國立交通大學

統計學研究所

碩 士 論 文

家族性病例對照資料之統計分析

Statistical Analysis for Familial Case-Control Data

研 究 生：蘇筱嵐

指導教授：王維菁　博士

中 華 民 國 九 十 九 年 六 月

家族性病例對照資料之統計分析

Statistical Analysis for Familial Case-Control Data

研 究 生：蘇筱嵐　　Student：Hsiao-Lan Su

指導教授：王維菁　博士　　Advisor：Dr. Weijing Wang

國 立 交 通 大 學

統 計 學 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Statistics

College of Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年六月

家族性病例對照資料之統計分析

學生：蘇筱嵐　　指導教授：王維菁 博士

國立交通大學統計學研究所

摘　　要

家族性病例對照資料研究方法近年來常使用於探討疾病與致病因子之關係。本論文回顧了分析家族資料的統計文獻方法：針對得病與否的反應變數，考慮了邏輯斯迴歸模型; 針對得病時間的反應變數，考慮了 Cox 等比風險模型。我們討論如何將建立在個別性資料上之研究方法推廣至家族性資料，並探討如何將適用於前瞻性資料的方法修改為分析病例對照資料所需做的假設與調整。此外，我們也透過模擬實驗來驗證推論過程中所需要之條件與比較參數估計之表現。

關鍵字：家族性病例對照資料研究；前瞻性研究；邏輯斯迴歸模型；Cox PH 模型；Clayton 模型

Statistical Analysis for Familial Case-Control Data

Student: Hsiao-Lan Su
Advisor: Dr. Weijing Wang

Institute of Statistics
National Chiao Tung University
Hsinchu, Taiwan

Abstract

Familial case-control data are frequently used to study the relationship between disease and risk factors. In the thesis, we review literature for analyzing familial data. The logistic model is applied to model the probability of disease incidence. The Cox proportional hazards model is applied to model the age at onset of the disease. For each model, we discuss how to extend the method and model developed for individual data to familial data. In addition, we discuss the criteria and modification from prospective data to case-control data. We also propose simulation algorithms for generating case-control data and then, based on simulated data, examine parameter estimates and crucial properties of the inference procedure.

Keywords: Familial case-control study; Prospective study; Logistic regression; Cox PH model; Clayton model

誌　　謝

碩士生涯的終點，也代表了十八年學生旅程的結束。這兩年能有所成果，首先必須感謝我的指導老師—王維菁教授，我在理論的推導與邏輯思考上，一直不是很敏感，反應也很慢，是老師像母親般不厭其煩的引導我從大方向思考，也讓我學會了該如何去統整概念和如何論述的技巧，更幫助我撐起整篇論文的架構，老師給予的訓練在我往後的人生中將會受用無窮。

再來要感謝所有交大統研 97 與悶騷的研究室同學們，因為有你們的一同努力與分享，學習之路不至於乏味而無援。還有高中與大學的好友們，你們的支持與鼓勵是我向前邁進的動力。

也必須感謝交大統計所所有的老師與郭姐，給予了我們良好的學習環境，讓我在碩士這兩年學到了許多知識。還有我的口試委員徐南蓉教授、黃信誠教授與洪慧念教授，謝謝你們的協助與討論，使論文更加完善。

最後，我要感謝我的母親，我知道我不是個貼心的好女兒，這幾年來讓您辛苦了，您無條件的支持使我求學過程中完全無後顧之憂。最後，我想將這篇論文獻給我們永遠懷念的父親，我會盡力達成您的期望。

感謝一路上曾幫助過我的貴人，希望大家未來的人生都順利而快樂。

蘇筱嵐　　謹誌于
國立交通大學統計學研究所
中華民國九十九年六月

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Motivation

Scientists are interested in studying the roles of genetic and environmental factors on the development of a disease. Besides the information about whether the disease is present or not, age-at-onset has been viewed as a useful quantitative trait for some commonly-seen complex diseases. For example, early onset of breast cancer has been viewed as an important hallmark for genetic predisposition. Figure 1 highlights the scientific background which motivates this thesis. For a quantitative trait, statisticians can perform regression analysis which the effects of the explanatory variables on the response.

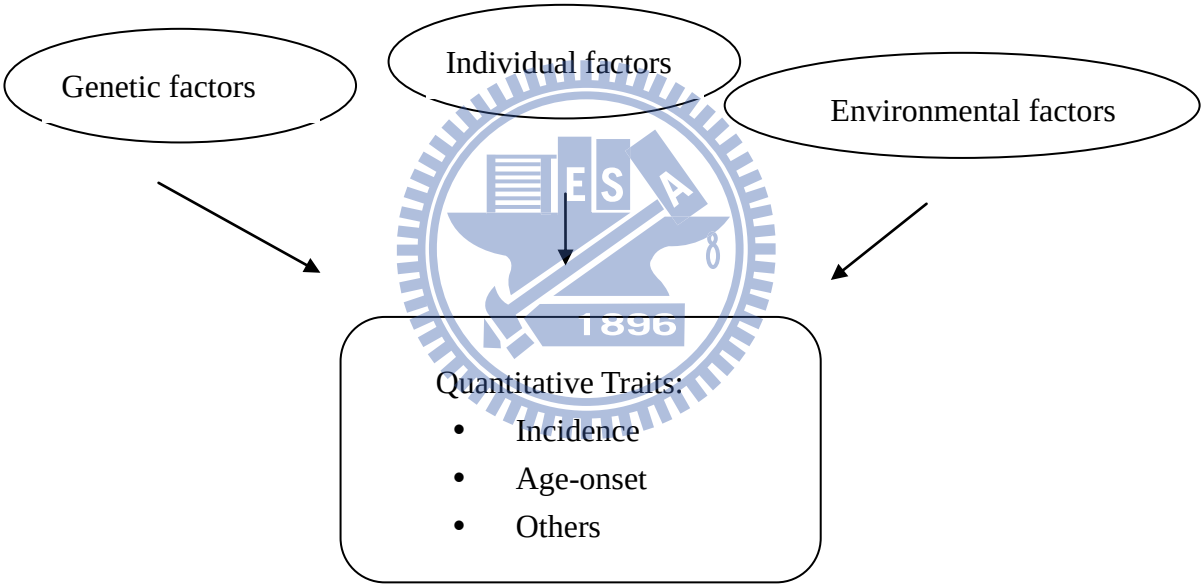

Figure 1: Scientific Background

We focus on two quantitative traits, namely disease incidence and age-at-onset. Disease incidence can be coded as a binary variable. Age-onset variables are continuous but may be censored due to termination of the study or loss to follow-up. Genetic, environmental and individual factors are treated as observed covariates. Their influences on the chosen response variable are of major interest. If disease incidence is the response variable, logistic regression models can be adopted. If age-at-onset is studied, failure-time regression models such as Cox proportional hazards models can be applied. When genetic information is not directly

measured, familial data can be used to detect its influence. Familial aggregation often indicates that genetic or shared environmental factors play some role in the development of the disease.

From the aspect of data design, the case-control sampling study is often applied to gather the information of rare diseases. It has the advantage that sufficient number of cases can be obtained and hence is cheaper and more convenient in comparison with a prospective study. In recent years, familial case-control designs have become a popular choice in genetic epidemiology. However statistical inference based on familial case-control data deserves careful investigation since the underlying probability structure is not straightforward.

1.2 Outline

The purpose of the thesis is to review related literature under a unified framework and examine some theoretical statements via simulations. In Chapter 2, we provide some background for different types of case-control designs. In Chapter 3, we review literature on logistic regression for familial prospective studies and case-control designs. Chapter 4 contains some simulation results which are conducted to verify crucial probability statements for logistic regression analysis. In Chapter 5, we review literature on familial age-onset data based on case-control designs. Chapter 6 contains simulation studies for checking the assumptions that are required in analysis of age-onset data from case-control family studies. Concluding remarks are contained in Chapter 7.

Chapter 2 An Overview of Case-control Designs

Case-control designs are preferable because they are cheaper and more convenient. In this chapter, we focus on two common case-control designs: namely the conventional and familial designs.

2.1 Conventional Case-control Designs

Conventional case-control designs begin by recruiting a group of individuals with a specific disease as “cases” and the other group of non-diseased individuals as “controls”. Cases and controls are compared based on risk factors including familial history of the disease. Here positive familial history is defined as presence of the disease in one or more first-degree relatives. However potential bias may arise due to incorrect information of recall. Furthermore individuals may differ in their family sizes so that positive family history is more likely to occur in a larger family. The family sizes differ in cases and controls can lead to false results. Liang (2000) discussed potential biases for conventional case-control designs in details.

2.2 Familial Case-Control Designs

Familial data obtained from case-control designs are frequently used to detect disease aggregation in families. This design begins by identifying a sample of diseased cases and an independent sample of disease-free controls, and for each individual, hereafter called a “proband”, determines his/her covariates, the family structure, and the disease status and covariates of relatives in the family. The disease status of relatives is treated as one part of the responses in the model.

A major difference between the two designs lies in the sampling unit. The sampling unit in familial case-control designs is a pre-defined set of family members. Compared to the conventional design, familial case-control designs provide direct evaluations of the relatives and can avoid misclassification of family history. It is also useful for genetic counseling. However familial case-control designs are more expensive.

2.3 The Issue of Matching in Case-Control Designs

In case-control designs, there are some confounding variables that may affect the evaluation of the association between disease incidence and risk factors. So, sometimes we must consider the necessity to match these confounding variables in the design stage. The purpose of matching is to let the units between cases and controls have more comparability.

The matching method includes frequency matching and individual matching. In conventional case-control design, if individual matching is part of the design, the conditional logistic regression method mentioned in Breslow and Day (1980) may be adopted. When in a familial case-control design, we note that the sampling units are families. So the matching between case probands and control probands doesn't guarantee the matching between case relatives and control relatives. Thus the matching procedure in such studies should be subject to some modification. First, the matching in design stage must be run under the condition that the confounding variables are familial, for example: races. Second, correlations among relatives have to be dealt with. Liang (1987) proposed a method for analyzing the matched designs which accounts for the within-family correlation. For age-onset responses, Li et al. (1998) also discussed situations under familial structure and matched procedure.

