Statistical Inference under Latent Class Models, with Application to Risk Assessment in Cancer Survivorship Studies

by

Huijing Wang

M.Sc., Simon Fraser University, 2012
B.Sc., Simon Fraser University, 2010
M.Sc., China Research Institute of Daily Chemical Industry, 2008
B.Eng., Beijing Technology and Business University, 2005

Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Actuarial Science and Statistics
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2015

Approval

Name:	**Huijing Wang**
Degree:	**Doctor of Philosophy (Statistics)**
Title:	***Statistical Inference under Latent Class Models, with Application to Risk Assessment in Cancer Survivorship Studies***
Examining Committee:	**Chair:** Dr. Tim B. Swartz Professor

Dr. X. Joan Hu
Senior Supervisor
Professor _____

Dr. John J. Spinelli
Co-Supervisor
Adjunct Professor _____

Ms. Mary L. McBride
Supervisor
Distinguished Scientist
Cancer Control Research
BC Cancer Agency _____

Dr. Lawrence C. McCandless
Internal Examiner
Associate Professor _____

Dr. Zhezhen Jin
External Examiner
Associate Professor
Department of Biostatistics
Columbia University _____

Date Defended: 12 November 2015

Abstract

Motivated by a cancer survivorship program, this PhD thesis aims to develop methodology for risk assessment, classification, and prediction.

We formulate the primary data collected from a cohort with two underlying categories, the at-risk and not-at-risk classes, using latent class models, and we conduct both cross-sectional and longitudinal analyses. We begin with a maximum pseudo-likelihood estimator (pseudo-MLE) as an alternative to the maximum likelihood estimator (MLE) under a mixture Poisson distribution with event counts. The pseudo-MLE utilizes supplementary information on the not-at-risk class from a different population. It reduces the computational intensity and potentially increases the estimation efficiency. To obtain statistical methods that are more robust than likelihood-based methods to distribution misspecification, we adapt the well-established generalized estimating equations (GEE) approach under the mean-variance model corresponding to the mixture Poisson distribution. The inherent computing and efficiency issues in the application of GEEs motivate two sets of extended GEEs, using the primary data supplemented by information from the second population alone or together with the available information on individuals in the cohort who are deemed to belong to the at-risk class. We derive asymptotic properties of the proposed pseudo-MLE and the estimators from the extended GEEs, and we estimate their variances by extended Huber sandwich estimators. We use simulation to examine the finite-sample properties of the estimators in terms of both efficiency and robustness. The simulation studies verify the consistency of the proposed parameter estimators and their variance estimators. They also show that the pseudo-MLE has efficiency comparable to that of the MLE, and the extended GEE estimators are robust to distribution misspecification while maintaining satisfactory efficiency. Further, we present an extension of the favourable extended GEE estimator to longitudinal settings by adjusting for within-subject correlation.

The proposed methodology is illustrated with physician claims from the cancer program. We fit different latent class models for the counts and costs of the physician visits by applying the proposed estimators. We use the parameter estimates to identify the risk of subsequent and ongoing problems arising from the subjects' initial cancer diagnoses. We perform risk classification and prediction using the fitted latent class models.

Keywords: Cross-Sectional Analysis; Event Count; Extended GEE Approach; Likelihood and Pseudo-Likelihood Estimation; Longitudinal Analysis; Medical Cost; Medical Insurance Information; Physician Claims; Risk Factor; Robust Variance Estimation

Dedication

To my parents, Mr. Wang Tong-Bao and Mrs. Chen Bao-Ling, with my gratitude and love.

Acknowledgements

I wish to express my deepest gratitude to my supervisor, Dr. Joan Hu, who inspired me to begin this journey, who taught me how to think deeply and broadly, and who always guided me in the right direction in both academia and real life. I admire her passion for statistics and her immense knowledge. I will always remember her patience, tolerance, and strictness throughout my graduate study. Our weekly meeting was always the time I learned the most. I could not have completed this work without her support, encouragement, and trust.

I am also grateful to all the professors in our department, who are both knowledgeable and kind. They inspire me to strive to become a better statistician. Moreover, I would like to thank the members of my thesis committee. My co-supervisor, Dr. John Spinelli, has been supportive throughout my graduate study: revising my papers, providing good advice, and inviting me to his lab meetings with other students. I am also grateful to the PI of the CAYACS program, Ms. Mary McBride, who provided me support, data access, and expert advice. Thanks to Dr. Lawrence McCandless for being willing to serve on my committee during his sabbatical, and to Dr. Tim Swartz for his kindness and willingness to chair my defence. Finally, I would like to thank Dr. Zhezhen Jin of the Columbia University for taking time from his busy schedule to serve as my external examiner.

My fellow graduate students, through their help and friendship, have enriched my life: thanks especially to Fei Wang, Dongdong Li, Yi Xiong, Shirin Golchi, Oksana Chkrebtti, Audrey Beliveau, Biljana Stojkova, Mike Grosskopf, Luyao Lin, Andrew Henrey, Jack Davis, Elena Szefer, Zheng Sun, Jean Shin, Joslin Goh, and Ruth Joy. It was always fun to discuss life and statistics with these super friendly and intelligent individuals; because of them I will cherish my graduate experience forever.

