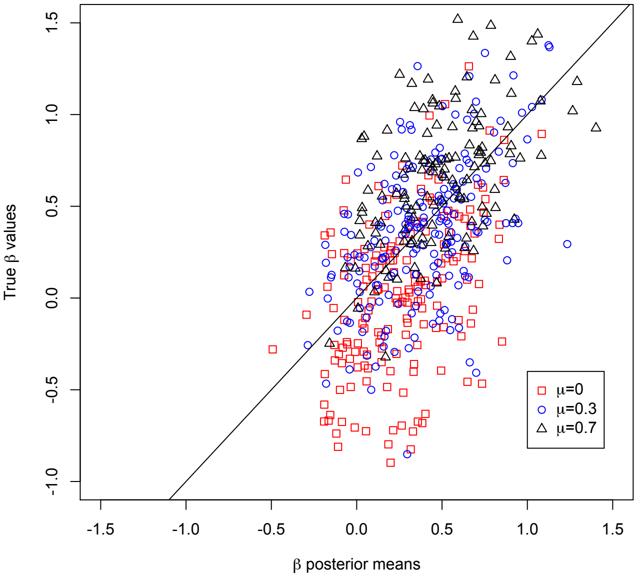

Figure 5.4: Binary exposure simulation. $\mu_{\beta_{2i}} = \{0, 0.3, 0.7\}$. Plot of true β_{2i} against posterior means. The symbol of the plotted point designates which of the Gaussian distributions the true value was generated from.

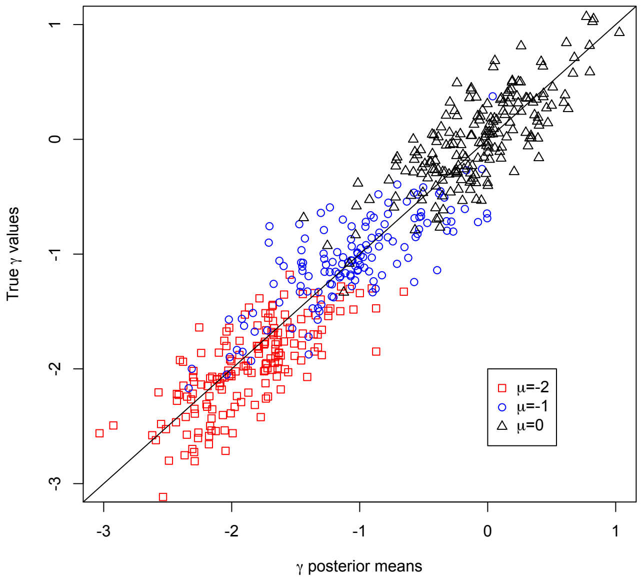

Figure 5.5: Continuous exposure simulation. Plot of true γ_{0i} against posterior means. The symbol of the plotted point designates which of the Gaussian distributions the true value was generated from.

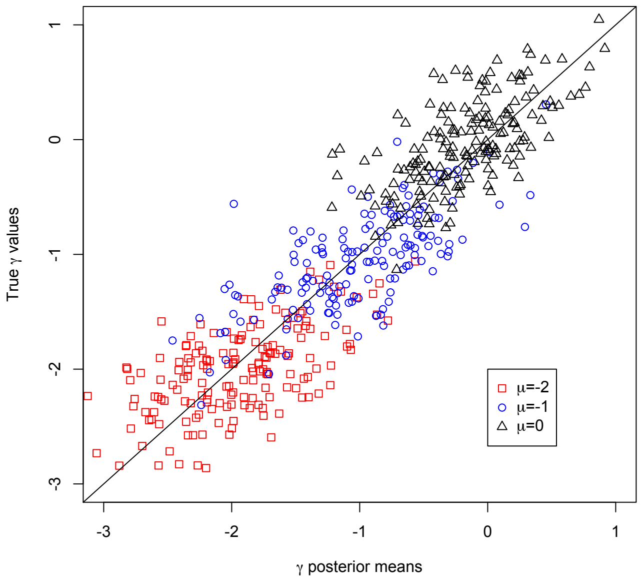

Figure 5.6: Binary exposure simulation. Plot of true γ_{0i} against posterior means. The symbol of the plotted point designates which of the Gaussian distributions the true value was generated from.

Prior parameter values that were used are shown in for the simulation studies are in Appendix 5.7.4. A sensitivity analysis was conducted, and the conclusion was the model was not sensitive to the prior specifications when prior parameter values were within reason. For the fixed effect parameters of β_1 and γ_1, prior values investigated were between -2 and 2 for the mean of the specified normal prior, and between 3-10 for the variance of the specified normal prior. By putting a higher variance on the normal priors, a larger space was being explored in the Metropolis-Hastings sampling, but this resulted in a lower acceptance rate. For the fixed hyperprior parameters in the DP, the inverse gamma parameter values investigated were between 1-15 for a and between 1-15 for b. The parameter values investigated for the prior of α ranged from 1-20 for a and 1-20 for b.

In the scenario of a continuous exposure, the prior on σ^2, which was an IG(a, b) was partially sensitive to prior values. Choosing values for a and b when the conditional variance of X given Z and D $(X|Z, D)$ is small (i.e. in a standardized case), such as $a = 2$ and $b = 50$ (resulting in a prior mean of 50 and a majority of the density between 18.5 and 52), would cause the chain to be stagnant for multiple iterations across the MCMC draws. A solution is to conduct a prior summary of the unconditional variance of X and choose parameters for the inverse gamma prior accordingly.

Chain convergence was assessed using the method proposed by Gelman and Rubin [1992]. This diagnostic method runs multiple chains (10 in this case) whose starting values are sampled from an over-dispersed distribution. The within chain and across chain variances are computed. A statistic is created which is the ratio of a weighted sum of the within chain and across chain variances to the average within chain variances. If this statistic is close to 1, then it is evident that the chains have converged as they would have if the number of MCMC draws were infinite. All the statistics obtained from the separate simulations used above returned values very close to 1. Between 50,000-100,000 MCMC samples were taken for any given dataset, and an autocorrelation summary was used to determine the amount of thinning needed (between 5-15 thinning was employed). A burn-in period of 2000 samples was used.

5.5 Application to the Asthma-Related Hospital Encounters Data

The BSP method described in this chapter is applied to the hospital admission data obtained from the air pollution study that has been the focus of the applied examples throughout the dissertation. A review of this study is in Chapter 1.1. The dataset used here was restricted to only those subjects with more than 2 observed events given the focus on subject-specific parameter estimates (781 subjects). This is done to highlight the advantages of the random effects model presented, and that requires numerous observations from each cluster/subject. For scaling purposes, the exposure and covariate used were standardized. Adjustment covariates of relative humidity and temperature were combined to create a single heat index covariate. The formula to create the heat index based on relative humidity and temperature is the one set forth by the National Weather Service.

In all models, the coefficient of β_1 is the log odds ratio for a unit increase in the heat index and β_{2i} is for the exposure. Prior information was used to obtain hyper-parameter values based on previous research by Chang et al. [2009] and Delfino et al. [2014] on similar air pollution studies. These values are given in the appendix.

Method		$OR_{\beta_1}{}^2$	$OR_{\widetilde{\beta}_2}{}^1$	$\widetilde{\gamma}_0{}^1$	γ_1	σ^2
BSP						
	Mean	0.9638	1.0714	-0.0329	-0.0400	0.8300
	S.D.	1.0510	**	**	0.0198	0.0250
	95% PI	(0.8724,1.0597)	**	**	(-0.0811,-0.0029)	(0.7820,0.8794)
CLR						
	Est.	0.9768	1.0661	—	—	—
	S.E.	1.0511	1.0432	—	—	—-
	95% CI	(0.9185,1.0445)	(0.9842,1.1136)	—	—	—

Table 5.9: Applied output using exposure of $PM_{2.5}$ lag 1. Data stratified to subjects in zipcodes which have the percent of the population living below the poverty line being above the median of Orange county. — for PSB method means no suitable estimate, as a PI is computed for each subject's parameter. ** for CLR method means this parameter is not estimated in the model.

Method		$OR_{\beta_1}{}^2$	$OR_{\widetilde{\beta}_2}{}^1$	$\widetilde{\gamma}_0{}^1$	γ_1	σ^2
BSP						
	Mean	0.9716	1.1264	-0.0509	-0.0407	0.8500
	S.D.	1.0653	**	**	0.0262	0.0320
	95% PI	(0.8583,1.1042)	**	**	(-0.0407,0.0634)	(0.7841,0.9087)
Ind.						
	Est.	0.9910	1.1275	—	—	—
	S.E.	1.0685	1.0618	—	—	—
	95% CI	(0.8703,1.1282)	(1.0025,1.2682)			

Table 5.10: Predictor of interest $PM_{2.5}$ lag 0, on stratified white non-Hispanics data. — for PSB method means no suitable estimate, as a PI is computed for each subject's parameter. ** for CLR method means this parameter is not estimated in the model.

[1] $\widetilde{\beta}_2$ and $\widetilde{\gamma}_0$ estimates for the BSP method computed as the mean of the posterior means of the subject specific effects across all subjects.

[2] β_1 represents the effect of the heat index covariate on the risk of an event.