Finally, Sturmer and Brenner (2000) discussed the issue of the balance between power gain and extra costs for doing the matching.

Chapter 3 Logistic Regression on Different Designs

Logistic regression models are commonly adopted for modeling the relationship between a binary response and covariates. We first discuss statistical inference based on prospective studies which can be easily understood. Then we discuss how to construct the likelihood function if the sample is obtained from a case-control design. Finally we will review the literature on logistic regression analysis for familial case-control studies.

Denote Y as a binary indicator for disease status. Specifically $Y=1$ represents that the individual is diseased while $Y=0$ indicates that the individual is free of the disease. Denote Z as a $p\times 1$ vector of covariates. Consider the following logistic regression model:

$$\Pr(Y=1\mid Z)=\frac{\exp(\alpha+\beta^T Z)}{1+\exp(\alpha+\beta^T Z)}. \tag{3.1}$$

Let $\{(Y_i,Z_i)\ (i=1,...,n)\}$ denote the observed sample. If the data are collected from a prospective design, the likelihood function can be written as

$$\prod_{i=1}^{n}\left\{\frac{\exp(\alpha+\beta^T Z_i)}{1+\exp(\alpha+\beta^T Z_i)}\right\}^{Y_i}\left\{\frac{1}{1+\exp(\alpha+\beta^T Z_i)}\right\}^{1-Y_i}. \tag{3.2}$$

A case-control study, by contrast, identify a sample of diseased cases: $Y=1$ and another independent sample of non-diseased controls: $Y=0$. The covariate Z is measured afterwards. Notice that the distribution of data from a case-control study is based on $\Pr(Z\mid Y)$ instead of $\Pr(Y\mid Z)$ as given in (3.1). However logistic regression analysis can still be applied to both sampling designs (Prentice and Pyke, 1979). In Sections 3.1 and 3.2, we will review the results of Whittemore (1995) in which the probability structure under conventional and familial case-control designs is well examined.

3.1 Conventional Case-Control Designs

Let Ω be the target population. The logistic regression model in (3.1) is equivalent to

$$\log\frac{\Pr(Y=1\mid Z,\Omega)}{\Pr(Y=0\mid Z,\Omega)}=\alpha+\beta^T Z \tag{3.3}$$

where α is the intercept that represents the log odds for developing the disease of the baseline group, and β is the log odds ratio between a subject with covariate Z and a subject of the baseline group. Since β reflects the effect of Z on Y, it is the parameter of major interest.

As mention earlier, a sample based on a case-control design involves $\Pr(Z \mid Y, \Omega)$. Applying Baye's rule, we obtain

$$\frac{\Pr(Z \mid Y=1,\Omega)}{\Pr(Z \mid Y=0,\Omega)} = \frac{\Pr(Y=0 \mid \Omega)}{\Pr(Y=1 \mid \Omega)} \exp(\alpha) \exp(\beta^T Z). \tag{3.4a}$$

Whittemore (1995) mentioned that one can imagine a hypothetical population denoted as Ω^* in which the covariate distribution is the same as in Ω such that

$$\frac{\Pr(Z \mid Y=1,\Omega^*)}{\Pr(Z \mid Y=0,\Omega^*)} = \frac{\Pr(Z \mid Y=1,\Omega)}{\Pr(Z \mid Y=0,\Omega)} = \frac{\Pr(Y=0 \mid \Omega)}{\Pr(Y=1 \mid \Omega)} \exp(\alpha) \exp(\beta^T Z). \tag{3.4b}$$

Define

$$\exp(\delta) = \left\{ \exp(\alpha) \frac{\Pr(Y=0 \mid \Omega)}{\Pr(Y=1 \mid \Omega)} \right\} \frac{\Pr(Y=1 \mid \Omega^*)}{\Pr(Y=0 \mid \Omega^*)}.$$

One can rewrite (3.4b) as

$$\frac{\Pr(Z \mid Y=1,\Omega^*)}{\Pr(Z \mid Y=0,\Omega^*)} = \frac{\Pr(Y=0 \mid \Omega^*)}{\Pr(Y=1 \mid \Omega^*)} \exp(\delta) \exp(\beta^T Z). \tag{3.5}$$

From (3.5), we can construct the following logistic model based on Ω^*:

$$\log \frac{\Pr(Y=1 \mid Z,\Omega^*)}{\Pr(Y=0 \mid Z,\Omega^*)} = \delta + \beta^T Z. \tag{3.6}$$

Now we discuss the implication of the above analysis. Comparing the two models in (3.3) and (3.6), they differ in the intercept parameter but have the same slope parameter, which is of major interest. In a case-control design, the sampling distribution is based on $\Pr(Z \mid Y=1,\Omega^*)$ and $\Pr(Z \mid Y=0,\Omega^*)$, where

$$\Pr(Z \mid Y=1,\Omega^*) = \Pr(Y=1 \mid Z,\Omega^*) \times \frac{\Pr(Z \mid \Omega^*)}{\Pr(Y=1 \mid \Omega^*)};$$

$$\Pr\left(Z \mid Y=0,\Omega^*\right)=\Pr\left(Y=0 \mid Z,\Omega^*\right)\times\frac{\Pr\left(Z \mid \Omega^*\right)}{\Pr\left(Y=0 \mid \Omega^*\right)}.$$

Notice that $\frac{\Pr\left(Z \mid \Omega^*\right)}{\Pr\left(Y \mid \Omega^*\right)}$ is independent with parameters. The likelihood function for case-control data can be constructed based on model (3.6). Accordingly case-control data can be treated as prospective data from Ω^* if the following condition holds:

$$\frac{\Pr(Z \mid Y=1,\Omega^*)}{\Pr(Z \mid Y=0,\Omega^*)}=\frac{\Pr(Z \mid Y=1,\Omega)}{\Pr(Z \mid Y=0,\Omega)}. \tag{3.7}$$

As long as (3.7) is satisfied in collecting the case-control sample, one can proceed the regression analysis, by pretending that the sample is from a prospective study, to obtain an estimate of β which is still reliable. We will examine the crucial condition in (3.7) via simulations.

3.2 Familial Case-Control Designs

In analysis of familial data, some studies ignored probands' information and only focus on relatives' data. Such an approach may lose efficiency by ignoring useful information in probands' data. Whittemore (1995) applied multivariate techniques to analyze familial case-control data. Specifically she proposed a two-stage sampling procedure. Specifically in the first stage, two types of probands (case and control) are sampled and then, in the second stage, their relatives are sampled. To simplify the discussion, we focus on bivariate analysis which means that only one relative is sampled based on each proband. The resulting likelihood analysis contains two components. One involves the logistic model on probands as introduced earlier. The other component is related to the model which measures the dependence between a proband and his/her relatives.

Let (Y_p, Z_p) and (Y_r, Z_r) be the disease status and covariates for a proband and his/her relative respectively. Denote $Y=(Y_p, Y_r)$ and $Z=(Z_p, Z_r)$. We will first discuss likelihood inference based on a prospective design and then the modification based on a case-control

design.

3.2.1 Likelihood analysis based on familial prospective data

A prospective study involves sampling from

$$\Pr(Y, Z_r \mid Z_p) = \Pr(Y \mid Z)\Pr(Z_r \mid Z_p). \tag{3.8}$$

When only one relative is involved, $\Pr(Y \mid Z) = \Pr(Y_p = y_p, Y_r = y_r \mid Z_p, Z_r)$ for $y_* = 0,1$ and $* = p, r$. Note that

$$\Pr(Y \mid Z) = \Pr(Y_p \mid Z_p, Z_r)\Pr(Y_r \mid Y_p, Z).$$

Whittemore (1995) mentioned that a reasonable joint model should satisfy the so-called "reproducible" assumption such that

$$\sum_{y_r=0}^{y_r=1} \Pr(Y_p = y_p, Y_r = y_r \mid Z_p, Z_r) = \Pr(Y_p = y_p \mid Z_p); \tag{3.9a}$$

$$\sum_{y_p=0}^{y_p=1} \Pr(Y_p = y_p, Y_r = y_r \mid Z_p, Z_r) = \Pr(Y_r = y_r \mid Z_r). \tag{3.9b}$$

That is, the covariate of a person is sufficient to determine his/her disease status and hence the relative's covariate does not contribute extra information. The paper examines the plausibility of the reproducible assumption. Suppose that the dependence between Y_p and Y_r within the same family may also be attributed to some un-measured latent variable denoted as U. If $\Pr(U \mid Z) = \Pr(U)$, the reproducible assumption can be achieved. Whether this assumption makes sense depends on the scientific problem at hand.

When (3.9a) is true, it follows that

$$\Pr(Y \mid Z) = \Pr(Y_p \mid Z_p)\Pr(Y_r \mid Y_p, Z). \tag{3.10}$$

Notice that $\Pr(Y_p \mid Z_p)$ can be modeled as

$$\log \frac{\Pr(Y_p = 1 \mid Z_p)}{\Pr(Y_p = 0 \mid Z_p)} = \alpha + \beta^T Z_p .$$

The second quantity $\Pr(Y_r \mid Y_p, Z) = \Pr(Y_r \mid Y_p, Z_p, Z_r)$ in (3.10) involves the dependence between a proband and his/her relative which is of major interest. Denote observed data as $\{(Y_{pi}, Y_{ri}, Z_{pi}, Z_{ri})\ (i = 1, ..., n)\}$. If the data is collected from a prospective sampling design, the likelihood function can be written as

$$\begin{aligned} L(\rho, \alpha, \beta) &= \prod_{i=1}^{n} \Pr\left(Y_{p_i}, Y_{r_i}, Z_{r_i} \mid Z_{p_i}\right) \\ &\propto \prod_{i=1}^{n} \Pr\left(Y_{p_i} \mid Z_{p_i}\right) \prod_{i=1}^{n} \Pr\left(Y_{r_i} \mid Y_{p_i}, Z_i\right) \\ &= L^{(1)}(\alpha, \beta) \times L^{(2)}(\rho, \alpha, \beta), \end{aligned} \tag{3.11}$$

where $L^{(1)}(\alpha, \beta)$ has the form as in (3.2) and ρ denotes additional parameter in $\Pr(Y_r \mid Y_p, Z)$. Additional joint model assumption is required to specify the form of $\Pr(Y_r \mid Y_p, Z)$.