Last but not least, I would like to thank my boyfriend and best friend, Voleak Choeurng, who is also a statistician. With his love, belief and support, I never feel mentally alone during the entire PhD journey. He is always there for me during good and bad times.

Special Acknowledgement I gratefully acknowledge the BC Cancer Registry, BC Cancer Agency, BC Children's Hospital, and BC Ministry of Health, for allowing the use of their data for this thesis project, which was facilitated by Population Data BC. Funding for this project came from the Canadian Cancer Society (CCS) Research Institute and the

CCS BC & Yukon Division. Disclaimer statement: all inferences, opinions, and conclusions drawn in this thesis are those of the author, and do not reflect the opinions or policies of the Data Stewards.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The population of cancer survivors has been increasing rapidly as a result of advances in treatment. In particular, approximately 80% of Canadian children and adolescents diagnosed with cancer now survive five or more years from diagnosis (McBride *et al.*, 2010). They are often at risk of subsequent and ongoing health problems that are primarily treatment-related. The evaluation of strategies for survivors' long-term management requires risk assessments. The risk assessment of the later effects involves finding vulnerable and normal subgroups within the survivor cohort; identifying the risk factors associated with later effects and numerically evaluating the effects of these factors; and estimating the overall burden on the public health system for both subgroups. Risk assessments provide long-term information for policy makers regarding survivor care, assisting with the determination of ongoing support needs, the application of new knowledge, and the identification of new problems as survivors age and treatments change. This enables the best use of public resources to maximize the quality of life in cancer survivors. For individual survivors, risk assessments allow the patients and their families to make better long-term care plans. Physicians also need risk-assessment information so that they can provide appropriate care for different survivor groups.

1.1 Background and Motivation

We need studies on the cancer survivorship of young people to improve and evaluate their care and quality of life. McBride *et al.* (2010) recommended an ongoing and systematic follow-up of large cohorts of survivors. The standard methods of cohort research include tracking and contacting cohort members, obtaining consent, maintaining ongoing contact for long-term follow-up, and administering questionnaires. It is difficult to recruit a complete study group, retain the individuals, and collect comprehensive quality data. The use of geographically defined population databases and record-linkage methodology is a cost-

effective way to assemble cohorts and to collect detailed long-term data for a population of survivors, with minimal loss of contact.

1.1.1 General Background

Public health care is mandatory in Canada. According to the Federal Health Act 1984, the health-care insurance plan of each province must insure all services that are "medically necessary." Therefore, provincial, person-based, longitudinal health administrative databases are available for each province. BC started introduction of computerized medical services plan (MSP) database in January 1986. It includes outpatient physician-visit claim data for every residence. The BC Cancer Agency (BCCA) is a provincial government agency responsible for cancer treatment, research, and control in the province of BC. BCCA has maintained a population-based cancer registry since 1969. These population-based databases provide comprehensive information on cancer survivorship. The advantages of using Canadian health administrative databases are that they are intended to capture all medically necessary care and to eliminate the participant and recall bias of self-report studies (McBride *et al.*, 2011).

1.1.2 The CAYACS Program

The Childhood, Adolescent, Young Adult Cancer Survivorship (CAYACS) research program (*http://www.cayacs.ca*) was established at BCCA to carry out research into later effects and survivor care in multiple domains and to inform policy and practice (McBride *et al.*, 2010). Using existing population-based registries, administrative databases, and record-linkage methodology, researchers have conducted a series of epidemiological, clinical, and health-service studies relating to the survivorship issues of young cancer survivors.

The CAYACS program developed an ongoing population-based database for survivorship research by identifying a survivor cohort consisting of all the individuals in BC diagnosed with cancer before the age of 20 who survived more than five years and linking their records to the longitudinal administrative databases of outcomes (of Health, 2013). The survivor cohort was compared with the general BC population. A comparison sample that was 10 times the sample size of the survivor cohort was randomly selected from the population of BC, matched by birth-year and sex to the cohort; this was also linked to the administrative databases of outcomes. The use of health services, and especially physician care, is an indication of health status. Physician claims and demographic information were available from the MSP. We study the physician claims in this dissertation.

The CAYACS physician-claim data were collected from January 1, 1986 to December 31, 2006, i.e., from the start of the computerized MSP to the end of the study follow-up. The survivor cohort had $n = 1962$ subjects diagnosed with cancer from 1981 to 2000. The sample from the general population of BC contained $m = 19,620$ people. The physician

claims have a longitudinal data structure. Figure 1.1 illustrates the longitudinal claims of a hypothetical survivor. This individual was diagnosed with cancer on May 18, 1995. After five years, he/she became a cancer survivor and his/her physician-visit claims were followed from May 18, 2000 to the end of the study on December 31, 2006. During this period, he/she visited physicians (GPs or specialists) a total of 68 times. CAYACS recorded each visit date and the corresponding "fee for services" paid by the government; we will refer to this as the medical cost. For any subject i in the survivor cohort, suppose the follow-up period is $[L_i, R_i]$ where L_i is five years after the cancer diagnosis and R_i is December 31, 2006, the date of departure from BC, or the date of death, whichever occurs first. For any subject j