Table 5.9 presents estimates output for a model using data from subjects living in zip codes that have the percent of residents living below the poverty line to be above the median for Orange county. The exposure is $\mathrm{PM}_{2.5}$ with a lag of 1 day and the covariate is heat index with a lag of 1 day. This stratified dataset could be viewed as representing a lower socio-economic demographic. Observe that the BSP and CLR methods used in this chapter will maintain the matched case-control pair bonds, and as a result it is no longer needed to stratify the dataset on season.

It can be seen in Table 5.9 that both the BSP and CLR method result in similar estimates for β_1 and $\widetilde{\beta}_2$. Similar inferences of the marginal parameter will be obtained using either method. Based on the simulation results, it is reasonable to think using the BSP method will obtain more accurate marginal estimates than the CLR method. A strength of the BSP method is mentioned in the simulation studies. This method can estimate the subject specific effects across all subjects. Figure 5.7 plots a random sample of subjects 95% probability intervals for the estimated odds ratio of an event with respect to the exposure (sorted by magnitude of the posterior mean of β_{2i}). The probability intervals for this model suggest that not all subjects exhibit the same effect of the exposure on the risk of an event. Across iterates, a majority of subjects belonged to a single cluster (roughly 80%), but approximately 15% of subjects were members of another cluster. The posterior mean of the concentration parameter, α_β, of the DP prior on the prior mean of the β_{2i}'s was 1.75. This suggests that there could possibly be 2 subpopulations within the sample with regards to the effect of the exposure on the risk of an event.

Table 5.10 presents another example. The dataset used in this example contains only subjects who identify themselves as being white non-Hispanics. As with the previous example, both the BSP method and CLR method will result in similar marginal parameter estimates. Figure 5.8 displays a random sample of subjects 95% probability intervals for the estimated odds ratio of an event with respect to the exposure (sorted by magnitude of the posterior mean of β_{2i}). Similar to the previous example, this plot suggests that the effect of the exposure on the risk of an event is not constant across subjects. Across iterates, 75% of subjects are clustered together, while %15 of subjects create another cluster. The posterior mean of the concentration parameter, α_β, of the DP prior on the mean of the β_{2i}'s was 2.5 across the in this model.

Dendrograms presented in Appendix 5.7.8 highlight the clustering described for the two illustrated examples presented. A dendrogram is a branching diagram that represents the relationships of similarity among a group of entities, where the relative drop in the dendrogram represents the relative reduction in prediction error conditional upon the clustering at that branch. Both figures show that a majority of subjects are being grouped into 2 clusters.

Across these 2 illustrations, the posterior means of the individual subject effects estimates are not equal across all subjects. As discussed, there is evidence of clustering occurring among the subjects, indicating the possibility of effect modification being induced by the presence of 2 subpopulations of subjects in the sample. Additionally some subjects have some credible intervals completely above 0 while others don't. The scientific significance of this illustration is that obtaining a single marginal parameter will implicitly assume a constant effect of the exposure on the risk of an event across all subjects (all subjects are from the same population with respect to the effect). A single marginal parameter will not accurately address the effect of the exposure on the risk of an event for all subjects. The BSP method presented in this chapter is shown to alleviate the issue of using a single marginal parameter to make inference when effect modification is present across subjects.

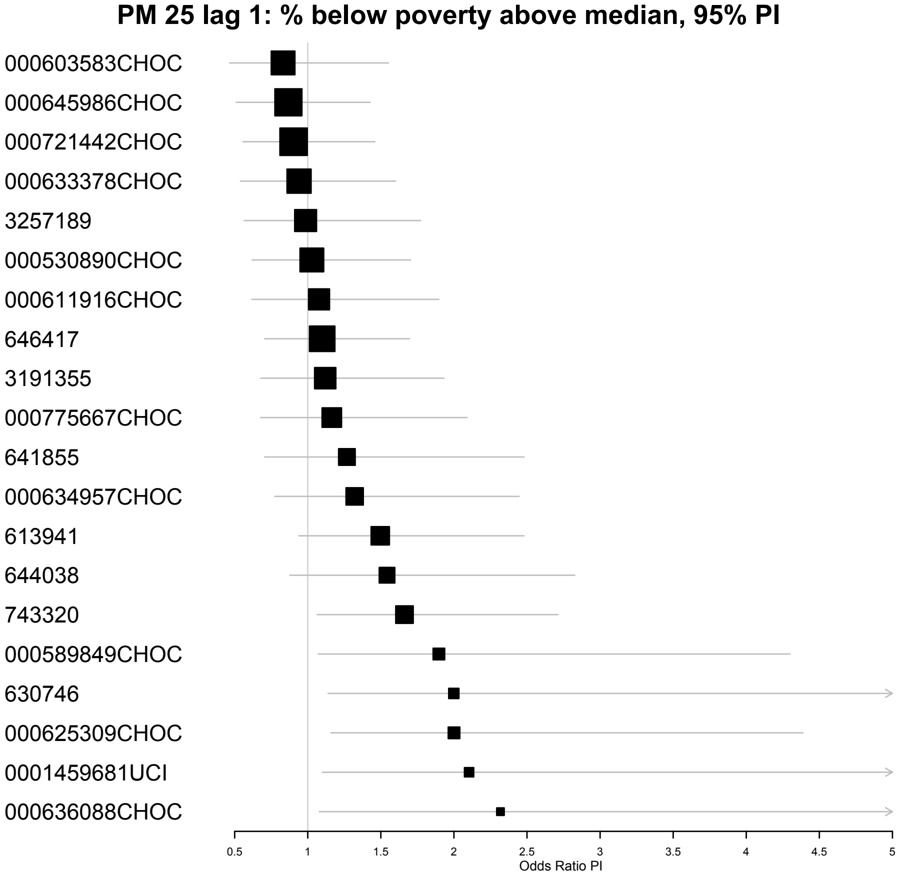

Figure 5.7: Plot of probability intervals of event odds ratios with respect to a change in pm2.5 lag 1 for randomly selected subjects. Dataset used was stratified to subjects living in zipcodes that have the percent of residents living below the poverty line being above the median for Orange county. The y-axis represents the id of the randomly selected subjects.

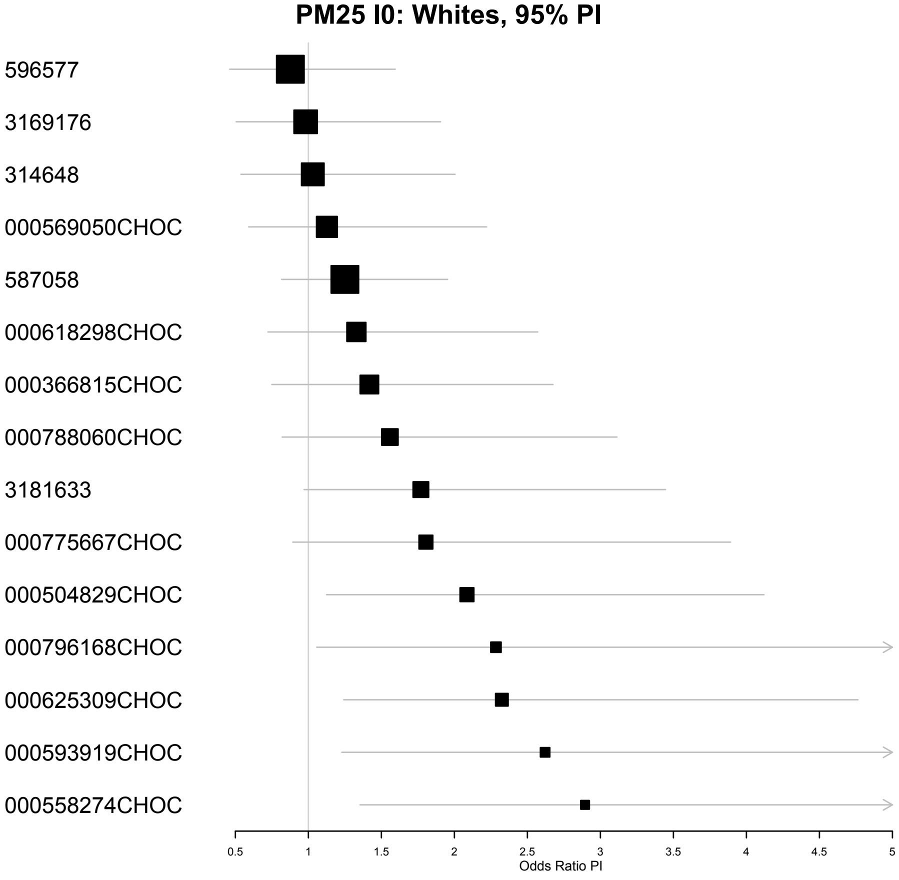

Figure 5.8: Plot of probability intervals of event odds ratios with respect to a change in pm2.5 lag 0 for randomly selected subjects. Dataset used was stratified to subjects who are white non-Hispanics. The y-axis represents the id of the randomly selected subjects.