One model choice is the following model first proposed by Bahadur (1961):

$$\Pr((Y_p, Y_r) \mid (Z_p, Z_r)) = (p_p)^{y_p} (1 - p_p)^{1 - y_p} (p_r)^{y_r} (1 - p_r)^{1 - y_r} (1 + \rho t_p t_r) \tag{3.12}$$

where

$$t_* = \frac{y_* - p_*}{\sqrt{p_* (1 - p_*)}} \qquad * = p, r ,$$

and

$$p_* \equiv \Pr(Y_* = 1 \mid Z_*) = \frac{\exp(\alpha + \beta^T Z_*)}{1 + \exp(\alpha + \beta^T Z_*)} \qquad * = p, r .$$

The coefficient ρ satisfies the following constraint:

$$-\min\left\{\sqrt{\frac{p_p p_r}{(1 - p_p)(1 - p_r)}}, \sqrt{\frac{(1 - p_p)(1 - p_r)}{p_p p_r}}\right\} \le \rho \le \min\left\{\sqrt{\frac{p_p (1 - p_r)}{(1 - p_p) p_r}}, \sqrt{\frac{p_r (1 - p_p)}{(1 - p_r) p_p}}\right\}.$$

We will check whether this model satisfies the reproducible assumption via simulations. The following table summarizes the joint probability of (Y_p, Y_r) given (Z_p, Z_r).

	$Y_r = 1$	$Y_r = 0$
$Y_p = 1$	$P_1 = (p_p)(p_r)$ $(1+\rho\frac{1-p_p}{\sqrt{p_p(1-p_p)}}\frac{1-p_r}{\sqrt{p_r(1-p_r)}})$	$P_3 = (p_p)(1-p_r)$ $(1+\rho\frac{1-p_p}{\sqrt{p_p(1-p_p)}}\frac{-p_r}{\sqrt{p_r(1-p_r)}})$
$Y_p = 0$	$P_2 = (1-p_p)(p_r)$ $(1+\rho\frac{-p_p}{\sqrt{p_p(1-p_p)}}\frac{1-p_r}{\sqrt{p_r(1-p_r)}})$	$P_4 = (1-p_p)(1-p_r)$ $(1+\rho\frac{-p_p}{\sqrt{p_p(1-p_p)}}\frac{-p_r}{\sqrt{p_r(1-p_r)}})$

Table 3.1 Joint probability for (Y_p, Y_r) given (Z_p, Z_r)

3.2.2 Likelihood analysis based on familial case-control data

A case-control study involves two independent samples from $\Pr(Y_r, Z \mid Y_p = 1)$ and $\Pr(Y_r, Z \mid Y_p = 0)$. Notice that $\Pr(Y_r, Z \mid Y_p) = \Pr(Y_r \mid Z, Y_p)\Pr(Z \mid Y_p)$ and

$$\Pr(Z \mid Y_p) = \Pr(Z_p, Z_r \mid Y_p) = \Pr(Z_p \mid Y_p)\Pr(Z_r \mid Y_p, Z_p).$$

The reproducible assumption implies that, given Z_p, Y_p and Z_r are independent. Hence $\Pr(Z \mid Y_p) = \Pr(Z_p \mid Y_p)\Pr(Z_r \mid Z_p)$. In summary we have

$$\begin{aligned}\Pr\left(Y_r, Z \mid Y_p\right) &= \Pr\left(Y_r, Z \mid Y_p\right)\Pr\left(Z_p \mid Y_p\right)\Pr\left(Z_r \mid Z_p\right) \\ &= \left\{\Pr(Z_p \mid Y_p)\right\}\left\{\Pr(Y_r \mid Z, Y_p)\Pr(Z_r \mid Z_p)\right\}. \qquad (3.13)\end{aligned}$$

Recall that $\Pr(Z_p \mid Y_p)$ can be analyzed assuming that the data is from a prospective sample from the model

$$\log \frac{\Pr(Y_p = 1 \mid Z_p, \Omega^*)}{\Pr(Y_p = 0 \mid Z_p, \Omega^*)} = \delta + \beta^T Z_p .$$

Accordingly the retrospective likelihood function is given by

$$\begin{aligned} L^*(\rho, \delta, \alpha, \beta) &= \prod_{i=1}^{n} \Pr\left(Y_{r_i}, Z_i \mid Y_{p_i}\right) \\ &\propto \prod_{i=1}^{n} \Pr\left(Z_{p_i} \mid Y_{p_i}\right) \prod_{i=1}^{n} \Pr\left(Y_{r_i} \mid Y_{p_i}, Z_i\right) \\ &= L^{*(1)}(\delta, \beta) \times L^{(2)}(\rho, \alpha, \beta) . \end{aligned}$$

Notice that the information of α is also contained in familial case-control data when the reproducible assumption holds for the joint model.

Chapter 4 Simulations for Logistic Regression Analysis

In this chapter, we propose data generation algorithms to simulate case-control data for logistic regression analysis. Some crucial probability statements will be examined to verify whether the simulated data are reliable for statistical inference.

4.1 Data Generation for Individual Data

4.1.1 Prospective data of the true population

First of all, we generate population data from the model:

$$\Pr(Y=1|Z,\Omega)=\frac{\exp(\alpha+\beta Z)}{1+\exp(\alpha+\beta Z)}.$$

Then set the values of the parameters: α, β and p. The algorithm is summarized below:

- Step 1: Generate $Z_i \sim \text{Bernoulli}(p)$
- Step 2: Given Z_i, generate $Y_i \sim \text{Bernoulli}\left(\frac{\exp(\alpha+\beta Z_i)}{1+\exp(\alpha+\beta Z_i)}\right)$.

The procedure is repeated for $i=1,...,N$ for very large N, say $N=10000$. Denote $\Omega=\{(Y_i,Z_i)(i=1,...,N)\}$.

4.1.2 Case-control data from the true population

Suppose that we generate $n \ll N$ observations from the population with n_1 persons from the case group with $Y=1$ and $n_0=n-n_1$ persons from the control group $Y=0$.

The procedure is stated as follows.

- Step 1: Randomly select n_1 subjects from the case group and record their values of Z_i;
- Step 2: Randomly select n_0 subjects from the control group and record their values of Z_i.

We briefly discuss how to implement Step 1 since Step 2 follows a similar procedure. First identify the case population: $\Omega_1 = \{(Y_i = 1, Z_i)\ (i = 1, ..., N_1)\}$ where $N_1 = \sum_{i=1}^{N} I(Y_i = 1)$. The objective is to select n_1 observations from N_1 subjects. Label the subjects in Ω_1 from 1 to N_1. At the first time, generate $U \sim U(0,1)$ and define $s = [N_1 \times U]$, where $[\ ]$ is the Gauss function. A subject with label "s" is selected into the case-control sample and removed from Ω_1. The procedure is repeated n_1 times. Specifically at the kth time, generate $U \sim U(0,1)$ and a subject with a re-defined label $s = [(N_1 - k + 1) \times U]$ is selected from the remaining case population containing $N_1 - k + 1$ subjects. Finally the case sample is formed and denoted as $\{(Y_k = 1, Z_k)\ (k = 1, .., n_1)\}$. The control sample can be generated in a similar way.

4.2. Analysis on Individual Data

We examine whether the proposed case-control sampling procedure produces reliable data. We let

$$R_* = \frac{\sum_{i=1}^{N} I(Z_i = *, Y_i = 1) / \sum_{i=1}^{N} I(Y_i = 1)}{\sum_{i=1}^{N} I(Z_i = *, Y_i = 0) / \sum_{i=1}^{N} I(Y_i = 0)} \qquad (* = 0,1)$$

and

$$r_* = \frac{\sum_{i=1}^{n} I(Z_i = *, Y_i = 1) / n_1}{\sum_{i=1}^{n} I(Z_i = *, Y_i = 0) / n_1} \qquad (* = 0,)$$

be the empirical estimates of $\frac{\Pr(Z = * | Y = 1, \Omega)}{\Pr(Z = * | Y = 0, \Omega)}$ and $\frac{\Pr(Z = * | Y = 1, \Omega^*)}{\Pr(Z = * | Y = 0, \Omega^*)}$ respectively. The first criteria to evaluate the quality of data is checking whether r_* is close to R_*. The intention is to examine whether equation (3.7) holds. Then we run logistic regression based on

the combined case-control data: $\{(Y_k = 1, Z_k)\ (k = 1, ..., n_1)\}$ and $\{(Y_k = 0, Z_k)\ (k = 1, ..., n_0)\}$. The MLE of $\hat{\delta}$ and $\hat{\beta}$ are obtained. By checking whether $\hat{\beta}$ is close to the true value, we can examine whether the case-control data provide reliable information of β. The results are summarized in Tables 4.1. We observe that the empirical estimate r_* is close to R_* obtained from the population data, the estimations of $\hat{\beta}$ are also stable and close to the true value.