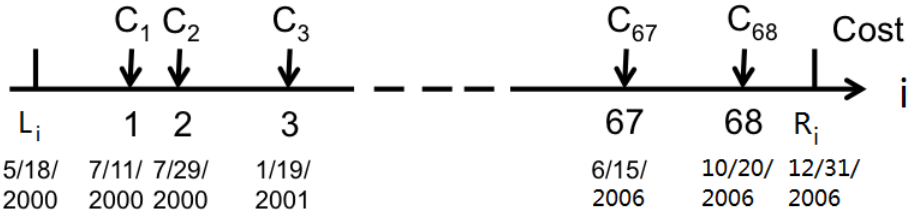

Figure 1.1: Physician visits and costs of a hypothetical CAYACS cancer survivor.

in the population sample, the longitudinal physician claims have the same structure as in Figure 1.1 except that, in the initial CAYACS design, L_j is the date of the fifth birthday or the date of arrival in BC, whichever occurs later. About 87% of the survivor cohort and 85% of the general population have an R_i/R_j date of December 31, 2006. The others either died or left BC before this date.

The age at study entry and aging potentially have a nonlinear relationship with the physician-visit pattern. Therefore, we decided to choose the L_j date for each individual in the general population according to that of a random survivor with the same birth-year and sex. This is a crucial step in our data clean-up. In Chapter 2, we compare the initial physician-claim data with our cleaned-up data. Thereafter, we use the cleaned-up data for our analyses. This approach to choosing the starting point of the comparison group may prove useful for other longitudinal studies.

The strengths of the CAYACS program are 1) it provides a population-based database with a matched general population as a comparison group, and 2) the data are longitudinal, so we can monitor changes in the outcomes over time. The scientific goals of the CAYACS physician-claims project are to evaluate the long-term physician visit frequency and the associated medical costs for young cancer survivors in BC, to identify risk factors and to conduct risk assessments for later effects, and to compare the results with those for the general population. A further goal is to assess and predict the long-term health service utilization and the continuity of care. In cancer survivorship research, it is also necessary to monitor changes over time.

1.1.3 Motivation

A recent CAYACS project on physician claims, summarized in McBride *et al.* (2011), provides a cross-sectional analysis of physician visits from 1998 to 2000 for a young survivor cohort of individuals diagnosed with cancer between 1970 and 1992. The analysis compared the physician-visit patterns of the cohort to those of the general population and identified factors associated with frequent physician visits. It showed that the demand for physician care among the survivors is considerably higher than that for a similar age and sex group in the population, and this need continues for many years after the diagnosis. About 97% of the survivors had at least one physician visit in the three-year period, and they saw physicians approximately twice as often as did the individuals in the general population. This study did not consider the cost of care.

The analysis of McBride *et al.* (2011) provides insight into the physician-visit patterns of the survivors and also raises further issues. For example, a comparison of the survivors as a whole with the general population may implicitly reveal whether a significant number of survivors are at risk of subsequent or ongoing problems. However, it does not explicitly relate this risk to the consequences of the original diagnosis. Moreover, the analysis indicates that the physician-visit frequency of the females in the cohort is significantly higher than that of the males. It is not clear whether this identifies sex as an important risk factor or simply reflects an overall pattern of physician visits; this pattern is also seen in the general population. Another finding is that older survivors visited physicians more frequently than younger survivors did. This is concluded from physician-visit counts in the three-year period for subjects of different ages. It would however be interesting to use the longitudinal data to study the aging effect over time, especially since a long follow-up period of up to 21 years is available.

Preliminary analyses indicated that while many cancer survivors visit physicians rather frequently, some survivors in the cohort have physician-visit patterns similar to those of the population. Many researchers believe that some survivors have the same health utilization as people without a cancer diagnosis. This motivated us to model the survivor cohort as a mixture of two latent classes: the groups *at-risk* and *not-at-risk* of later effects of the original diagnoses. A subject's membership is latent. The individuals in the at-risk class have a potentially higher frequency of physician visits, resulting in higher medical costs; while the individuals in the not-at-risk class have the same physician-visit patterns as the general population.

The formulation of two latent classes provides us with a convenient framework to study the features of physician visits due to the subsequent and ongoing treatment-related problems of survivors, and to numerically evaluate the survivors' risk of later effects. This may lead to a better risk assessment of the survivors. The survivor cohort can be classified into two strata. This allows us to evaluate separately the visit frequencies or medical costs of

the two latent classes. The model also leads to a natural comparison of the survivors in the at-risk class to the general population, if the not-at-risk class in the cohort is defined as the class that has the same physician-visit frequency or medical costs as the population.

This dissertation develops inference procedures for analysis under latent class models (LCMs) of physician claims from both the survivor cohort and the general population.

1.2 Literature Review

This dissertation adopts LCMs for cross-sectional and longitudinal data and develops estimating procedures based on likelihood and generalized estimating equations (GEEs). This section reviews these three topics.

1.2.1 Latent Class Models

Schlattmann (2009) explained why mixture models are particularly useful in medical applications. Patients are not alike, and finite mixture models can help to explain unobserved heterogeneity in the variables of interest. Schlattmann presented theory for finite mixture models and gave a large range of applications. A common method for handling overdispersed counts is negative binomial regression where the Poisson parameter is assumed to follow a Gamma distribution. This can be conceptualized as an LCM with an infinite number of classes where the latent variable follows a Gamma distribution. The choice of Gamma is for the sake of convenience.