5.6 Discussion

In matched case-control studies, it is often reasonable to hypothesize effect modification across matched sets (subjects in the case-crossover design). In such a scenario, a single marginal parameter estimated in the analysis will not universally address all the matched sets (subjects). It was shown in Chapter 4 that if the matched set sizes differ and it is no longer acceptable to break the individual matched case-control pair bonds, then researchers must decide on the estimand of interest prior to conducting the analysis. Depending on the estimand chosen, the estimation procedure will assign equal weight to the matched sets or equal weight to the individual matched pairs. When there is effect modification across the matched sets, the two weighting schemes will result in substantially different estimates being obtained.

An inefficient approach would be to conduct an analysis for each subject. This would result in each subject specific parameter to be estimated with only a few observations. The ideal approach would obtain estimates of the subject specific coefficients for each subject, and additionally would group similar subjects with respect to their effect of exposure on risk of event (i.e. the coefficient value). This would allow for clustering of subjects and borrowing of information across similar subjects. In a likelihood maximization approach, a fully specified parametric distribution would need to be specified for the subject specific effects (random effects). The augmented data likelihood would then have to be marginalized over the random effects (by numerical integration) to obtain estimates for the fixed effects in the model, as well as the parameters that specify the distribution of the random effects. Misspecification of the specified distribution will lead to non-robust estimates, and also the numerical integration can be intensive when dealing with non-standard likelihood. In this chapter, a Bayesian semi-parametric hierarchical model (BSP) was proposed to best address these issues. The BSP method will allow for a flexible class of distribution to be the prior of the subject specific effects.

A Dirichlet process mixture prior is specified as the prior of the subject specific effects. By specifying a prior of the form of a Gaussian distribution on the subject specific effects, and then specifying a Dirichlet process prior on the parameters of this distribution, a Dichlet process mixture model is specified as the prior of the subject specific effects. It was discussed in Section **??** that a Dirichlet process prior would be specified for the the parameters of the normal prior of γ_{0i} and for the normal prior of β_{2i}. The γ_{0i}'s capture the stratum effects on the exposure distribution, and the β_{2i}'s are the stratum specific coefficients on the exposure in the prospective probability of an event.

The simulation results presented in Section 5.4.2 highlight that the BSP method has lower MSE for the marginal parameter estimates over the CLR method across all simulations settings of exposure type and covariate type. The marginal values for the subject specific parameters were described in (5.9) and (5.10). An advantage of the BSP method is it will result in obtaining estimates of the subject specific effects. This will allow researchers to investigate the presence of effect modification, while being able to obtain a single marginal estimate (by taking the mean of the individual subject specific effect estimates).

When the mixture of normals that the subject specific effects arise from do not have considerable overlapping densities, then the BSP method is able to cluster the subject specific effects accurately (Figure 5.1 and Figure 5.3). When the means of the normals in the mixture are not a substantially distant from one another (or the variances of the individual normals is large) then there is considerable overlap of the densities. In this scenario, it is not apparent that the density generated is a mixture. The BSP method still results in fairly accurate estimates, as shown in Figure 5.2 and Figure 5.4. The illustrative applied examples in section showed that across subjects, the posterior means of the subject specific effects varied with similarities between some subjects. The clustering of the subjects across iterations suggest that the sample of subjects are arising from two different populations.

In the presence of multicollinearity between covariate and exposure, the BSP method will explicitly model this association (through the $\boldsymbol{\gamma_1}$ parameter). By obtaining estimates for the stratum specific effects of γ_{0i}, this method will allow for the modeling of the exposure distribution within each strata. Additionally since the BSP method discussed in this chapter does not break the bonds, it has the added advantage of not needing to stratify the air pollution data based on season.

In the case of a single distribution generating the random effects, the BSP method will only be slightly less efficient than a full parametric Bayesian model. Prior to conducting an analysis, one cannot be certain if a single parametric distribution can suffice for a prior of the subject specific effects. The only loss in efficiency with the Dirichlet process prior is that across the MCMC iterations, there is a small probability that some subjects will be moved out of the cluster. In general, a majority of the subjects will belong to a single cluster across iterations. Therefore, the model proposed in this chapter is a viable method to alleviate the issues presented in Chapters 3 and 4.

5.7 Chapter Appendix

5.7.1 Full Conditional Posterior Distributions for Fixed Effect Parameters

Let $f(w|.)$ denote the probability density function for the random variable w conditional on all other variables in the model. Additionally for ease of notation, set:

$$g_{ijl} = \exp[\boldsymbol{\beta_1(Z_{ijl} - Z_{ij1})} + \xi_{ijl}\{b(\theta^*_{ijl}) - b(\theta_{ijl})\} - \xi_{ij1}\{b(\theta^*_{ij1}) - b(\theta_{ij1})\}].$$

Observe that θ_{ijl} is a function of γ_{0i} and $\boldsymbol{\gamma_1}$. Also, θ^*_{ijl} is a function of γ_{0i}, $\boldsymbol{\gamma_1}$, and β_{2i} (and in the continuous exposure case also a function of σ^2). In the scenario of a continuous exposure, ξ_{ijl} is a function of σ^2.

A prior distribution for $\boldsymbol{\beta_1}$, $\pi(\boldsymbol{\beta_1})$, is specified to be a normal distribution: i.e. $\boldsymbol{\beta}_1 \sim \mathrm{N}(\boldsymbol{\mu}_{\beta_1}, \Sigma_{\beta_1})$. The posterior for $\boldsymbol{\beta}_1$ is:

$$\begin{aligned} f(\boldsymbol{\beta}_1|.) &\propto L_c \times \pi(\boldsymbol{\beta_1}) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{m_i}\left(1 + \sum_{l=2}^{M+1} g_{ijl}\right)^{-1} \exp\left[-\frac{1}{2}(\boldsymbol{\beta_1} - \boldsymbol{\mu_{\beta_1}})'\Sigma^{-1}_{\beta_1}(\boldsymbol{\beta_1} - \boldsymbol{\mu_{\beta_1}})\right]. \end{aligned}$$

The prior distribution for $\boldsymbol{\gamma}_1$, $\pi(\boldsymbol{\gamma}_1)$, is specified to be a normal distribution: i.e. $\boldsymbol{\gamma}_1 \sim \mathrm{N}(\boldsymbol{\mu}_{\gamma_1}, \Sigma_{\gamma_1})$. The posterior for $\boldsymbol{\gamma_1}$ is:

$$
\begin{aligned}
f(\gamma_{1q}|.) &\propto L_c \times \pi(\boldsymbol{\gamma_1}) \\
&\propto \exp\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\xi_{ij1}\{\theta_{ij1}X_{ij1} - b(\theta^*_{ij1})\}\right] \exp\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{l=1}^{M+1}\xi_{ijl}\{\theta_{ijl}X_{ijl} - b(\theta^*_{ijl})\}\right] \\
&\quad \times \prod_{i=1}^{n}\prod_{j=1}^{m_i}\left(1 + \sum_{l=2}^{M+1} g_{ijl}\right)^{-1} \exp\left[-\frac{1}{2}(\boldsymbol{\gamma_1} - \boldsymbol{\mu_{\gamma_1}})'\Sigma^{-1}_{\gamma_1}(\boldsymbol{\gamma_1} - \boldsymbol{\mu_{\gamma_1}})\right].
\end{aligned}
$$

For the continuous exposure models only, there is also the parameter σ^2. Specify a prior distribution for σ^2, $\pi(\sigma^2)$, to be an inverse gamma: i.e. $\sigma^2 \sim \mathrm{IG}(a, b)$. The posterior for σ^2 is:

$$
\begin{aligned}
f(\sigma^2|.) &\propto L_c \times \pi(\sigma^2) \\
&\propto \prod_{i=1}^{n}\prod_{j=1}^{m_i}\exp\left[\frac{1}{\sigma^2}\left\{(\theta_{ij1} + \sigma^2\beta_{2i})X_{ij1} - \frac{1}{2}(\theta_{ij1} + \sigma^2\beta_{2i})^2\right\}\right] \\
&\quad \times \prod_{i=1}^{n}\prod_{j=1}^{m_i}\prod_{l=2}^{M+1}\exp\left[\frac{1}{\sigma^2}\left\{\theta_{ijl}X_{ijl} - \frac{\theta^2_{ijl}}{2}\right\}\right] \exp\left(-\frac{b}{\sigma^2})(\sigma^2\right)^{\left(\sum_{i=1}^{n} m_i\right)\left(-\frac{n+nM}{2}\right)-a-1}.
\end{aligned}
$$

5.7.2 Full Conditional Posterior Distributions in the Dirichlet Process

Let c_i be the cluster indicator for i ($c_i \in \{1, 2, \ldots, K\}$) and let n_j be the number of c_i's equal to j ($j = 1, 2, \ldots, K$).

Let $n = \sum_j n_j$ and $n_\#$ be the number of unique c_i's. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and c_U be the set of unique c_i's. Again, let $f(w|.)$ define the density of w given all other variables.