Table 4.1: Logistic regression analysis of case-control data

$\beta = 0.5, N = 10000, \text{Replications} = 100$				
$n_1 = 100, n_0 = 100$				
	$\frac{\overline{r_0}}{R_0}$	$\frac{\overline{r_1}}{R_1}$	$\left(\bar{\hat{\beta}} - \beta\right) \times 10^3$	SE of $\hat{\beta}$
$p = 0.3$	1.013132	1.018201	−22.336209	0.034299
$p = 0.5$	1.024408	0.992022	−33.755393	0.027108
$p = 0.7$	1.021804	1.003206	−1.477217	0.028456
$n_1 = 50, n_0 = 150$				
	$\frac{\overline{r_0}}{R_0}$	$\frac{\overline{r_1}}{R_1}$	$\left(\bar{\hat{\beta}} - \beta\right) \times 10^3$	SE of $\hat{\beta}$
$p = 0.3$	0.991638	1.044951	31.116517	0.033287
$p = 0.5$	0.978128	1.033878	56.818868	0.032211
$p = 0.7$	1.013979	1.004441	20.400237	0.036247
$n_1 = 150, n_0 = 50$				
	$\frac{\overline{r_0}}{R_0}$	$\frac{\overline{r_1}}{R_1}$	$\left(\bar{\hat{\beta}} - \beta\right) \times 10^3$	SE of $\hat{\beta}$
$p = 0.3$	1.018330	1.038011	−17.547938	0.036607
$p = 0.5$	1.037552	1.021409	−27.831059	0.038455
$p = 0.7$	1.042732	1.009384	−6.449529	0.038106

4.3 Data Generation for Familial Data

4.3.1 Familial prospective data of the true population

We first generate data following the model proposed by Bahadur (1961). First set the values of the parameters: α, β, ρ and p. The algorithm is summarized below:

- Step 1: Generate Z_{p_i} following Bernoulli(p) and Z_{r_i} independently also following Bernoulli(p);
- Step 2: Given Z_{p_i} and Z_{r_i}, compute P_1, P_2, P_3, P_4 mentioned in Table 3.1;
- Step 3: Generate $U_i \sim \text{Uniform}(0,1)$;
- Step 4: Set $\begin{cases} Y_{p_i}=1,\ Y_{r_i}=1 \text{ if } 0 \le U_i \le P_1 \\ Y_{p_i}=0,\ Y_{r_i}=1 \text{ if } P_1 \le U_i \le P_1+P_2 \\ Y_{p_i}=1,\ Y_{r_i}=0 \text{ if } P_1+P_2 \le U_i \le P_1+P_2+P_3 \\ Y_{p_i}=0,\ Y_{r_i}=0 \text{ if } P_1+P_2+P_3 \le U_i \le 1 \end{cases}$.

The procedure is repeated for $i=1,...,N$ for $N=10000$.

4.3.2 Familial case-control data from the true population

The procedure is stated as follows.

- Step 1: Randomly select n_1 probands from the case families with $Y_{p_i}=1$ and record the values of $(Z_{p_i}, Y_{r_i}, Z_{r_i})$;
- Step 2: Randomly select n_0 probands from the control families with $Y_{p_i}=0$ and record their values of $(Z_{p_i}, Y_{r_i}, Z_{r_i})$.

4.4. Analysis on Familial Data

We first examine whether the algorithm for generating perspective data satisfies the reproducible assumption. For $z_p, z_r = 0,1$, define

$$q_1(y_p, z_p, z_r) = \frac{\sum_{i=1}^{N} I(Y_{pi} = y_p, Z_{pi} = z_p, Z_{ri} = z_r)}{\sum_{i=1}^{N} I(Z_{pi} = z_p, Z_{ri} = z_r)}, \text{ and}$$

$$\tilde{q}_1(y_p, z_p) = \frac{\sum_{i=1}^{N} I(Y_{pi} = y_p, Z_{pi} = z_p)}{\sum_{i=1}^{N} I(Z_{pi} = z_p)};$$

$$q_2(y_r, z_p, z_r) = \frac{\sum_{i=1}^{N} I(Y_{ri} = y_r, Z_{pi} = z_p, Z_{ri} = z_r)}{\sum_{i=1}^{N} I(Z_{pi} = z_p, Z_{ri} = z_r)}, \text{ and}$$

$$\tilde{q}_2(y_r, z_r) = \frac{\sum_{i=1}^{N} I(Y_{ri} = y_r, Z_{ri} = z_r)}{\sum_{i=1}^{N} I(Z_{ri} = z_r)}.$$

The reproducible condition should imply that $q_1(y_p, z_p, z_r) \approx \tilde{q}_1(y_p, z_p)$ and $q_2(y_r, z_p, z_r) \approx \tilde{q}_2(y_r, z_r)$. The results of these quantities based on prospective data and case-control data from the true population are recorded in Table 4.2 ~ 4.4. In analyzing the case-control familial data, we assume α and β are known and then obtain the MLE of ρ. By checking whether $\hat{\rho}$ is close to the true value, we can examine whether the familial case-control data provide reliable information of the association in a family. This result is given in Table 4.5.

We observe that the performance of the reproducible properties is good in our population data which means that the model is appropriate. But sometimes the reproducible properties do not reflected in the simulated case-control data. Accordingly the estimations of $\hat{\rho}$ will have worse results in these situations.

Table 4.2.A: Checking reproducible properties of the population data (p=0.3)

$N=10000$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.734649$	$q_1=0.630283$	$q_1=0.265351$	$q_1=0.369717$
$z_r=0$	$q_1=0.729743$	$q_1=0.638770$	$q_1=0.270257$	$q_1=0.361230$
	$\tilde{q}_1=0.731267$	$\tilde{q}_1=0.636183$	$\tilde{q}_1=0.268733$	$\tilde{q}_1=0.363817$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.718202$	$q_2=0.607213$	$q_2=0.281798$	$q_2=0.392787$
$z_p=0$	$q_2=0.752438$	$q_2=0.642232$	$q_2=0.247562$	$q_2=0.357768$
	$\tilde{q}_2=0.742251$	$\tilde{q}_2=0.632012$	$\tilde{q}_2=0.257749$	$\tilde{q}_2=0.367988$

Table 4.2.B: Checking reproducible properties based on case-control data (p=0.3)

$N=10000, n_1=100, n_0=100$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.545455$	$q_1=0.547170$	$q_1=0.454545$	$q_1=0.452830$
$z_r=0$	$q_1=0.696970$	$q_1=0.391304$	$q_1=0.303030$	$q_1=0.608696$
	$\tilde{q}_1=0.636364$	$\tilde{q}_1=0.448276$	$\tilde{q}_1=0.363636$	$\tilde{q}_1=0.551724$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.590909$	$q_2=0.575758$	$q_2=0.409091$	$q_2=0.424242$
$z_p=0$	$q_2=0.698113$	$q_2=0.521739$	$q_2=0.301887$	$q_2=0.478261$
	$\tilde{q}_2=0.666667$	$\tilde{q}_2=0.536000$	$\tilde{q}_2=0.333333$	$\tilde{q}_2=0.464000$

Table 4.2.C: Checking reproducible properties based on case-control data (p=0.3)

$N=10000, n_1=50, n_0=150$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.363636$	$q_1=0.282609$	$q_1=0.636364$	$q_1=0.717391$
$z_r=0$	$q_1=0.222222$	$q_1=0.218750$	$q_1=0.777778$	$q_1=0.781250$
	$\tilde{q}_1=0.275862$	$\tilde{q}_1=0.239437$	$\tilde{q}_1=0.724138$	$\tilde{q}_1=0.760563$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.590909$	$q_2=0.222222$	$q_2=0.409091$	$q_2=0.777778$
$z_p=0$	$q_2=0.652174$	$q_2=0.427083$	$q_2=0.347826$	$q_2=0.572917$
	$\tilde{q}_2=0.632353$	$\tilde{q}_2=0.371212$	$\tilde{q}_2=0.367647$	$\tilde{q}_2=0.628788$

Table 4.2.D: Checking reproducible properties based on case-control data (p=0.3)

$N=10000, n_1=150, n_0=50$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.782609$	$q_1=0.763158$	$q_1=0.217391$	$q_1=0.236842$
$z_r=0$	$q_1=0.818182$	$q_1=0.690476$	$q_1=0.181818$	$q_1=0.309524$
	$\tilde{q}_1=0.807692$	$\tilde{q}_1=0.713115$	$\tilde{q}_1=0.192308$	$\tilde{q}_1=0.286885$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.695652$	$q_2=0.527273$	$q_2=0.304348$	$q_2=0.472727$
$z_p=0$	$q_2=0.842105$	$q_2=0.619048$	$q_2=0.157895$	$q_2=0.380952$
	$\tilde{q}_2=0.786885$	$\tilde{q}_2=0.582734$	$\tilde{q}_2=0.213115$	$\tilde{q}_2=0.417266$

Table 4.3.A: Checking reproducible properties of the population data (p=0.5)

$N=10000$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.725902$	$q_1=0.611177$	$q_1=0.274098$	$q_1=0.388823$
$z_r=0$	$q_1=0.739654$	$q_1=0.627195$	$q_1=0.260346$	$q_1=0.372805$
	$\tilde{q}_1=0.732735$	$\tilde{q}_1=0.619038$	$\tilde{q}_1=0.267265$	$\tilde{q}_1=0.380962$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.729869$	$q_2=0.611089$	$q_2=0.270131$	$q_2=0.388911$
$z_p=0$	$q_2=0.726092$	$q_2=0.616170$	$q_2=0.273908$	$q_2=0.383830$
	$\tilde{q}_2=0.727973$	$\tilde{q}_2=0.613609$	$\tilde{q}_2=0.272027$	$\tilde{q}_2=0.386391$

Table 4.3.B: Checking reproducible properties based on case-control data (p=0.5)

$N=10000, n_1=100, n_0=100$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.500000$	$q_1=0.390244$	$q_1=0.500000$	$q_1=0.609756$
$z_r=0$	$q_1=0.548387$	$q_1=0.529412$	$q_1=0.451613$	$q_1=0.470588$
	$\tilde{q}_1=0.527778$	$\tilde{q}_1=0.467391$	$\tilde{q}_1=0.472222$	$\tilde{q}_1=0.532609$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.695652$	$q_2=0.532258$	$q_2=0.304348$	$q_2=0.467742$
$z_p=0$	$q_2=0.609756$	$q_2=0.470588$	$q_2=0.390244$	$q_2=0.529412$
	$\tilde{q}_2=0.655172$	$\tilde{q}_2=0.504425$	$\tilde{q}_2=0.344828$	$\tilde{q}_2=0.495575$