LCMs were first introduced by Lazarsfeld and Henry (1968) in the field of social science. Goodman (1974) formalized the model and derived the maximum likelihood estimation (MLE) procedure. LCMs were developed for finding unobserved subgroups/latent classes in multivariate categorical data. The multivariate responses were correlated via the latent variable; in other words, they were conditionally independent. The MLE procedure is typically used. LCMs are a way to deal with unobserved heterogeneity that cannot be explained by the observed covariates; they can be considered a type of finite mixture model. They are also closely related to many other statistical models, such as other latent structure models, mixture-of-experts models, and random effects models (Lindsay *et al.*, 1991; Muthén, 2002; Vermunt and Magidson, 2003). They are analogous to cluster analysis models, sometimes with a different focus.

LCMs have many applications; see, for example, Magidson and Vermunt (2002), Pepe and Janes (2007), and Vermunt (2008). The popular zero-inflated Poisson (ZIP) model (e.g., Lambert, 1992; Hall, 2000) is a special case. LCMs have also been used in many health applications, such as disease diagnosis without a reference standard (van Smeden *et al.*, 2014), healthcare costs (Shih and Tai-Seale, 2012), genetic effects (Ekholm *et al.*, 2012), and medical treatment appropriateness (Uebersax, 1993). Other areas of application include finance (De Angelis, 2013), market research (Grisolía and Willis, 2012; Varki and

Chintagunta, 2004), economics (Boxall and Adamowicz, 2002), survey research (Reboussin *et al.*, 1999), and sociology (Formann and Kohlmann, 1998).

Originally, LCMs did not include any covariates in the latent class probability. These have been added more recently (Formann and Kohlmann, 1998; Ghosh *et al.*, 2011; Yang *et al.*, 2011; Wang *et al.*, 2014) to identify important predictors of latent classes. This development introduced the issue of variable selection in addition to the choice of the number of latent classes. Bayesian techniques for model diagnosis and model selection have been proposed (Garrett and Zeger, 2000; Ghosh *et al.*, 2011; Yang *et al.*, 2011), with corresponding estimation procedures by Bayesian methods (Yang *et al.*, 2011; Desantis *et al.*, 2012). Varki and Chintagunta (2004) applied an LCM to longitudinal data. Ever since Goodman (1974) derived the MLE procedure for LCMs, statistical inference for these models has relied on a distributional assumption for the responses. The computationally complex MLE procedures limited the use of LCMs. Therefore, Reboussin *et al.* (1999) developed a GEE procedure for a latent transition model with longitudinal multivariate categorical data.

The above examples all concern multivariate categorical data, mostly multivariate binary data. Wang *et al.* (2014) proposed an LCM for count data with risk assessment for the latent classes based on MLE and pseudo-MLE inference procedures. Researchers have often discussed the practical value of LCMs, and in particular how to determine the number of latent classes (Reboussin *et al.*, 1999). In this dissertation, we define two latent classes: the at-risk class and the not-at-risk class.

1.2.2 Cross-Sectional Analysis vs. Longitudinal Analysis

A longitudinal study observes a response variable repeatedly over time. In contrast, cross-sectional studies measure a single outcome for each individual. In both cases, the scientific objectives can usually be formulated as regression problems: the goal is to describe the dependence of the response on the explanatory variables and to evaluate the response for given values of the explanatory variables. While it is often possible to address the same scientific questions with a longitudinal study or cross-sectional study, a major advantage of longitudinal studies is the separation of what in the context of population studies are called *cohort* and *ageing* effects (Diggle *et al.*, 2002). In other words, longitudinal studies can distinguish changes over time within individuals (ageing effects) from differences among people at their baseline levels (cohort effects). For instance, consider the CAYACS physician-claim data. When we summarize the frequency of the physician visits during the follow-up period into cross-sectional counts and estimate how the counts change with time, we must assume that the mean of the counts changes proportionally to the observation length. With a longitudinal summary, this assumption is unnecessary since ageing effects can be estimated at any time point.

A study may focus on inference of the average response in the population. With one observation on each subject, we can model the population average of response. Not only can do this, longitudinal studies can borrow strength across time for the person of interest as well as across people. Thus, longitudinal studies can conduct subject-specific inference, and cross-sectional studies cannot. In longitudinal studies, each person can be thought of as serving as his or her own control. The subjects can usually be assumed to be independent of one another. Therefore, another merit of longitudinal studies is their ability to distinguish the degree of variation in responses across time for one person from the variation in responses among people. Moreover, longitudinal studies can deal with time-varying covariates.

The choice of the statistical model depends on the type of outcome. For example, continuous responses can be adequately described by linear regression models, perhaps after transformation. Generalized linear models (GLMs) will be adopted for binary or count responses. Any statistical methods for univariate data, e.g., general linear regression and GLMs, can be applied to cross-sectional analysis, depending on the outcome type and the objectives.