The full conditional posterior for the probability that subject i belongs to cluster l ($l =$

$1, 2, \ldots, K)$ is:

$$\begin{aligned} P(c_i = l|.) &\propto P(c_i = l|\pi) f(\theta_i|\mu_{c_i}, \sigma^2, c_i = l) \\ &\propto \pi_l \exp\left(\frac{-(\theta_i - \mu_l)^2}{2\sigma^2}\right). \end{aligned}$$

The full conditional posterior for the sticks are $(l = 1, 2, \ldots, K)$:

$$\begin{aligned} f(v_l|.) &\propto f(v_l|\alpha) P(c|v_l) \\ &\propto v_l^{1-1}(1 - v_l)^{\alpha-1} v_l^{n_l}(1 - v_l)^{\sum\limits_{j=l+1}^{K} n_j} \\ &\propto v_l^{n_l+1-1}(1 - v_l)^{\sum\limits_{j=l+1}^{K} n_j+\alpha-1}. \end{aligned}$$

since $P(c|v) = \pi_1^{n_1} \ldots \pi_K^{n_K}$. Thus $v_l|. \sim \text{Beta}(n_l + 1, \sum\limits_{j=l+1}^{K} n_j + \alpha)$.

$$\begin{aligned} f(\mu_l|.) &\propto f(\mu_l|\mu_0, \sigma_0^2) f(\theta|c, \mu_l, \sigma^2) \\ &\propto \exp\left[\frac{-(\mu_l - \mu_0)^2}{2\sigma_0^2}\right] \prod_{i:c_i=l} \exp\left[\frac{-(\theta_i - \mu_l)^2}{2\sigma^2}\right]. \end{aligned}$$

Thus $\mu_l|. \sim \text{Normal}\left((\frac{n_l}{\sigma^2} + \frac{1}{\sigma_0^2})^{-1} \left[\frac{\sum\limits_{i:c_i=1} \theta_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right], (\frac{n_l}{\sigma^2} + \frac{1}{\sigma_0^2})^{-1} \right)$.

$$\begin{aligned} f(\sigma_\theta^2|.) &\propto f(\sigma_\theta^2|a_0, b_0) f(\theta|\mu, \sigma_\theta^2) \\ &\propto (\sigma_\theta^2)^{-(a_0+1)} \exp\left(\frac{-b_0}{\sigma_\theta^2}\right) (\sigma_\theta^2)^{\frac{-n}{2}} \exp[-\frac{1}{\sigma_\theta^2} \sum_{i=1}^{n} (\theta_i - \mu_{c_i})^2. \end{aligned}$$

Therefore $\sigma_\theta^2|. \sim \text{IG}(a_0 + \frac{n}{2}, \frac{1}{2}\sum\limits_{i=1}^{n}(\theta_i - \mu_{c_i})^2 + b_0)$.

$$\begin{aligned} f(\sigma_0^2|.) &\propto f(\sigma_0^2|a_{00}, b_{00}) f(\mu|\sigma_0^2, \mu_0, c) \\ &\propto (\sigma_0^2)^{-(a_{00}+1)} \exp\left(-\frac{b_{00}}{\sigma_0^2}\right) (\sigma_0^2)^{-\frac{n_\#}{2}} \exp\left[-\frac{1}{2\sigma_0^2} \sum_{l \in c_U} (\mu_l - \mu_0)^2\right]. \end{aligned}$$

Thus $\sigma_0^2|. \sim \mathrm{IG}(a_{00} + \frac{n_\#}{2}, \frac{1}{2}\sum\limits_{l \in c_U}(\mu_l - \mu_0)^2 + b_{00})$.

$$\begin{aligned} f(\mu_0|.) &\propto f(\mu_0|\sigma_{00}^2, \mu_{00}) f(\mu|\sigma_0^2, \mu_0, c) \\ &\propto \exp\left[\frac{-(\mu_0 - \mu_{00})2}{2\sigma_{00}^2}\right] \exp\left[-\frac{1}{2\sigma_0^2}\sum_{l \in c_U}(\mu_l - \mu_0)^2\right]. \end{aligned}$$

Thus $\mu_0|. \sim \mathrm{Normal}\left((\frac{n_\#}{\sigma_0^2} + \frac{1}{\sigma_{00}^2})^{-1}\left(\frac{\sum\limits_{l \in c_U}\mu_l}{\sigma_0^2} + \frac{\mu_{00}}{\sigma_{00}^2}\right), (\frac{n_\#}{\sigma_0^2} + \frac{1}{\sigma_{00}^2})^{-1}\right)$.

$$\begin{aligned} f(\alpha|.) &\propto f(\alpha|a_\alpha, b_\alpha) f(v|\alpha) \\ &\propto \alpha^{a_\alpha - 1}\exp(-\alpha b_\alpha)\prod_{l=1}^{K}\alpha(1 - v_l)^{\alpha - 1}. \end{aligned}$$

Thus $\alpha|. \sim \mathrm{G}(a_\alpha + K, -\sum\limits_{l=1}^{K}\log(1 - v_l) + b_\alpha)$.

5.7.3 Sampling Method

Once the parameter values for the hierarchical components are drawn according to Section 5.7.2, a Metropolis-Hastings algorithm is implemented within the Gibbs sampling to draw values for β_1, γ_1, and in the continuous case σ^2. Let θ denote a general parameter.

The proposal density is set to be the same as the prior. That is, $g(\theta) = \pi(\theta)$, where $g(.)$ represents the proposal distribution and $\pi(.)$ is the prior distribution.

With a full conditional of the form $\pi(\theta|.) \propto \pi(\theta) * L_c(\theta|.)$, where $L_c(\theta|.)$ is the likelihood of the data evaluated at θ. The acceptance probability for jumping from θ to θ^* is:
$p^* = \frac{\pi(\theta^*|.)/g(\theta^*)}{\pi(\theta|.)/g(\theta)} = \frac{L_c(\theta^*)}{L_c(\theta)}$

For the subject specific effects γ_{0i} and β_{2i}, it is the same as for the shared parameters but the likelihood to accept a single subjects effects proposed value will only involve the likelihood contribution for that subject. Thus for a subject specific parameter θ_i, to jump from θ_i to the proposed θ_i^*, the acceptance probability is $p_i^* = \frac{L_{c_i}(\theta^*)}{L_{c_i}(\theta)}$.

To be specific, the Metropolis-Hastings algorithm to sample from the full conditional distribution of θ is as follows.

1. Draw a proposal draw θ^* from the proposal distribution $g(\theta^*)$.
2. Compute the acceptance probability for accepting the proposal as $p = \min\left(1, \frac{\pi(\theta^*|.)/g(\theta^*)}{\pi(\theta|.)/g(\theta)}\right)$.
3. Draw a $u \sim \text{uniform}(0, 1)$ random variable. Accept θ^* if $u < p$.
4. Repeat steps 1-3 to reach desired number of samples.

5.7.4 Hyper-parameter Values

The values for the fixed hyper-parameters in the simulation study were as follows. For the fixed parameters: $\mu_{\beta_1} = 0$, $\sigma^2_{\beta_1} = 5$, $\mu_{\gamma_1} = 0$, $\sigma^2_{\gamma_1} = 10$, $a = 5$, and $b = 4$. For the subject specific effects β_{2i} and γ_{0i} a single set of values were used for each: $a_0 = 5$, $b_0 = 4$, $a_{00} = 5$, $b_{00} = 4$, $\mu_{00} = 0$, $\sigma^2_{00} = 5$, $a_\alpha = 5$, and $b_\alpha = 4$.

Based on previous research, some prior information was used to choose hyper-parameter values. Such information stated that exposure parameter coefficients from air pollution studies are rarely large in magnitude, as they represent a change in the log odds ratio of an event for a daily exposure. As a result, the variance hyper-priors were slightly smaller than those in the simulation study. For the applied example, the following were used: $\mu_{\beta_1} = 0$, $\sigma^2_{\beta_1} = 3$, $\mu_{\gamma_1} = 0$, $\sigma^2_{\gamma_1} = 5$, $a = 5$, and $b = 4$. For the subject specific parameters β_{2i} and γ_{0i} a single set of values were used for each: $a_0 = 5$, $b_0 = 4$, $a_{00} = 5$, $b_{00} = 4$, $\mu_{00} = 0$, $\sigma^2_{00} = 3$, $a_\alpha = 5$, and $b_\alpha = 4$.

5.7.5 Bivariate DP

The stratum specific effects were modeled jointly, and identical results were obtained. The bivariate model is as follows:

$$\begin{pmatrix} \gamma_{0i} \\ \beta_{2i} \end{pmatrix} | \boldsymbol{\mu}_i, \Sigma \sim \mathrm{N}(\boldsymbol{\mu}_i, \Sigma).$$

$$\Sigma \sim \text{Inv-Wishart}(\Psi, v).$$

$$\boldsymbol{\mu}_i | G \sim G.$$

$$G \sim \mathrm{DP}(\alpha, G_0).$$

$$G_0 \equiv \mathrm{N}(\boldsymbol{\mu}_0, \Sigma_0).$$

$$\Sigma_0 \sim \text{Inv-Wishart}(\Lambda, w).$$

$$\boldsymbol{\mu}_0 \sim \mathrm{N}(\boldsymbol{\mu}_{00}, \Sigma_{00}).$$

5.7.6 Simulating $1 : M$ Data

Again, suppress the j subscript. To generate m_i many events for subject i, repeat the following m_i many times. For 1:2 matching let $\boldsymbol{Z_i} = \boldsymbol{(Z_{i1}, Z_{i2}, \ldots, Z_{i,M+1}})$ where $M = 2$.