Table 4.3.C: Checking reproducible properties based on case-control data (p=0.5)

$N=10000, n_1=50, n_0=150$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.325581$	$q_1=0.245614$	$q_1=0.674419$	$q_1=0.754386$
$z_r=0$	$q_1=0.306122$	$q_1=0.137255$	$q_1=0.693878$	$q_1=0.862745$
	$\tilde{q}_1=0.315217$	$\tilde{q}_1=0.194444$	$\tilde{q}_1=0.684783$	$\tilde{q}_1=0.805556$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.604651$	$q_2=0.244898$	$q_2=0.395349$	$q_2=0.755102$
$z_p=0$	$q_2=0.614035$	$q_2=0.215686$	$q_2=0.385965$	$q_2=0.784314$
	$\tilde{q}_2=0.610000$	$\tilde{q}_2=0.230000$	$\tilde{q}_2=0.390000$	$\tilde{q}_2=0.770000$

Table 4.3.D: Checking reproducible properties based on case-control data (p=0.5)

$N=10000, n_1=150, n_0=50$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.825397$	$q_1=0.636364$	$q_1=0.174603$	$q_1=0.363636$
$z_r=0$	$q_1=0.733333$	$q_1=0.787879$	$q_1=0.266667$	$q_1=0.212121$
	$\tilde{q}_1=0.780488$	$\tilde{q}_1=0.701299$	$\tilde{q}_1=0.219512$	$\tilde{q}_1=0.298701$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.777778$	$q_2=0.550000$	$q_2=0.222222$	$q_2=0.450000$
$z_p=0$	$q_2=0.727273$	$q_2=0.727273$	$q_2=0.272727$	$q_2=0.272727$
	$\tilde{q}_2=0.757009$	$\tilde{q}_2=0.612903$	$\tilde{q}_2=0.242991$	$\tilde{q}_2=0.387097$

Table 4.4.A: Checking reproducible properties of the population data (p=0.7)

$N=10000$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.725744$	$q_1=0.631256$	$q_1=0.274256$	$q_1=0.368744$
$z_r=0$	$q_1=0.713138$	$q_1=0.625140$	$q_1=0.286862$	$q_1=0.374860$
	$\tilde{q}_1=0.721928$	$\tilde{q}_1=0.629445$	$\tilde{q}_1=0.278072$	$\tilde{q}_1=0.370555$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.718154$	$q_2=0.607750$	$q_2=0.281846$	$q_2=0.392250$
$z_p=0$	$q_2=0.731822$	$q_2=0.622896$	$q_2=0.268178$	$q_2=0.377104$
	$\tilde{q}_2=0.722294$	$\tilde{q}_2=0.612238$	$\tilde{q}_2=0.277706$	$\tilde{q}_2=0.387762$

Table 4.4.B: Checking reproducible properties based on case-control data (p=0.7)

$N=10000, n_1=100, n_0=100$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.505155$	$q_1=0.428571$	$q_1=0.494845$	$q_1=0.571429$
$z_r=0$	$q_1=0.577778$	$q_1=0.437500$	$q_1=0.422222$	$q_1=0.562500$
	$\tilde{q}_1=0.528169$	$\tilde{q}_1=0.431034$	$\tilde{q}_1=0.471831$	$\tilde{q}_1=0.568966$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.659794$	$q_2=0.422222$	$q_2=0.340206$	$q_2=0.577778$
$z_p=0$	$q_2=0.642857$	$q_2=0.375000$	$q_2=0.357143$	$q_2=0.625000$
	$\tilde{q}_2=0.654676$	$\tilde{q}_2=0.409836$	$\tilde{q}_2=0.345324$	$\tilde{q}_2=0.590164$

Table 4.4.C: Checking reproducible properties based on case-control data (p=0.7)

$N=10000, n_1=50, n_0=150$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.261364$	$q_1=0.111111$	$q_1=0.738636$	$q_1=0.888889$
$z_r=0$	$q_1=0.363636$	$q_1=0.260870$	$q_1=0.636364$	$q_1=0.739130$
	$\tilde{q}_1=0.295455$	$\tilde{q}_1=0.161765$	$\tilde{q}_1=0.704545$	$\tilde{q}_1=0.838235$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.431818$	$q_2=0.386364$	$q_2=0.568182$	$q_2=0.613636$
$z_p=0$	$q_2=0.600000$	$q_2=0.521739$	$q_2=0.400000$	$q_2=0.478261$
	$\tilde{q}_2=0.488722$	$\tilde{q}_2=0.432836$	$\tilde{q}_2=0.511278$	$\tilde{q}_2=0.567164$

Table 4.4.D: Checking reproducible properties based on case-control data (p=0.7)

$N=10000, n_1=150, n_0=50$				
	$(y_p, z_p)=(1,1)$	$(y_p, z_p)=(1,0)$	$(y_p, z_p)=(0,1)$	$(y_p, z_p)=(0,0)$
$z_r=1$	$q_1=0.724490$	$q_1=0.575758$	$q_1=0.275510$	$q_1=0.424242$
$z_r=0$	$q_1=0.857143$	$q_1=0.900000$	$q_1=0.142857$	$q_1=0.100000$
	$\tilde{q}_1=0.768707$	$\tilde{q}_1=0.698113$	$\tilde{q}_1=0.231293$	$\tilde{q}_1=0.301887$
	$(y_r, z_r)=(1,1)$	$(y_r, z_r)=(1,0)$	$(y_r, z_r)=(0,1)$	$(y_r, z_r)=(0,0)$
$z_p=1$	$q_2=0.734694$	$q_2=0.755102$	$q_2=0.265306$	$q_2=0.244898$
$z_p=0$	$q_2=0.787879$	$q_2=0.800000$	$q_2=0.212121$	$q_2=0.200000$
	$\tilde{q}_2=0.748092$	$\tilde{q}_2=0.768116$	$\tilde{q}_2=0.251908$	$\tilde{q}_2=0.231884$

Table 4.5: The MLE of ρ based on case-control familial data

	$N=10000, \beta=0.5$ $\rho=0.5$		
	$n_1=100, n_0=100$	$n_1=50, n_0=150$	$n_1=150, n_0=50$
$p=0.3$	$\hat{\rho}=0.522000$	$\hat{\rho}=0.400637$	$\hat{\rho}=0.473103$
$p=0.5$	$\hat{\rho}=0.467715$	$\hat{\rho}=0.614172$	$\hat{\rho}=0.519416$
$p=0.7$	$\hat{\rho}=0.562765$	$\hat{\rho}=0.534639$	$\hat{\rho}=0.399379$

Chapter 5 Regression Analysis Based on Familial Data

For those who have developed the disease, the age at onset may be informative. As mentioned in Li et al. (1998), early age of onset has been a hallmark for genetic predisposition in most of diseases that aggregate in families. When age-at-onset is chosen as the primary response, the effect of censoring has to be considered in the analysis.

In this chapter, we discuss several important issues on analyzing familial age-onset data. Specifically denote T as the age-onset variable and Z as a $p\times 1$ vector of covariates. The Cox proportional hazards model is the most well-known model for failure time variables which can be written as

$$\lambda(t\mid Z)=\lambda_0(t)\exp\left(\beta^T Z\right), \tag{5.1}$$

where $\lambda_0(t)$ is the baseline hazard function and β measures the effect of Z on the hazard and is of major interest. In familial failure-time analysis, the Cox model is imposed on probands. For inference of β, we first review the analysis based on a prospective sample and then extend the discussion to a valid case-control sample. Finally we will discuss the modeling and inference frameworks when familial case-control data are collected.

5.1 Likelihood Analysis Based on Probands

Under right censoring, let C be the censoring variable. One observes that $X=T\wedge C$, $\delta=I(T\le C)$ and covariates Z. In prospective studies, we identify a sample of individuals with specified covariates: Z_i and then determine their observed time and disease status: (X_i,δ_i) for $i=1,...,n$. At time t, the risk set can be denoted as $R(t)=\{i: X_i\ge t, i=1,...,n\}$. Given the risk set information, a subject failing at time t with covariate Z_j given $j\in R(t)$ will contribute to the partial likelihood by

$$\frac{\lambda\left(t\mid Z_j\right)}{\sum_{i\in R(t)}\lambda\left(t\mid Z_i\right)}=\frac{\lambda_0(t)\exp\left(\beta^T Z_j\right)}{\sum_{i\in R(t)}\lambda_0(t)\exp\left(\beta^T Z_i\right)}=\frac{\exp\left(\beta^T Z_j\right)}{\sum_{i\in R(t)}\exp\left(\beta^T Z_i\right)}. \tag{5.2}$$

Prentice and Breslow (1978) discussed the likelihood formulation based on case-control age-onset data. It is important to first introduce the sampling procedure which involves how to match a case subject with a control subject. Specifically at time t, $m(t)$ observations are sampled from the case population containing those who develop the disease at time t and, independently, $n(t)$ observations are sampled from the control population containing those who have not developed the disease up to time t. Observed data can be summarized in Table 5.1.