Longitudinal data require special statistical methods because multiple observations on one subject are likely to be correlated, and this correlation must be taken into account to draw valid statistical inference. One can view balanced longitudinal data as realizations of a multivariate variable and apply well-developed multivariate inferential methods. However, longitudinal studies typically have unbalanced designs, time-varying covariates, and other characteristics that make standard multivariate procedures unsuitable (Ware, 1985). The observations of each individual may not occur at specific time points but instead in a given interval, as for the CAYACS physician claims.

For cross-sectional data, only the dependence of the response on the covariates must be specified; there is no correlation. For repeated measurements there are three approaches to the modelling of correlation within a subject. Each approach models both the dependence of the response on the explanatory variables and the correlation among the responses for the same subject. The first approach is a type of marginal moment model, which specifies only the conditional mean and covariance structure of the longitudinal outcomes. This approach has the advantage of separately modelling the mean and covariance. Valid inferences about parameters in the mean model can sometimes be made even when an incorrect form is assumed for the covariance function (Liang and Zeger, 1986; Zeger and Liang, 1986). The second approach is the random effects model. It assumes that correlation arises among repeated responses because the regression coefficients vary across individuals. Here, the random variation is explicitly decomposed into between-subject and within-subject variation; see for example McCulloch *et al.* (2008). With a combination of fixed and random effects, these models are referred to as linear mixed models and generalized linear mixed models (GLMMs) depending on the type of the response. The relationship of these two approaches will be discussed in detail in Chapter 5. The third approach is a transition model (Ware

et al., 1988). It specifies a regression model for the conditional expectation as an explicit function of the covariates and of the past responses. It models the within-subject variation via Markov structures to account for the correlation of observations within the same subject. Transition models combine the assumptions about the dependence of the responses on the covariates and the correlation among multiple responses into a single equation.

Longitudinal analysis introduces theoretical and computational complications. We begin by studying LCMs for cross-sectional data, and we then extend the approach to longitudinal data.

1.2.3 Likelihood-Based vs. GEE-Based Estimation

This section explores two popular categories of inferential methods, likelihood-based and GEE-based estimation.

Likelihood inference is the foundational estimation approach of classical statistics; the distribution function of the response variable must be fully specified. The likelihood function has proved to be such a powerful tool for inference and it has been extended in many ways. For example, quasi-likelihood and quasi-score functions have been developed to overcome dispersion in GLMs, and various pseudo-likelihood functions have been proposed for more complicated models. Finding the MLE for the full likelihood function can be computationally intensive. For GLMs the MLE can be found by an iteratively reweighted least-squares fit. Another common strategy for finding the MLE is approximations from profile likelihood functions for component parameters of high dimension. For more examples, see Reid (2010). Pseudo-likelihood (also called composite likelihood) methods provide an approximation to the likelihood function. Varin *et al.* (2011) discuss the theory and applications of this approach.

Most inferential methods for longitudinal analysis fall into two categories: likelihood-based approaches and GEE approaches. The maximum likelihood and its variants are standard for GLMs and linear mixed models. For example, restricted maximum likelihood (REML) estimates are particularly important for estimating variance components in linear models with random effects. Finding the MLE for GLMMs involves numerically evaluating high-dimensional integrals. McCulloch (1997) proposed a Monte Carlo expectation-maximization (EM) algorithm and a Monte Carlo Newton–Raphson algorithm to obtain the MLE in GLMMs. Further details of likelihood-based estimation are provided in McCullagh and Nelder (1989) and McCulloch *et al.* (2008).

A somewhat different approach to the likelihood-based analysis of complex data is based on the quasi-likelihood of Wedderburn (1974). This approach starts by specifying parametric forms for the mean and variance of the response with an additional scale parameter for the variance function. Therefore, there is no need for a distributional assumption for the response. The estimating equation would be the score equation for a GLM with these first two moments, if such a model existed. Therefore, it is also called the quasi-score equation. The

theory of quasi-likelihood inference was developed by McCullagh (1983). Liang and Zeger extended the quasi-likelihood and GLMs to the analysis of longitudinal data (Liang and Zeger, 1986; Zeger and Liang, 1986), developing the GEE approach. They proposed using a "working covariance" function, and they showed that the estimates of the parameters in the mean were consistent even if the working covariance function was not correct. For longitudinal data, both subject-specific/mixed effects models and population-averaged/marginal models can be estimated by the GEE-based approach. When the random effects are assumed to follow a Gaussian distribution, simple relationships between the parameters in the mean function of marginal models and the fixed effects in mixed effects models are available (Zeger *et al.*, 1988).

1.3 Notation and Framework

We study LCMs for different types of response variables. This section presents the notation and framework used throughout the thesis. Specific notation will be introduced at the beginning of each chapter.

Consider a cohort of young cancer survivors, i.e., individuals diagnosed with cancer before the age of 20 who have survived for at least five years. A random sample of the general population was selected by matching birth-year and gender with the survivor cohort. Every subject has longitudinal physician-visit data as shown in Figure 1.1. Let $\mathbb{Y}$ be the summary response variable for the physician data of each individual. The data of $\mathbb{Y}$ can be binary (visit or no visit), discrete (visit counts), or continuous (medical costs). Moreover, $\mathbb{Y}$ can be a univariate variable ($\mathbb{Y} = Y$; cross-sectional summary for the entire follow-up period), a vector of variables ($\mathbb{Y} = \mathbf{Y}$; longitudinal summary under a discrete time scale, e.g., yearly) or a stochastic process ($\mathbb{Y} = Y(t)$ under a continuous time scale).