Compute the probability the first covariate $\boldsymbol{Z_{i1}}$ generates an event, $P(D_{i1} = 1 | D_{i1} + D_{i2} + D_{i3} = 1, \boldsymbol{Z_{i.}, S_i})$:

$$\begin{aligned}
&= \frac{p(D_{i1} = 1, D_{i2} = 0, D_{i3} = 0|\boldsymbol{Z_{i.}, S_i})}{p(D_{i1} + D_{i2} + D_{i3} = 1|\boldsymbol{Z_{i.}, S_i})} \\
&= \frac{p(D_{i1} = 1|\boldsymbol{Z_{i1}, S_i}) * p(D_{i2} = 0|\boldsymbol{Z_{i2}, S_i}) * p(D_{i3} = 0|\boldsymbol{Z_{i3}, S_i})}{\sum\limits_{k=1}^{3} P(D_{ik} = 0|\boldsymbol{Z_{ik}, S_i}) \prod\limits_{h \neq k} P(D_{ih} = 0|\boldsymbol{Z_{ih}, S_i})}.
\end{aligned}$$

All the pieces needed are derived in the chapter. Divide numerator and denominator by the numerator quantity to simplify. Once D_{i1} is drawn and if it is equal to 1, then stop. All other D_{ik} such that $k \neq 1$ will be set to 0. $\boldsymbol{Z_{i1}}$ will be the covariate value for the case, and all other $\boldsymbol{Z_{ik}}$, $k \neq 1$ will be for the controls. If $D_{i1} = 0$, then the process continues to draw D_{i2}. The probability that $D_{i2} = 1$ given that $D_{i1} = 0$ is as follows:

$$\begin{aligned}
P(D_{i2} = 1|\sum_{k=1}^{3} D_{ik} = 1, D_{i1} = 0, \boldsymbol{Z_{i.}, S_i}) &= \frac{P(D_{i1} = 0, D_{i2} = 1, D_{i3} = 0|, D_{i1} = 0, \boldsymbol{Z_{i.}, S_i})}{P(D_{i1} + D_{i2} + D_{i3} = 1|, D_{i1} = 0, \boldsymbol{Z_i, S_i})} \\
&= \frac{P(D_{i2} = 0|\boldsymbol{Z_{i2}, S_i}) P(D_{i3} = 0|\boldsymbol{Z_{i3}, S_i})}{\sum\limits_{k=1}^{3} P(D_{ik} = 0|\boldsymbol{Z_{ik}, S_i}) \prod\limits_{\substack{h \neq k \\ h \neq 1}} P(D_{ih} = 0|\boldsymbol{Z_{ih}, S_i})}
\end{aligned}$$

Draw a value for D_{i2} based on the probability above. If $D_{i2} = 1$, then stop and set $D_{i3} = D_{i1} = 0$. $\boldsymbol{Z_{i2}}$ will be the covariate value for the case, and all other $\boldsymbol{Z_{ik}}$, $k \neq 2$ will be for the controls.. If $D_{i3} = 0$, then $P(D_{i3}) = 1$ using the same idea as above. Thus $D_{i3} = 1$ and $D_{i1} = D_{i2} = 0$. $\boldsymbol{Z_{i3}}$ will be the covariate value for the case, and all other $\boldsymbol{Z_{ik}}$, $k \neq 3$ will be for the controls.

Once D_{il} and $\boldsymbol{Z_{il}}$ are simulated, generate the exposures for each $l = 1, 2, 3$ accordingly. The process for simulating $1 : M$ matching with $M > 2$ is similar.

5.7.7 Trace Plots

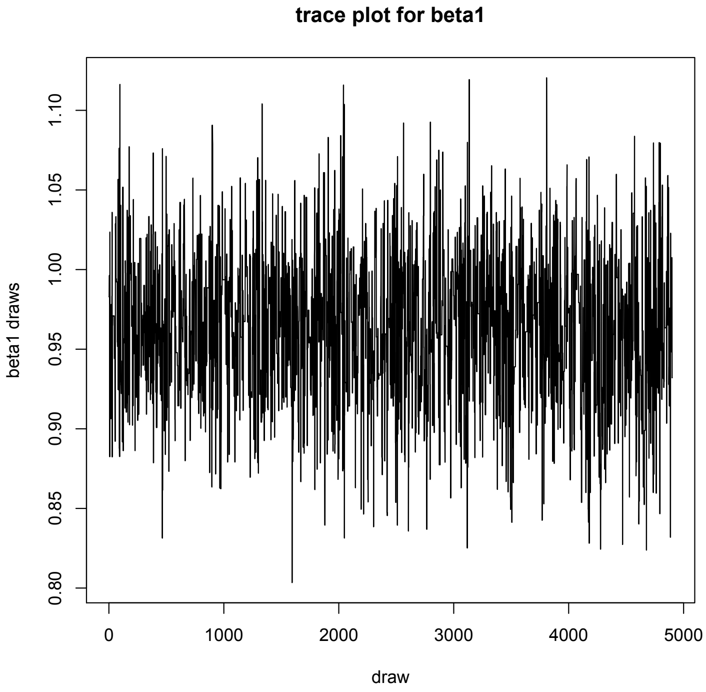

Figure 5.9: Representative trace plot for β_1 for the continuous exposure simulation study.

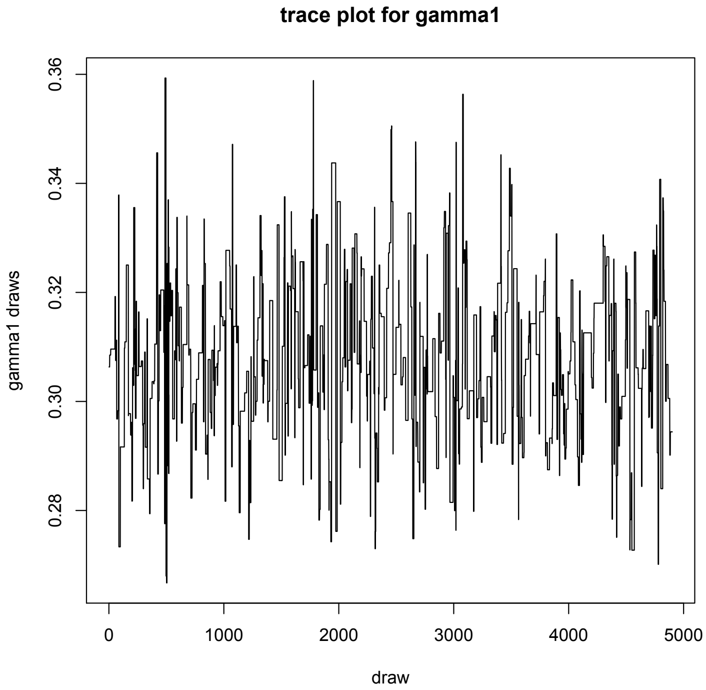

Figure 5.10: Representative trace plot for γ_1 for the continuous exposure simulation study.

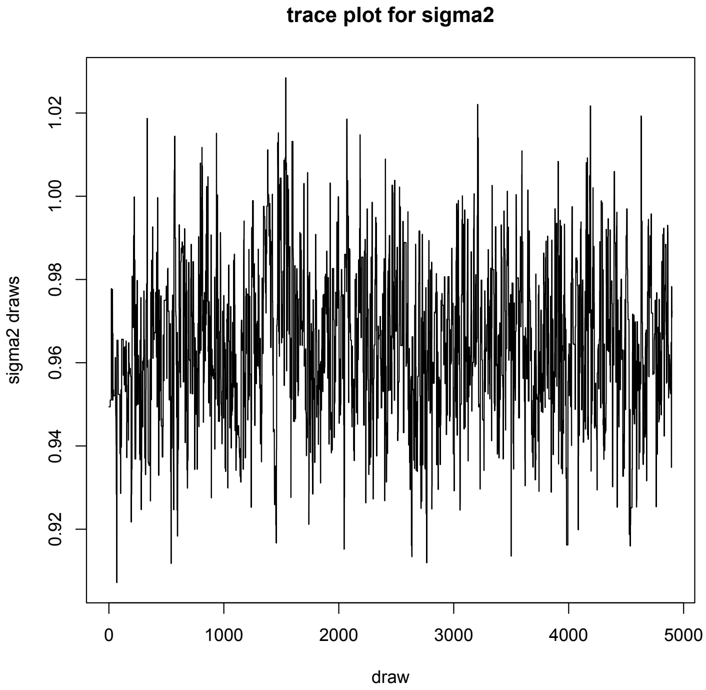

Figure 5.11: Representative trace plot for σ^2 for the continuous exposure simulation study.