Time: t_i	Case: $(X = t_i, \delta = 1)$	Control: $(X = t_i, \delta = 0)$
t_1	$m(t_1)$ individuals	$n(t_1)$ individuals
⋮	⋮	⋮
t_i	$m(t_i)$ individuals	$n(t_i)$ individuals
⋮	⋮	⋮
t_k	$m(t_k)$ individuals	$n(t_k)$ individuals

Table 5.1 Age-matched case-control data

The case-control design for collecting age-onset data considers sampling from the conditional distribution of Z based on (X, δ). To establish the relationship between prospective and retrospective samples, Prentice and Breslow (1978) extended the result of Cornfield (1951) to age-onset data and derived the following condition:

$$\frac{\Pr(Z \mid X = t, \delta = 1) / \Pr(Z = 0 \mid X = t, \delta = 1)}{\Pr(Z \mid X = t, \delta = 0) / \Pr(Z = 0 \mid X = t, \delta = 0)} = \frac{\Pr(X = t, \delta = 1 \mid Z) / \Pr(X = t, \delta = 1 \mid Z = 0)}{\Pr(X = t, \delta = 0 \mid Z) / \Pr(X = t, \delta = 0 \mid Z = 0)}. \tag{5.3}$$

Notice that when C is independent of both T and Z, we have

$$\frac{\Pr(X = t, \delta = 1 \mid Z)}{\Pr(X = t, \delta = 0 \mid Z)} = \frac{\Pr(T = t, C > t \mid Z)}{\Pr(T > t, C = t \mid Z)} = \frac{\Pr(T = t \mid Z)}{\Pr(T > t \mid Z)} \frac{\Pr(C > t)}{\Pr(C = t)} = \frac{\lambda(t)}{\lambda_C(t)};$$

$$\frac{\Pr(X = t, \delta = 1 \mid Z = 0)}{\Pr(X = t, \delta = 0 \mid Z = 0)} = \frac{\Pr(T = t \mid Z = 0)}{\Pr(T > t \mid Z = 0)} \frac{\Pr(C > t)}{\Pr(C = t)} = \frac{\lambda_0(t)}{\lambda_C(t)}.$$

Hence the right-hand side of (5.3) equals $\lambda(t)/\lambda_0(t)$ and, under the proportional hazard model, (5.3) becomes

$$\frac{\Pr(Z \mid X = t, \delta = 1)/\Pr(Z = 0 \mid X = t, \delta = 1)}{\Pr(Z \mid X = t, \delta = 0)/\Pr(Z = 0 \mid X = t, \delta = 0)} = \exp(\beta^T Z). \tag{5.4}$$

Rearranging (5.4), we obtain

$$\Pr(Z \mid X = t, \delta = 1) = \Pr(Z \mid X = t, \delta = 0)\exp(\beta^T Z)\frac{\Pr(Z = 0 \mid X = t, \delta = 1)}{\Pr(Z = 0 \mid X = t, \delta = 0)}$$

which is equation (2) in Li et al. (1998). The left-hand side of (5.4) is identifiable based on case-control data which implies that β is also identifiable based on such data.

Prentice and Breslow (1978) proposed a conditional likelihood approach for estimating β based on case-control data. At time t, define $R(m(t), n(t))$ as a set of all subsets of size $m(t)$ from a total of $m(t)+n(t)$ subjects. Given this risk set information, the first $m(t)$ subjects with covariates $Z_1, ..., Z_{m(t)}$ respectively actually belonging to the case group will contribute the probability

$$\frac{\prod_{i=1}^{m(t)} \lambda(t \mid Z_i)}{\sum_{l \in R(m(t),n(t))} \prod_{j=1}^{m(t)} \lambda(t \mid Z_{lj})}, \tag{5.5}$$

where Z_{lj} denotes the covariate value for j th subject in the l th combinations. Notice that

$$\prod_{i=1}^{m(t)} \lambda(t \mid Z_i) = \{\lambda_0(t)\}^{m(t)} \exp\{\beta^T (Z_1 + ... + Z_{m(t)})\}$$

and

$$\prod_{j=1}^{m(t)} \lambda(t \mid Z_{lj}) = \{\lambda_0(t)\}^{m(t)} \exp\{\beta^T (Z_{l1} + ... + Z_{lm(t)})\}.$$

It follows that

$$\frac{\prod_{i=1}^{m(t)} \lambda(t \mid Z_i)}{\sum_{l \in R(m(t),n(t))} \prod_{j=1}^{m(t)} \lambda(t \mid Z_{lj})} = \frac{\exp(s^T \beta)}{\sum_{l \in R(m(t),n(t))} \exp(s_l^T \beta)}, \tag{5.6}$$

where $s = Z_1 + ... + Z_{m(t)}$ and $s_l = Z_{l1} + ... + Z_{lm(t)}$. Finally the likelihood can be written as

$$\prod_{j=1}^{k} \frac{\exp(s^T \beta)}{\sum_{l \in R(m(t_j), n(t_j))} \exp(s_l^T \beta)}, \tag{5.7}$$

where $t_1 < ... < t_k$ denote observed failure times for the case group. It is important to note that $R(m(t), n(t))$ only includes subjects who are sampled from the retrospective study at time t. Hence it does not have the nested property of a regular risk set such as $R(t) \subset R(t-)$. It is important to mention that computation of (5.7) involves all possible permutations in the denominator which is very time-consuming if $m(t) + n(t)$ is not small. Several authors proposed algorithms to approximate the likelihood.

We provide a numerical example to illustrate construction of $R(m, n)$ in which the label t is ignored to simply the presentation. Suppose the case-sample contains subjects with covariates Z_1 and Z_2 respectively and the matched control-sample contains subjects with covariates Z_3 and Z_4 respectively. Hence $R(2,2)$ consists of $\binom{4}{2}$ combinations which can be labeled by $l = 1, ..., 6$ corresponding to $(Z_1, Z_2), (Z_1, Z_3), (Z_1, Z_4), (Z_2, Z_3), (Z_2, Z_4), (Z_3, Z_4)$ sets of covariates respectively. For example $Z_{22} = Z_3$ corresponds to the second covariate with $l = 2$. It follows that $s = Z_1 + Z_2$, $s_1 = Z_1 + Z_2$, $s_2 = Z_1 + Z_3$, $s_3 = Z_1 + Z_4$, $s_4 = Z_2 + Z_3$, $s_5 = Z_2 + Z_4$ and $s_6 = Z_3 + Z_4$.

5.2 Likelihood Analysis Based on Familial Data

Table 5.2 summarizes observed case-control familial data in which probands' times (onset or censored) are matched. Specifically at time t_i, we sample $m(t_i)$ case probands and their relatives and matched with $n(t_i)$ control probands and their relatives. Denote $X = (X_p, X_r)$ as observed times and $\delta = (\delta_p, \delta_r)$ as the corresponding indicators for a proband and his/her relative respectively.

Time	**Case family:** $\left(X_p = t_i, \delta_p = 1\right)$	**Control family:** $\left(X_p = t_i, \delta_p = 0\right)$
t_1	$m(t_1)$ probands and their relatives	$n(t_1)$ probands and their relatives
⋮	⋮	⋮
t_i	$m(t_i)$ probands and their relatives	$n(t_i)$ probands and their relatives
⋮	⋮	⋮
t_k	$m(t_k)$ probands and their relatives	$n(t_k)$ probands and their relatives

Table 5.2 Age-matched case-control familial data

To simply the presentation, assume that there are two members in one family (one proband and one relative). Observed information for a case subject includes $(X_p = t, \delta_p = 1, Z_p, X_r, \delta_r, Z_r)$ while the information for the corresponding age-matched control subject includes $(X_p = t, \delta_p = 0, Z_p, X_r, \delta_r, Z_r)$. Two samples from $\Pr((X_r, \delta_r), Z_p, Z_r \mid (X_p = t, \delta_p = 1))$ and $\Pr((X_r, \delta_r), Z_p, Z_r \mid (X_p = t, \delta_p = 0))$ are drawn independently.

Li et al. (1998) extended the discussions in Whittemore (1995) from binary data to age-onset data. The model assumption consists of two stages. In the first stage, the model on the proband namely $\Pr(Z_p \mid (X_p, \delta_p))$, is assumed to follow the Cox model. In the second stage, the model on $\Pr((X_r, \delta_r) \mid (X_p, \delta_p), Z)$ is constructed where $Z = (Z_p, Z_r)$. It should

be mentioned that a crucial reproducible assumption to simply the analysis is given by

$$\Pr\{(X_*, \delta_*) \mid Z_p, Z_r) = \Pr\{(X_*, \delta_*) \mid Z_*) \quad (* = p,r). \tag{5.8}$$

This assumption only holds when covariates Z within a family do not depend on any unmeasured variables responsible for the correlation of age-onset. One can write

$$\Pr((X_r, \delta_r), Z_r, Z_p \mid (X_p, \delta_p)) = \Pr((X_r, \delta_r), Z_r \mid (X_p, \delta_p), Z_p) \Pr(Z_p \mid (X_p, \delta_p)).$$

Note that

$$\Pr((X_r, \delta_r,) Z_r \mid (X_p, \delta_p), Z_p) = \Pr((X_r, \delta_r) \mid (X_p, \delta_p), Z_p, Z_r) \Pr(Z_r \mid (X_p, \delta_p), Z_p).$$

By (5.8), we know that, given Z_p, (X_p, δ_p) and Z_r are independent. So, we have

$$\Pr(Z_r \mid (X_p, \delta_p), Z_p) = \Pr(Z_r \mid Z_p).$$

Therefore, formulated the likelihood function for familial case-control data can be based on the following decompositions:

$$\Pr(Z_p \mid (X_p, \delta_p)) \Pr((X_r, \delta_r) \mid (X_p, \delta_p), Z) \Pr(Z_r \mid Z_p). \tag{5.9}$$

The first component of (5.9) can be analyzed based on probands' data which can be performed in the first stage. Recall the under the Cox model assumption discussed in Section 5.1, the form of $\Pr(Z_p \mid (X_p, \delta_p))$ can be specified and equation (5.7) can be applied to estimate β. The third component $\Pr(Z_r \mid Z_p)$ in (5.9) can be treated as a constant.