Let $\mathbf{Z}$ be a vector of p covariates/potential risk factors, which can be a combination of time-independent and time-dependent variables. $\mathbf{Z}$ may include both demographic and cancer-related variables for the survivor cohort; it includes only demographic variables for the general population.

To formulate the two strata with unobservable membership, corresponding to the at-risk and not-at-risk classes in the cohort, we introduce a latent binary variable η to indicate whether or not a subject belongs to the at-risk class.

LCM specification

- **Risk model for latent indicator η**:
  The conditional risk probability of η can be specified as a parametric functional form of $\mathbf{Z}$ up to α. Let $E(\eta|\mathbf{Z}) = P(\eta = 1|\mathbf{Z}) = p(\mathbf{Z};\alpha)$. In practice, a common form is a logit$(\cdot)$ transformation of $p(\mathbf{Z};\alpha)$.

- **Regression model for at-risk class ($\eta = 1$):**
  The regression model of response variable $\mathbb{Y}$ in the $\eta = 1$ class can be specified as a function of $\mathbf{Z}$ up to parameter β. Let $E(\mathbb{Y}|\eta = 1, \mathbf{Z}) = \mu_1(\mathbf{Z}; \beta)$. In practice, $l(\mu_1(\mathbf{Z}; \beta))$ is commonly considered, where the link function $l(\cdot)$ is the $\log(\cdot)$ function for count responses and the identity function for continuous responses.

- **Regression model for not-at-risk class ($\eta = 0$):**
  The regression model of response variable $\mathbb{Y}$ in the $\eta = 0$ class can be specified as a function of $\mathbf{Z}$ up to parameter θ. Let $E(\mathbb{Y}|\eta = 0, \mathbf{Z}) = \mu_0(\mathbf{Z}; \theta)$.

Many studies have a cohort of interest and independent information from other sources, e.g., a comparison group. We define the data from the survivor cohort to be primary data, denoted $\mathcal{P}$, and we define the information from the general population to be supplementary information, denoted $\mathcal{Q}$, where $\mathcal{P} \perp \mathcal{Q}$. Unless stated otherwise, for $\mathcal{Q}$ the start of the follow-up is chosen according to the survivor cohort. In Chapter 2, we compare analysis under $\mathcal{Q}$ to analysis under the original CAYACS general population data, referred to as $\mathcal{Q}_{\text{orig}}$. Moreover, we define $[\mathbb{Y}|\cdot]$ to be a conditional probability function of $\mathbb{Y}$ in general, which can be a pdf or pmf according to the type of data.

Our primary interest lies in estimating the parameters α, β, and θ in the LCMs, $p(\mathbf{Z}; \alpha)$, $\mu_1(\mathbf{Z}; \beta)$, and $\mu_0(\mathbf{Z}; \theta)$ using available data from $\mathcal{P} = \{(\mathbb{Y}_i, \mathbf{Z}_i) : i = 1, \ldots, n\}$ and $\mathcal{Q} = \{(\mathbb{Y}_i, \mathbf{Z}_i) : i = 1, \ldots, m\}$. For the CAYACS application, a consistent estimator of α gives a consistent estimator of the risk probability $p(\mathbf{Z}; \alpha)$ and then yields a measure of how likely the survivors with covariates of $\mathbf{Z}$ are to have later effects of the original diagnoses. The estimator of α can also be used to identify risk factors directly associated with the later effects. Consistent estimators of β and θ, on the other hand, can be used to identify factors associated with frequent visits and high medical costs of the at-risk class and infrequent visits and lower medical costs of the not-at-risk class. Moreover, comparisons of β and θ based on their estimates can detect differences in the visit frequency or medical costs between the two classes.

In addition to risk assessments, our scientific goals are (1) to compare the patterns of the physician visits and medical costs $\mathbb{Y}$ between the survivor cohort, especially the at-risk class, and the general population overall, as well as in different strata according to $\mathbf{Z}$; (2) to develop applications of LCMs to conduct risk classification and risk prediction for η within the cohort based on the available knowledge $\mathbb{Y}$, $\mathbf{Z}$.

1.4 Outline

This dissertation is motivated by the CAYACS data, and we illustrate the proposed LCMs and the associated inference procedures using this dataset. The statistical methodology

is not limited to this specific program and can be applied more broadly. The rest of this dissertation is organized as follows.

Chapter 2 summarizes the CAYACS physician claims into cross-sectional counts. Likelihood-based inference procedures, including an MLE and a maximum pseudo-likelihood estimation, are proposed for a mixture Poisson LCM. Continuing with cross-sectional counts, Chapter 3 introduces partially available membership information, bridged by likelihood-based estimations, and develops extended GEE inference procedures for the LCM, which is distribution-free. In Chapter 4, the CAYACS physician claims are summarized into longitudinal counts and costs. Conventional approaches for longitudinal data are adopted to analyze the survivor cohort and the general population separately and together. Chapter 5 extends the extended GEE methods of Chapter 3 to the longitudinal counts and continuous variables. We analyze the CAYACS physician claims in each chapter using the methodology described in that chapter. Chapter 6 provides a summary and a discussion of future work.