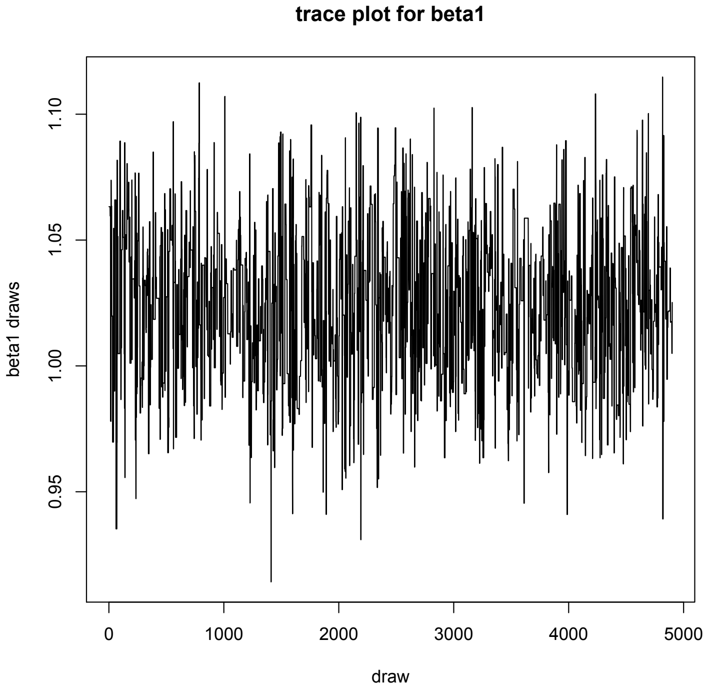

Figure 5.12: Representative trace plot for β_1 for the binary exposure simulation study.

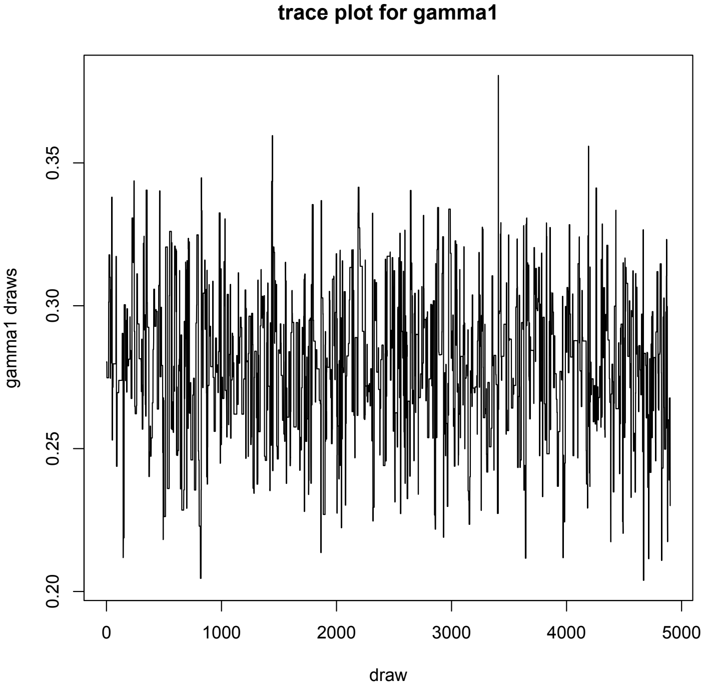

Figure 5.13: Representative trace plot for γ_1 for the binary exposure simulation study.

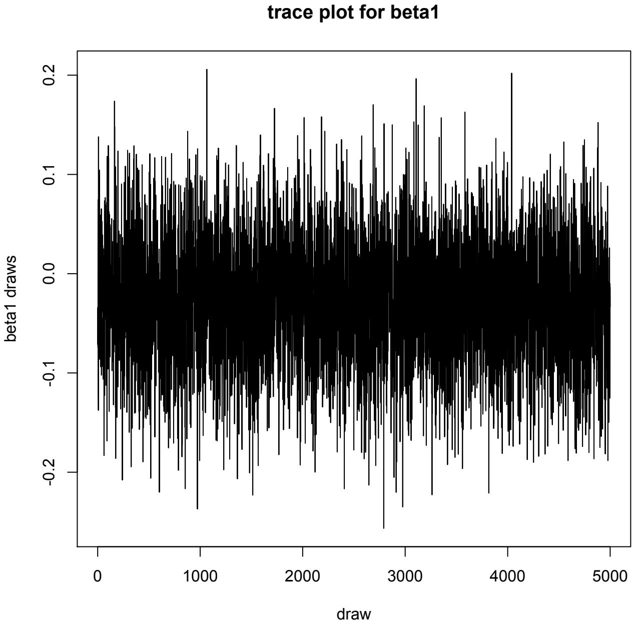

Figure 5.14: Trace plot for β_1 for the applied illustration using the dataset comprising of white non-Hispanic subjects.

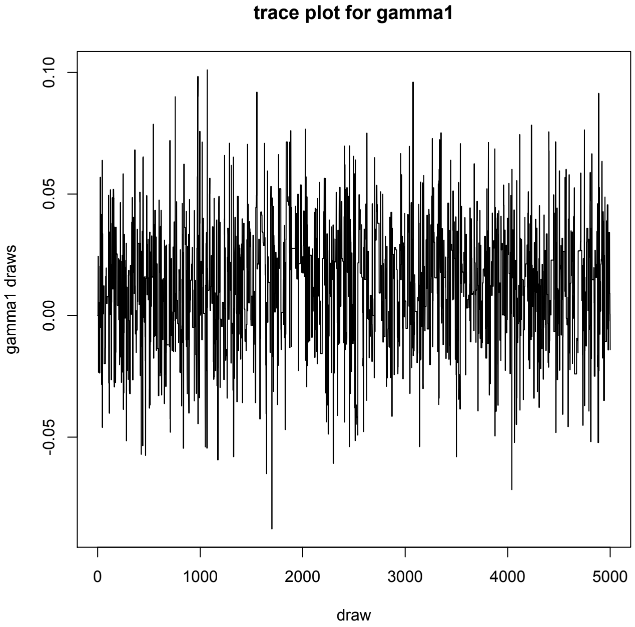

Figure 5.15: Trace plot for γ_1 for the applied illustration using the dataset comprising of white non-Hispanic subjects.

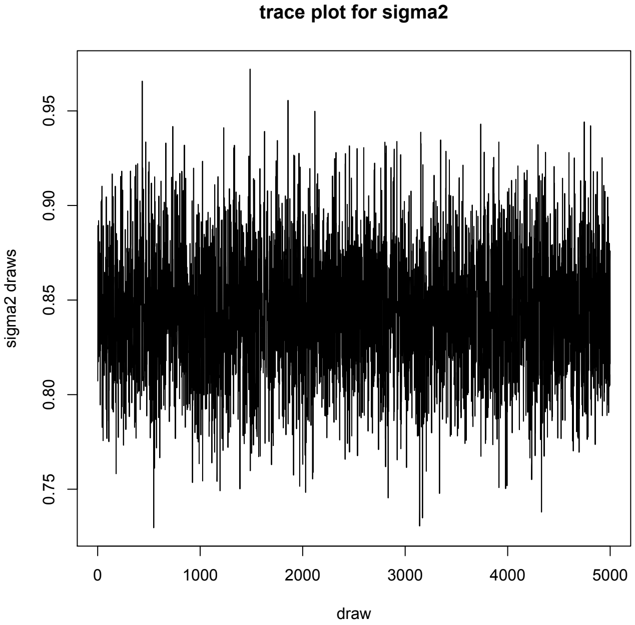

Figure 5.16: Trace plot for σ^2 for the applied illustration using the dataset comprising of white non-Hispanic subjects.

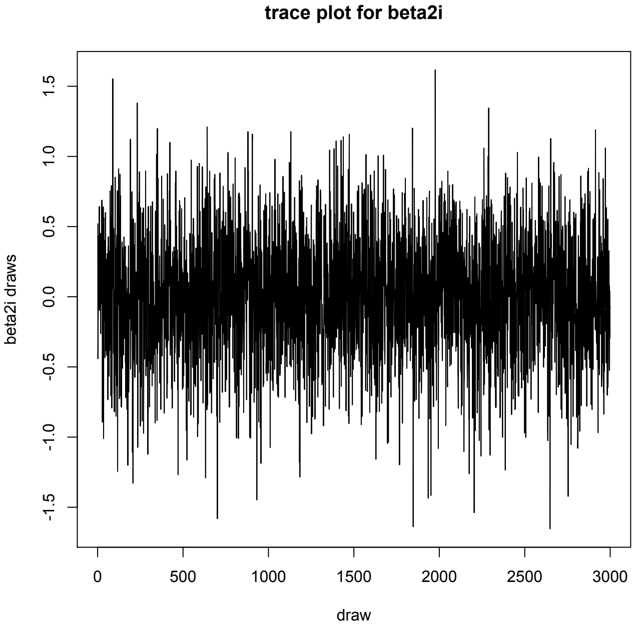

Figure 5.17: Representative trace plot for a randomly selected subject's β_{2i} for the applied illustration using the dataset comprising of white non-Hispanic subjects.