To specify the form of $\Pr((X_r, \delta_r) \mid (X_p, \delta_p), Z)$ in (5.9), Li et al. (1998) adopted the Clayton model (Clayton, 1978) which is the most popular assumption for bivariate failure-time data. Recall that T_p and T_r represent the failure times of a proband and his/her relative respectively. When (T_p, T_r) follows the Clayton model , the joint survival function can be written as

$$S(s,t)=\Pr(T_p>s,T_r>t)=\begin{cases}\left[S_p(s)^{1-\alpha}+S_r(t)^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}} & \text{if } \alpha>1\\ S_p(s)\times S_r(t) & \text{if } \alpha=1\end{cases} \qquad (5.10)$$

where $S_*(t)=\exp\left[-\int_0^t \lambda_0(v)\exp\left(\beta^T Z_*\right)dv\right]=S_0(t)^{\exp\left(\beta^T Z_*\right)}$ for $*=p,r$ assuming that the Cox model (5.1) and α is an association parameter measuring the increased risk of an individual if his/her relative has the disease rather than being disease free at a given age. Note that $\alpha=\frac{1+\tau}{1-\tau}$, where τ is Kendall's tau.

The next objective is to derive the joint probability of $(X,\delta)=(X_p,X_r,\delta_p,\delta_r)$ where $X_*=T_*\wedge C_*$, $\delta_*=I(T_*\le C_*)$ and (T_p,T_r) are subject to censoring independently by (C_p,C_r). Notice that

$$\frac{\partial S_*(t)}{\partial t}=\frac{\partial\{S_0(t)^{\exp\left(\beta^T Z_*\right)}\}}{\partial t}=\exp\left(\beta^T Z_*\right)\left\{S_0(t)^{\exp\left(\beta^T Z_*\right)-1}\right\}\left\{\frac{\partial S_0(t)}{\partial t}\right\},$$

and

$$\frac{\partial S_0(t)}{\partial t}=\frac{\partial\exp\left[-\int_0^t \lambda_0(s)ds\right]}{\partial t}=-\lambda_0(t)\exp\left[-\int_0^t \lambda_0(s)ds\right].$$

Accordingly, we have

$$\frac{\partial S_*(t)}{\partial t}=-\lambda_0(t)\exp\left(\beta^T Z_*\right)\left\{S_0(t)^{\exp\left(\beta^T Z_*\right)}\right\}. \qquad (5.11)$$

Define

$$C_\alpha(u,v)=\begin{cases}\left[u^{1-\alpha}+v^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}} & \text{if } \alpha>1\\ u\times v & \text{if } \alpha=1.\end{cases}$$

By (5.11), it follows that $\Pr(X_p=x_p,X_r=x_r,\delta_p=1,\delta_r=1\,|\,Z)$ is proportional to

$$\left.\frac{\partial^2 C_\alpha(u,v)}{\partial x_p \partial x_r}\right|_{(u,v)=\{S_p(x_p),S_r(x_r)\}}$$

$$=\alpha\left[u^{1-\alpha}+v^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}-2}\left\{u^{-\alpha}\frac{\partial u}{\partial x_p}\right\}\left\{v^{-\alpha}\frac{\partial v}{\partial x_r}\right\}\Bigg|_{(u,v)=\{S_p(x_p),S_r(x_r)\}}$$

$$=\alpha\left[S_p(x_p)^{1-\alpha}+S_r(x_r)^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}-2}$$
$$\times S_0(x_p)^{(1-\alpha)\exp(\beta^T Z_p)}\lambda_0(x_p)\exp(\beta^T Z_p)$$
$$\times S_0(x_r)^{(1-\alpha)\exp(\beta^T Z_r)}\lambda_0(x_r)\exp(\beta^T Z_r).$$

Similarly $\Pr(X_p=x_p, X_r=x_r, \delta_p=1, \delta_r=0\,|\,Z)$ is proportional to

$$-\left.\frac{\partial C_\alpha(u,v)}{\partial x_p}\right|_{(u,v)=\{S_p(x_p),S_r(x_r)\}}$$

$$=-\left[u^{1-\alpha}+v^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}-1}\left\{u^{-\alpha}\frac{\partial u}{\partial x_p}\right\}\Bigg|_{(u,v)=\{S_p(x_p),S_r(x_r)\}}$$

$$=\left[S_p(x_p)^{1-\alpha}+S_r(x_r)^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}-1}$$
$$\times\left(S_0(x_p)^{(1-\alpha)\exp(\beta^T Z_p)}\lambda_0(x_p)\exp(\beta^T Z_p)\right);$$

$$\Pr(X_p=x_p, X_r=x_r, \delta_p=0, \delta_r=1\,|\,Z)\propto -\left.\frac{\partial C_\alpha(u,v)}{\partial x_r}\right|_{(u,v)=\{S_p(x_p),S_r(x_r)\}}$$

$$=-\left[u^{1-\alpha}+v^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}-1}\left\{v^{-\alpha}\frac{\partial v}{\partial x_r}\right\}\Bigg|_{(u,v)=\{S_p(x_p),S_r(x_r)\}}$$

$$=\left[S_p(x_p)^{1-\alpha}+S_r(x_r)^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}-1}$$
$$\times\left(S_0(x_r)^{(1-\alpha)\exp(\beta^T Z_r)}\lambda_0(x_r)\exp(\beta^T Z_r)\right);$$

$$\Pr(X_p=x_p, X_r=x_r, \delta_p=0, \delta_r=0\,|\,Z)\propto C_\alpha(u,v)\Big|_{(u,v)=\{S_p(x_p),S_r(x_r)\}}$$

$$=\left[u^{1-\alpha}+v^{1-\alpha}-1\right]^{\frac{1}{1-\alpha}}\Bigg|_{(u,v)=\{S_p(x_p),S_r(x_r)\}}$$

$$= \left[S_p\left(x_p\right)^{1-\alpha} + S_r\left(x_r\right)^{1-\alpha} - 1 \right]^{\frac{1}{1-\alpha}}.$$

Accordingly we obtain

$$\Pr\left(X, \delta \mid Z\right) \propto \left[S_0\left(x_p\right)^{(1-\alpha)\exp\left(\beta^T Z_p\right)} + S_0\left(x_r\right)^{(1-\alpha)\exp\left(\beta^T Z_r\right)} - 1 \right]^{\frac{1}{1-\alpha}-D}$$
$$\times \prod_{j=1}^{D} \left\{ \left[\alpha\left(j-1\right)+2-j\right] \left[S_0\left(x_j\right)\right]^{(1-\alpha)\exp\left(\beta^T Z_j\right)} \lambda_0\left(x_j\right) \exp\left(\beta^T Z_j\right) \right\} \qquad (5.12)$$

where D is the number of individuals with disease among the proband and the relative. Let $\Lambda_0\left(t\right)$ be a baseline cumulative hazard function up to time t. The conditional probability $\Pr\left(\left(X_r, \delta_r\right) \mid \left(X_p, \delta_p = 0\right), Z\right)$ for a control family is obtained by dividing (5.12) by $\Pr\left(X_p = x_p, \delta_p = 0 \mid Z\right) = \exp\left(-\Lambda_0\left(x_p\right)\exp\left(\beta^T Z_p\right)\right)$. In addition, the probability $\Pr\left(\left(X_r, \delta_r\right) \mid \left(X_p, \delta_p = 1\right), Z\right)$ for a case family is dividing (5.12) by $\Pr\left(X_p = x_p, \delta_p = 1 \mid Z\right) = \exp\left(-\Lambda_0\left(X_p\right)\exp\left(\beta^T Z_p\right)\right)\lambda_0\left(X_p\right)\exp\left(\beta^T Z_p\right)$. Finally, the retrospective likelihood function at the time t_i is given as:

$$L\left(t_i\right) = \prod_{j=1}^{m\left(t_i\right)+n\left(t_i\right)} \Pr\left(\left(X_{r_j}, \delta_{r_j}\right), Z_j \mid \left(X_{p_j}, \delta_{p_j}\right)\right)$$
$$\propto \prod_{j=1}^{m\left(t_i\right)+n\left(t_i\right)} \Pr\left(Z_{p_j} \mid \left(X_{p_j}, \delta_j\right)\right) \prod_{j=1}^{m\left(t_i\right)+n\left(t_i\right)} \Pr\left(\left(X_{r_j}, \delta_{r_j}\right) \mid \left(X_{p_j}, \delta_{p_j}\right), Z_j\right)$$

$$= L_i^{(1)}\left(\beta\right) \times L_i^{(2)}\left(\beta, \alpha\right) \qquad (5.13)$$

The resulting likelihood is then the product of terms (5.13) over the k distinct times: $L = \prod_{i=1}^{k} L\left(t_i\right)$. So we can obtain the MLE of β and α through this likelihood for a case-control study.

Chapter 6 Simulations for Analysis of Age-onset Data from Case-control Family Studies

We examine some probability statements based on the case-control design by simulations. We generate probands' data based on the Cox model in (5.1) which follows $S(t)=\exp\left[-\int_0^t \lambda_0(s)\exp(\beta^T Z)ds\right]$. The generation procedure can be stated as follows. Let $S(T)=U\sim U(0,1)$. Under the assumption of Cox proportional model, we have $U=\exp\left[-\int_0^T \lambda_0(s)ds\times\exp(\beta^T Z)\right]$. It implies that $T=\Lambda_0^{-1}\left[\frac{-\log(U)}{\exp(\beta^T Z)}\right]$, where $\Lambda_0(t)=\int_0^t \lambda_0(s)ds$. If we set $T\,|\,Z=0\sim \text{Exp}(1)$. Hence it follows that $T=\frac{-\log(U)}{\exp(\beta^T Z)}$, where $-\log(U)$ follows Exp(1).

6.1 Data Generation for Individual Data

6.1.1 Prospective data of the true population

First of all, we generate the population data based on Cox PH model (5.1) and the afore mentioned assumptions. Set the values of β and p. The algorithm is summarized below:

- Step 1: Generate $Z_i\sim \text{Bernoulli}(p)$;
- Step 2: Generate $U_i\sim U(0,1)$;
- Step 3: Generate $T_i=\frac{-\log(U_i)}{\exp(\beta^T Z_i)}$;
- Step 4: Generate $C_i\sim U(0,K)$, K is a fixed number;
- Step 5: We have $\{(X_i,\delta_i),Z_i\}$, where $X_i=T_i\wedge C_i$ and $\delta_i=I(T_i\le C_i)$.