Chapter 2

Likelihood-Based Estimation with Cross-Sectional Counts

2.1 Introduction

In this chapter, we summarize the CAYACS physician claims into cross-sectional counts, i.e., a count of the physician visits during the entire follow-up period for each subject, for both the survivor cohort and the general population. Since the development of an MLE procedure for LCMs (Goodman, 1974), most LCMs have been estimated by the MLE. This is because one usually needs to specify the underlying probability model in a parametric form for each of the latent classes to avoid nonidentifiability problems. However, there are concerns about computational robustness when we implement likelihood-based procedures with LCMs (e.g., Hall and Shen, 2010). Moreover, the efficiency of the MLE will drop considerably because of the increased number of parameters. A model with two latent classes has almost three times as many parameters as a comparable marginal model. On the other hand, in many practical situations, information is readily available on one of the two latent classes. In the CAYACS case, the provincial medical insurance system collects rich information on the general population. These considerations led to a pseudo-MLE procedure, i.e., a way to estimate the model parameters using supplementary information. The procedure is potentially more efficient and robust, and it is relatively easy to implement.

This chapter formulates the CAYACS cross-sectional visit counts by an LCM. It also develops the associated likelihood-based estimating procedures, to evaluate the proportion of at-risk subjects in the cohort, assess the frequency of physician visits of the at-risk group, and identify the associated risk factors. We motivate and illustrate the proposed model and associated inference procedures using the CAYACS program.

The rest of this chapter is organized as follows. Section 2.2 introduces the notation and a mixture Poisson model for the physician visits of the cohort. In Section 2.3, we first present the MLE for the model parameters with the primary data and an application of the

EM algorithm to compute the MLE. We then propose a pseudo-MLE procedure using the additional information on the not-at-risk class, namely the physician-visit records for a collection of individuals selected from the general population. We establish the consistency and asymptotic normality of the pseudo-MLE and derive its asymptotic variance. Two variance estimators for the pseudo-MLE are presented. Section 2.4 reports the simulation studies of efficiency and robustness that we conducted to examine the finite-sample properties of the inference procedures and the two variance estimators. Section 2.5 presents an analysis of the CAYACS data via the proposed methodology. Section 2.6 provides concluding remarks.

2.2 Model Specification

In this chapter and the next, the response variable $\mathbb{Y} = Y$ will be a univariate count representing a subject's count of physician visits in the time period $(0, T]$, and $\mathbf{Z}$ will be his/her covariate vector, thus $[\mathbb{Y}|\cdot] = [Y|\cdot]$. The observation period varies from subject to subject.

As in Section 1.3, the latent variable η indicates the at-risk class. We denote $E(\eta|\mathbf{Z}) = P(\eta = 1|\mathbf{Z})$ by $p(\mathbf{Z};\alpha)$ and the conditional expectations of Y for the at-risk and not-at-risk classes by $E(Y|\eta, T, \mathbf{Z}) = \mu_\eta(T, \mathbf{Z})$ for $\eta = 1$ and 0, respectively. Thus, the expectation of Y conditional on T and $\mathbf{Z}$ is $E(Y|T, \mathbf{Z}) = \mu_1(T, \mathbf{Z};\beta)p(\mathbf{Z};\alpha) + \mu_0(T, \mathbf{Z};\theta)[1 - p(\mathbf{Z};\alpha)]$. This LCM is further specified as a finite mixture Poisson model as follows. We assume that the counts Y of the two classes follow a Poisson distribution with the conditional expectations $\mu_\eta(T, \mathbf{Z})$ for $\eta = 1, 0$. This formulation includes the popular ZIP model (e.g., Lambert, 1992; Hall, 2000) as a special case with $\mu_0(T, \mathbf{Z}) \equiv 0$.

The primary goal of this chapter is the estimation of the parameters α, β, and θ of $p(\mathbf{Z};\alpha)$, $\mu_1(T, \mathbf{Z};\beta)$, and $\mu_0(T, \mathbf{Z};\theta)$ using the cohort data $\mathcal{P} = \{(Y_i, T_i, \mathbf{Z}_i) : i = 1, \ldots, n\}$, a set of n independent and identically distributed realizations of $(Y, T, \mathbf{Z})$, and the general population data $\mathcal{Q} = \{(Y_i, T_i, \mathbf{Z}_i) : i = 1, \ldots, m\}$.

2.3 Likelihood-Based Inference Procedures

The event count Y conditional on T and $\mathbf{Z}$ follows the mixture Poisson distribution

$$[Y|T, \mathbf{Z};\alpha,\beta,\theta] = [Y|\eta = 1, T, \mathbf{Z};\beta]p(\mathbf{Z};\alpha) + [Y|\eta = 0, T, \mathbf{Z};\theta][1 - p(\mathbf{Z};\alpha)], \tag{2.1}$$

where $[Y|\eta, T, \mathbf{Z}]$ is the pmf of the Poisson distribution with a mean of $\mu_\eta(T, \mathbf{Z})$. We consider the MLE procedure based only on $\mathcal{P}$. The EM algorithm (Dempster *et al.*, 1977) is adapted to compute the MLE of the parameters. We then assume that there is a consistent estimator for θ in $\mu_0(T, \mathbf{Z};\theta)$, the event frequency model for the not-at-risk class. We propose a

pseudo-MLE procedure to estimate the parameters α in the risk model $p(\mathbf{Z};\alpha)$ and β in the regression model for the at-risk class, $\mu_1(T,\mathbf{Z};\beta)$.