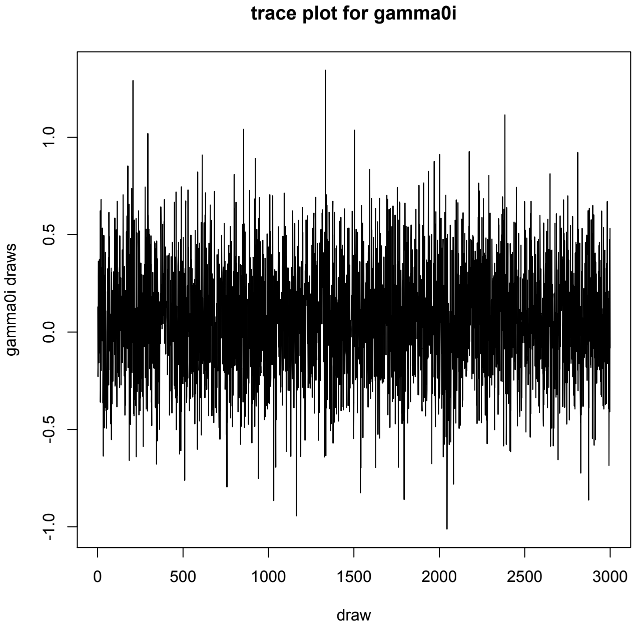

Figure 5.18: Representative trace plot for a randomly selected subject's γ_{0i} for the applied illustration using the dataset comprising of white non-Hispanic subjects.

5.7.8 Dendograms

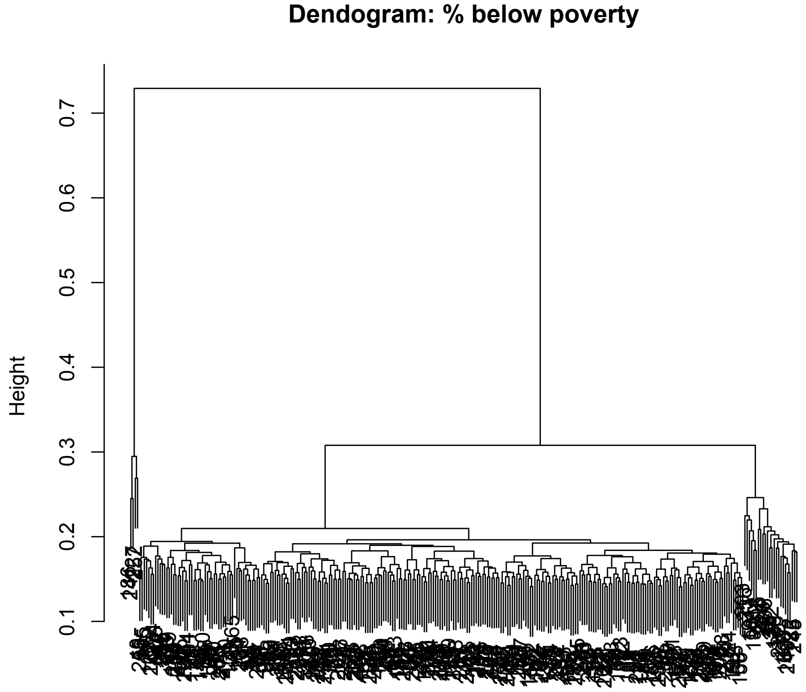

Figure 5.19: Dendogram for the applied illustration using the dataset comprising of subjects living in zipcodes where the percent living below the poverty line is above the median for Orange county.

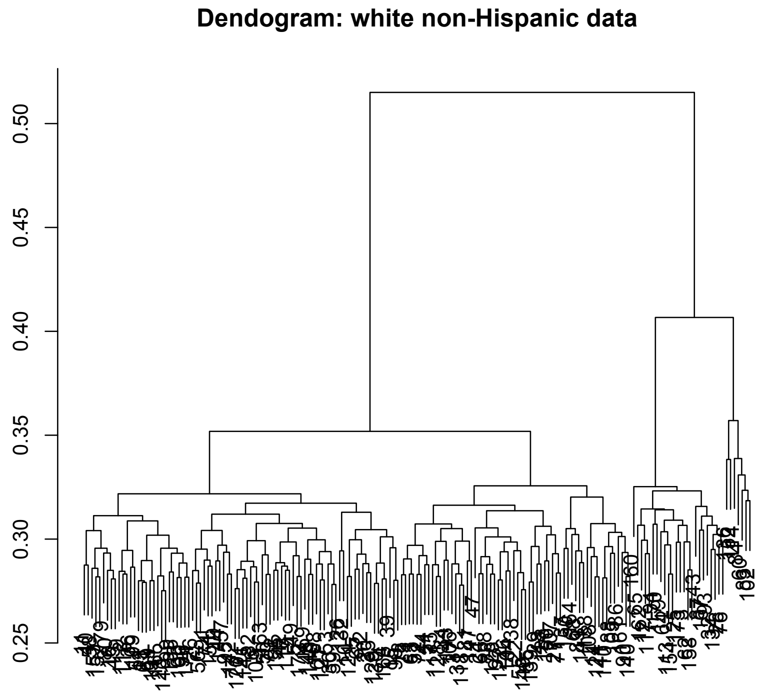

Figure 5.20: Dendogram for the applied illustration using the dataset comprising of white non-Hispanic subjects.

Chapter 6

Future Work

The work from this dissertation can be summarized as follow. Through an extensive simulation study, it was found that the appropriate method to obtain unbiased parameter estimates while accounting for the correlation within the dataset in a matched case-control study with numerous case-control pairs within each matched is the discrete method. When the matched set sizes vary, and effect modification exists among the matched sets, there is no single true estimand. It was shown and discussed that in the scenario of varying matched set sizes with effect modification, different methods used to obtain a marginal parameter estimate have substantially different estimands. The estimands differ in the weights assigned when computing a marginal parameter estimates. Prior to conducting the analysis, researchers need to determine if the matched sets should be weighted equally, or if the individual matched pairs within the sets should be weighted equally. The Bayesian semi-parametric (BSP) model proposed does not need the estimand to be specified prior to conducting the analysis. This method provides estimates of the stratum specific effects across all strata (matched sets).

Future work will expand on the BSP method proposed in this dissertation. With regards to the implementation of the method, to improve convergence time of the MCMC chains the Metropolis-Hastings algorithm can be replaced with a Hamiltonian Markov chain. With regards to the methodology, two extensions could be implemented. The first is to construct a clustered likelihood to make comparisons with the discrete method discussed in Chapter 3 and Chapter 4. That is, we could construct a likelihood of the form

$$L = \prod_{i=1}^{n} L_i,$$

where

$$L_i \propto p\left(\boldsymbol{D_{i\cdot\cdot}}|\boldsymbol{S_i}, \boldsymbol{Z_{i\cdot\cdot}}, \sum_{j=1}^{m_i}\sum_{l=1}^{M+1} D_{ijl} = t\right) p(\boldsymbol{X_{i\cdot\cdot}}|\boldsymbol{S_i}, \boldsymbol{Z_{i\cdot\cdot}}, \boldsymbol{D_{i\cdot\cdot}})$$

and where

$$p\left(\boldsymbol{D_{i\cdot\cdot}}|\boldsymbol{S_i}, \boldsymbol{Z_{i\cdot\cdot}}, \sum_{j=1}^{m_i}\sum_{l=1}^{M+1} D_{ijl} = t\right) = \frac{\prod\limits_{j=1}^{m_i}\left[p(D_{ij1} = 1)\prod\limits_{l=2}^{M+1} D_{ijl} = 0)\right]}{\sum\limits_{D_i^*:\sum\limits_j\sum\limits_l D^*_{ijl} = m_i}\left[\prod\limits_j\prod\limits_l p(D_{ijl} = D^*_{ijl})\right]}.$$

The second extension will be a primary focus of the future work to be implemented. This extension will construct a likelihood that is not conditional on the number of events known to happen in each matched pair. That is to say the likelihood will be constructed using the quantities $p(\boldsymbol{D}_{ij.}, \boldsymbol{X}_{ij.}, \boldsymbol{\delta}_{ij.}|\boldsymbol{Z}_{ij.}, \boldsymbol{S}_i)$ as opposed to $p(\boldsymbol{D}_{ij.}, \boldsymbol{X}_{ij.}, \boldsymbol{\delta}_{ij.}|\boldsymbol{Z}_{ij.}, \boldsymbol{S}_i, \sum_{l=1}^{M+1} D_{ijl} = 1)$. This will result in a likelihood that will not have the subject specific intercepts, β_{0i}, of the prospective probability of an events being factored out. These random effects can be modeled, leading one to be able to make inference about the individual subjects baseline risk of an event, and to also obtain prediction of the risk of an event given a specific level of the exposure.

Bibliography

P. K. Andersen and R. D. Gill. Coxs regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3): 349–355, 1955.

S. G. Baker. The multinomial-Poisson transformation. *Statistician*, 43(4):495–504, 1994.

T. F. Bateson and J. Schwarts. Control for seasonal variation and time trend in case crossover studies of acute effects of environmental exposures. *Epidemiology*, 10(5):539–544, 1999.

T. F. Bateson and J. Schwarts. Selection bias and confounding in case-crossover analyses of environmental time series. *Epidemiology*, 12(6):654–661, 2001.

D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

N. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, 43(1):1121–1134, 1975.

N. Breslow and N. Day. *Statistical Methods in Cancer Research: Volume 1.* International Agency for Research on Cancer, 1980.