The procedure is repeated for $i=1,...,N=10000$.

6.1.2 Case-control data from the true population

Suppose that we generate $n << N$ observations from the population with $n_1 = n/2$ persons from the case group with $(X_i, \delta_i = 1)$ and then we construct the matched $n_0 = n_1 = n/2$ control group with $(X_i, \delta_i = 0)$. The procedure is stated as follows.

- Step 1: Randomly select n_1 subjects from the case group and record their values of Z_i;
- Step 2: For each case we match with a control from control group with $|X_1 - X_0| < 0.1$, where X_1 and X_0 are observed times for a case and a control respectively.

The method of Step 1 is the same as Step 1 in section 4.1.2. And when a control has been collected to match a case, he/she will not be picked up again to match another case individual.

6.2 Analysis on Individual Data

Now we want to know whether the case-control sampling procedure produces reliable data. Let $Q = \dfrac{\sum_{i=1}^{n} I(Z_i = 1, X_i, \delta_i = 1) / \sum_{i=1}^{n} I(Z_i = 0, X_i, \delta_i = 1)}{\sum_{i=1}^{n} I(Z_i = 1, X_i, \delta_i = 0) / \sum_{i=1}^{n} I(Z_i = 0, X_i, \delta_i = 0)}$ be the empirical estimate of $\dfrac{\Pr(Z=1 \mid X=t, \delta=1) / \Pr(Z=0 \mid X=t, \delta=1)}{\Pr(Z=1 \mid X=t, \delta=0) / \Pr(Z=0 \mid X=t, \delta=0)}$. We then examine whether the equation (5.4) can be achieved by checking whether Q is close to $\exp(\beta \times 1)$.

Then we analyze the case-control data by solving the MLE of β and then check whether $\hat{\beta}$ is close to the true value of β. The results are summarized in Table 6.1. We obtain that when Q is close to $\exp(\beta^T Z)$, the estimation of $\hat{\beta}$ is also closer to our true value.

Table 6.1: Analysis of age-onset data
based on case-control studies
$N = 10000, n_1 = n_0 = 100, \text{Replications} = 100$
$\beta = 0.5$

	$\overline{Q - \exp(\beta)}$	$\left(\overline{\hat{\beta}} - \beta\right) \times 10^3$	SE of $\hat{\beta}$
$p = 0.3$	-0.047896	30.762727	0.029843
$p = 0.5$	-0.043819	24.623113	0.028702
$p = 0.7$	-0.020804	-18.990294	0.031158

6.3 Data Generation for Familial Data

6.3.1 Familial prospective data of the true population

We now generate familial data following the Clayton model of the form in (5.10). We also let $T \mid Z = 0 \sim \text{Exp}(1)$, so $S_0(t) = \exp(-t)$. Then we can modify the data generation procedure for the Clayton model originally proposed by Prentice and Cai (1992). The algorithm is summarized below:

- Step 1: Generate $Z_{*_i} \sim \text{Bernoulli}(p)$ for $* = p, r$;
- Step 2: Set $\tau = 0.5 \Rightarrow \theta = \dfrac{(1-\tau)}{2\tau}$;
- Step 3: Generate independent variables (U_{p_i}, U_{r_i}), $U_{*_i} \sim U(0,1)$ for $* = p, r$;
- Step 4: Generate (T_{p_i}', T_{r_i}'): the baseline group with $Z_{*_i} = 0$, and $T_{*_i}' \sim \text{Exp}(1)$ for $* = p, r$ by setting

$$a = \left(1 - U_{r_i}\right)^{-\frac{1}{\theta}},$$

$$T_{p_i}' = \theta \log\left\{(1-a) + a\left(1 - U_{p_i}\right)^{-\frac{1}{1+\theta}}\right\},$$

$$T_{r_i}' = -\log\left\{\left(1 - U_{r_i}\right)\right\};$$

- Step 5: Set $\beta = 0.5$;

- Step 6: Generate $\left(T_{p_i}, T_{r_i}\right)$: $T_{*_i} = \dfrac{T'_{*_i}}{\exp\left(\beta Z_{*_i}\right)}$ for $* = p, r$;
- Step 7: Generate $\left(C_{p_i}, C_{r_i}\right)$, $C_{*_i} \sim U\left(0, K\right)$ for $* = p, r$, where K is a fixed number;
- Step 8: We have $\left\{\left(X_{p_i}, X_{r_i}\right), \left(\delta_{p_i}, \delta_{r_i}\right), \left(Z_{p_i}, Z_{r_i}\right)\right\}$, where $X_{*_i} = T_{*_i} \wedge C_{*_i}$ and $\delta_{*_i} = I\left(T_{*_i} \le C_{*_i}\right)$ for $* = p, r$.

The procedure is repeated for $i = 1, ..., N$ for $N = 10000$.

6.3.2 Case-control data from the true population

Suppose that we generate $n << N$ families from the population with $n_1 = n/2$ families from the case families with $\left(X_{p_i}, \delta_{p_i} = 1\right)$. Then we match $n_0 = n_1 = n/2$ control families: $\left(X_{p_i}, \delta_{p_i} = 0\right)$ to the case families.

The procedure is stated as follows.

- Step 1: Randomly select n_1 probands from the case families and record their values of Z_{p_i} and data on his/her relative: $\left\{Z_{r_i}, \left(X_{r_i}, \delta_{r_i}\right)\right\}$;
- Step 2: Each case proband is matched with a control proband from the control group with $\left|X_1 - X_0\right| < 0.1$, where X_1 and X_0 are observed times for a case proband and a control proband respectively. And also record their values of Z_{r_i} and data on his/her relative: $\left\{Z_{r_i}, \left(X_{r_i}, \delta_{r_i}\right)\right\}$.

6.4 Analysis on Familial Data

We analyze the simulated familial case-control data by calculating the MLE of the parameters: β and α based on the likelihood function (5.13) and then check whether $\hat{\beta}$ and $\hat{\alpha}$ are close to their true value. We also check whether the probands' data can achieve equation (5.4). These results are summarized in Table 6.2. We observe that when Q is close

to $\exp\left(\beta^T Z\right)$, the estimation of $\hat{\beta}$ is also close to the true value, that means the conditional likelihood function can represent the first part of our decomposition of the likelihood function (5.13).

Table 6.2: Analysis of familial age-onset data
based on case-control studies

$N = 10000, n_1 = n_0 = 100, \text{Replications} = 100$

$\tau = 0.5, \beta = 0.5$

	$\overline{Q - \exp(\beta)}$	$\left(\bar{\hat{\beta}} - \beta\right) \times 10^3$	SE of $\hat{\beta}$	$\left(\bar{\hat{\tau}} - \tau\right) \times 10^3$	SE of $\hat{\tau}$
$p = 0.3$	-0.076737	23.221662	0.008890	-1.859457	0.004787
$p = 0.5$	0.023449	10.732187	0.009518	5.252649	0.004348
$p = 0.7$	0.018496	-6.228642	0.009259	-12.534723	0.004334

Chapter 7 Concluding Remarks

In the thesis, we study and review the literature on two regression models: logistic regression model and Cox proportional hazards model based on the prospective design and the case-control design to analyze the familial disease incidence data and familial age-onset data respectively. For inference, it has been shown that the data from a case-control design can be analyzed as if it is from a prospective design if some crucial properties like (3.7), (5.4) and reproducible properties hold. We perform simulations to examine these properties and check the properties of the parameter estimates. It allows us to see how the quality of generated data affects subsequent inference results.

The variables that we are interested in include disease incidence and age-onset data. For age-onset data, we treat an individual without the disease as being censored. But, these censored individuals may be actually censored or be non-susceptible. In such a situation, cure model can be adopted. This topic deserves future investigation.

References

[1] Bahadur, R. R. (1961). A Representation of the Joint Distribution of Responses to *n* Dichotomous Outcomes. *In Studies in Item Analysis and Prediction*, Stanford Mathematical Studies in the Social Sciences IV, Ed. H. Solomon, pp.158-168. Stanford University Press.

[2] Breslow, N. E. and Day N. E. (1980). *Statistical Methods in Cancer Research*, Vol. 1. Lyon: IARC Scientific Publication No. 32.

[3] Clayton, D. G. (1978). A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika* 65, 141-151.

[4] Li, H., Yang, P. and Schwartz, A. G. (1998). Analysis of Onset Data from Case-Control Family Studies. *Biometrics* 54, 1030-1039.

[5] Liang, K.-Y. (1987) Extended Mantel-Haenszel Estimating Procedure for Multivariate Logistic Regression Models. *Biometrics* 43, 289-299.

[6] Liang, K.-Y. and Beaty, T. H. (2000). Statistical Designs for Familial Aggregation. *Statistical Methods in Medical Research* 9, 543-562.

[7] Prentice, R. L. and Breslow, N. E. (1978). Retrospective Studies and Failure Time Models. *Biometrika* 65, 153-158.

[8] Prentice, R. L. and CAi, J. (1992). Covariance and Survivor Function Estimation Using Censored Multivariate Failure Time Data. *Biometrika* 79, 495-512.

[9] Prentice, R. L. and Pyke, R. (1979). Logistic Disease Incidence Models and Case-Control Studies. *Biometrika* 66, 403-411.

[10] Shih, J. H. and Chatterjee, N. (2002). Analysis of Survival Data from Case-Control Family Studies. *Biometrics* 58, 502-509.

[11] Sturmer, T. and Brenner, H. (2000) Potential Gain in Efficiency and Power to Detect Gene-Environment Interactions by Matching in Case-Control Studies. *Genetic*

Epidemiology 18, 63-80.

[12] Whittemore, A. S. (1995). Logistic Regression of Family Data from Case-Control Studies. *Biometrics* 82, 55-67.