N. E. Breslow. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association*, 91(433):14–28, 1996.

N. E. Breslow, N. E. Day, K. T. Halvorsen, R. L. Prentice, and C. Sabai. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, 108(4):299–307, 1978a.

N. E. Breslow, N. E. Day, K. T. Halvorsen, R. L. Prentice, and C. Sabai. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, 108(4):299–307, 1978b.

A. C. Broders. Squamous-cell epithelioma of the lip. A study of five hundred and thirty-seven cases. *Journal of the American Medical Association*, 74(10):656–664, 1920.

J. Chang, R. Delfino, D. Gillen, T. Tjoa, B. Nickerson, and D. Cooper. Repeated respiratory hospital encounters among children with asthma and residential proximity to traffic. *Occupational and Environmental Medicine*, 66(2):90–98, 2009.

W. G. Cochran. Some method for strengthening the common χ^2 tests. *Biometrics*, 10: 417–451, 1954.

J. Cornfield. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11(6): 1269–1275, 1951.

J. Cornfield, T. Gordon, and W. W. Smith. Quantal response curves for experimentally uncontrolled variables. *Bulletin of the international Statistical Institute*, 1961:97–115, 1961.

D. Cox. Regression models and life tables. *Journal of the Royal Statistical Society*, 34(2): 187–220, 1972.

D. Cox and D. Oakes. *Analysis of survival times data.* Chapman and Hall, 1984.

D. R. Cox. Some procedures connected with the logistic qualitative response curve. *Research Papers in Statistics: Festschrift for J. Neyman*, pages 55–71, 1966.

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

N. E. Day and D. F. Kerridge. A general maximum likelihood discriminant. *Biometrics*, 23 (2):313–323, 1967.

R. Delfino, J. Wu, T. Tjoa, S. Gulesserian, B. Nickerson, and D. Gillen. Asthma morbidity and ambient air pollution: effect modification by residential traffic-related air pollution. *Epidemiology*, 25(1):48–57, 2014.

D. DeLong, G. Guirguis, and C. Ying. Efficient computation of subset selection probabilities with application to Cox regression. *Biometrika*, 81(3):607–611, 1994.

P. J. Diggle. A point process modeling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *Journal of the American Statistical Association*, 153(3):349–362, 1990.

P. J. Diggle, K. Y. Liang, and S. L. Zeger. *Analysis of longitudinal data.* Oxford Science Publications, 1994.

P. J. Diggle, S. E. Morris, and J. C. Wakefield. Point-source modeling used matched case-control data. *Biostatistics*, 1(1):89–105, 2000.

B. Efron. The efficiency of Coxs likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565, 1977.

M. D. Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994.

M. D. Escobar and M. West. Bayesian density estimations and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

T. S. Ferguson. *Recent advances in statistics.* Academic Press, 1983.

T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis.* John Wiley, 1991.

D. Follman, M. Proschan, and E. Leifer. Multiple outputation: Inference for complex clustered data by averaging analyses from independent data. *Biometrics*, 59(2):420–429, 2003.

K. Fung, S. Khan, D. Krewski, and T. Ramsay. A comparison of methods for the analysis of recurrent health outcome data with environmental covariates. *Statistics in Medicine*, 26 (3):532–545, 2007.

M. Gail, J. Lubin, and L. Rubinstein. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika*, 68(3):703–707, 1981.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

M. Ghosh and M. H. Chen. Bayesian inference for matched case-control studies. *Sankhyā*, 12(2):107–127, 2002.

P. J. Heagerty. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3):688–698, 1999.

I. Hertz-Picciotto and B. Rockhill. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, 53(3):1121–1134, 1997.

E. Hoffman, P. Sen, and C. Weinberg. Within-cluster resampling. *Biometrika*, 88(4):1121–1134, 2001.

P. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233, 1967.

H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

J. Jaakola. Case-crossover design in air pollution epidemiology. *European Respiratory Journal*, 21(40):81–85, 2003.

H. Janes, L. Sheppard, and T. Lumley. Overlap bias in the case-crossover design, with application to air pollution exposures. *Statistics in Medicine*, 24(2):285–300, 2005a.

H. Janes, L. Sheppard, and T. Lumley. Case crossover analyses of air pollution exposure data: Referent selection strategies and their implications for bias. *Epidemiology*, 16(6): 717–726, 2005b.

J. Kalbfleisch and R. Prentice. Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60(2):1121–1134, 1976.

J. Kalbfleisch and R. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, 1980.

J. D. Kalbfleisch and R. L. Prentice. Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2):267–278, 1973.

I. Kim, H. K. Cheong, and H. Kim. Semiparametric regression models for detecting effect modification in matched case-crossover studies. *Statistics in Medicine*, 30(15):1837–1851, 2011.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38 (4):963–974, 1982.

J. E. Lane-Claypon. A further report on cancer on the breast. *Reports on Public Health and Medical Subjects*, 32, 1926.

D. Levy and T. Lumley. Bias in the case-crossover design: Implications for studies of air pollution. *Environmetrics*, 11(6):689–704, 2000.

N. M. Liang and S. L. Zeger. Longitudinal data analysis using generalizes linear models. *Biometrika*, 73(1):13–22, 1986.

A. M. Lileinfield and D. E. Lilienfield. A century of case-control studies: progress? *Journal of Chronic Diseases*, 32(1):5–13, 1979.

D. Lin and L. Wei. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989.

A. Y. Lo. On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

H. L. Lombard and C. R. Doering. Cancer studies in Massachusetts. 2. Habits, characteristics and environment of individuals with and without cancer. *New England Journal of Medicine*, 198:481–487, 1928.

X. Luo and G. Sorock. Analysis of recurrent event data under the case-crossover design with applications to elderly falls. *Statistics in Medicine*, 27(15):2890–2901, 2008.

S. MacEachern and P. Muller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

M. Maclure. The case-crossover design: A method for studying transient effect on the risk of acute events. *American Journal of Epidemiology*, 133(2):144–153, 1991.

N. Mantel. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 58:690–700, 1963.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

R. J. Marshall. Bayesian analysis of case-control studies. *Statistics in Medicine*, 7(12): 1223–1230, 1988.

Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

B. Mukherjee, L. Zhang, M. Ghosh, and S. Sinha. Semiparametric Bayesian analysis of case-control data under conditional gene-environment independence. *Biometrics*, 63(3): 834–844, 2007.

P. Müller and K. Roeder. A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, 84(3):523–537, 1997.

W. Navidi. Bidirectional case-crossover designs for exposures with time trends. *Biometrics*, 54(2):1107–1111, 1998.

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

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

J. Neuhaus and J. Kalbfleisch. Between and within cluster covariate effects in the analysis of clustered data. *International Biometric Society*, 54(2):638–645, 1998.

C. I. Neutel, S. Perry, and C. Maxwell. Medication use and risk of falls. *Pharmacoepidemiology and Drug Safety*, 11(2):97–104, 2002.

M. Nurminen and P. Mutanen. Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics*, 14(1):67–77, 1987.

K. L. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

A. Phillips and P. W. Holland. Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics*, 43:311–323, 1987.

M. C. Pike and R. H. Morrow. Statistical analysis of patient-control studies in epidemiology: Factor under investigation an all-or-non variable. *British Journal of Preventive amd Social Medicine*, 24(1):42–44, 1970.

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

J. Robins, N. Breslow, and S. Greenland. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, 42(2):311–323, 1986.

D. B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

G. A. Satten and R. J. Carroll. Conditional and unconditional categorical regression models with missing covariates. *Biometrics*, 56:384–388, 2000.

S. Seaman and S. Richardson. Bayesian analysis of case-control studies with categorical covariates. *Biometrika*, 88(4):1073–1088, 2001.

S. Seaman and S. Richardson. Equivalence of prospective and retrospective models in the bayesian analysis of case-control studies. *Biometrika*, 91(1):15–25, 2004.

D. G. Seigel and S. W. Greenhouse. Multiple relative risk function in case-control studies. *American Journal of Epidemiology*, 97(5):324–331, 1973.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

S. Sinha, B. Mukherjee, and M. Ghosh. Bayesian semiparametric modeling for matched case-control studies with multiple disease states. *Biometrics*, 60(1):41–49, 2004.

S. Sinha, B. Mukherjee, M. Ghosh, and R. Carroll. Semiparametric Bayesian modeling of matched case-control studies with with missing exposure. *Journal of the American Statistical Association*, 100(470):591–601, 2005.

L. Trasande and G. D. Thurston. The role of air pollution in asthma and other pediatric morbidities. *Journal of Allergy and Clinical Immunology*, 115(4):689–699, 2005.

S. Vines and C. Farrington. Within-subject exposure dependency in case-crossover studies. *Statistics in Medicine*, 20(20):3039–3049, 2001.

J. M. Williamson, M. N. Hocine, and F. C. P. Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59(1):36–42, 2003.

S. L. Zeger and N. M. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, 1986.

M. Zelen and R. Parker. Case-control studies and Bayesian inference. *Statistics in Medicine*, 5(3):261–269, 1986.

Y. Zhang, R. Woods, C. Chaisson, T. Neogi, T. McAlindon, and D. Hunter. Alcohol consumption as a trigger of recurrent gout attacks. *The American Journal of Medicine*, 119 (9):13–18, 2006.