2.3.1 Maximum Likelihood Estimation

Under the mixture Poisson model (2.1), the likelihood function of (α,β,θ) based on $\mathcal{P} = \{(Y_i,T_i,\mathbf{Z}_i) : i = 1,\ldots,n\}$ is

$$L(\alpha,\beta,\theta;\mathcal{P}) \propto \prod_{i=1}^{n} [Y_i|T_i,\mathbf{Z}_i;\alpha,\beta,\theta]. \tag{2.2}$$

The MLE of (α,β,θ) may be attained by directly maximizing (2.2) or its log-transformation. With the usual regularity conditions, the MLE $(\hat{\alpha},\hat{\beta},\hat{\theta})$ has asymptotic normality, i.e., as $n \to \infty$, $\sqrt{n}(\hat{\alpha}-\alpha,\hat{\beta}-\beta,\hat{\theta}-\theta)'$ converges in distribution to the multivariate normal distribution with mean zero and variance $FI(\alpha,\beta,\theta)^{-1}$. Here, $FI(\alpha,\beta,\theta)$ is the Fisher information matrix; it can be consistently estimated by $-n^{-1}\partial^2 \log L(\alpha,\beta,\theta;\mathcal{P})/\partial(\alpha,\beta,\theta)^2$ with the MLE plugged in.

Applying the EM algorithm gives us an alternative procedure for finding the MLE of (α,β,θ); this algorithm is potentially more intuitive and easier to implement. In particular, we consider the "complete data" $\{(Y_i,\eta_i,T_i,\mathbf{Z}_i) : i = 1,\ldots,n\}$. The log-likelihood function based on the complete data is

$$l(\alpha,\beta,\theta;\mathbf{Y},\boldsymbol{\eta}|\mathbf{T},\mathbf{Z}) = l_1(\alpha;\boldsymbol{\eta}|\mathbf{Z}) + l_2(\beta;\mathbf{Y},\boldsymbol{\eta}|\mathbf{T},\mathbf{Z}) + l_3(\theta;\mathbf{Y},\boldsymbol{\eta}|\mathbf{T},\mathbf{Z}), \tag{2.3}$$

where

$$l_1(\alpha;\boldsymbol{\eta}|\mathbf{Z}) = \sum_{i=1}^{n} \Big[\eta_i \log p(\mathbf{Z}_i;\alpha) + (1-\eta_i)\log\left[1-p(\mathbf{Z}_i;\alpha)\right]\Big], \tag{2.4}$$

$$l_2(\beta;\mathbf{Y},\boldsymbol{\eta}|\mathbf{T},\mathbf{Z}) = \sum_{i=1}^{n} \eta_i \log[Y_i|\eta_i = 1,T_i,\mathbf{Z}_i;\beta], \tag{2.5}$$

and

$$l_3(\theta;\mathbf{Y},\boldsymbol{\eta}|\mathbf{T},\mathbf{Z}) = \sum_{i=1}^{n} (1-\eta_i)\log[Y_i|\eta_i = 0,T_i,\mathbf{Z}_i;\theta]. \tag{2.6}$$

The EM algorithm iterates between an E-step and an M-step until convergence is achieved. The E-step estimates the unobserved η_i's with their conditional expectations using the current estimates of (α,β,θ). The M-step separately maximizes (2.4), (2.5), and (2.6) to update the estimates of (α,β,θ) using the most recent estimates of the η_i's. The computational advantage is obvious, since the complete-data log-likelihood is the sum of (2.4), (2.5), and (2.6), each of which depends on only one of the three parameter vectors.

Let the initial values be $\alpha^{(0)}$, $\beta^{(0)}$, and $\theta^{(0)}$. At the lth iteration ($l \geq 1$) of the algorithm, given the $(l-1)$th estimates $\alpha^{(l-1)}$, $\beta^{(l-1)}$, and $\theta^{(l-1)}$, the algorithm updates the estimates as follows:

E-Step. For $i = 1, \ldots, n$, calculate $\eta_i^{(l)} = E\{\eta_i | Y_i, T_i, \mathbf{Z}_i; \alpha^{(l-1)}, \beta^{(l-1)}, \theta^{(l-1)}\}$ by

$$E\{\eta | Y, T, \mathbf{Z}; \alpha, \beta, \theta\} = \frac{[Y|\eta = 1, T, \mathbf{Z}; \beta]p(\mathbf{Z}; \alpha)}{[Y|\eta = 1, T, \mathbf{Z}; \beta]p(\mathbf{Z}; \alpha) + [Y|\eta = 0, T, \mathbf{Z}; \theta][1 - p(\mathbf{Z}; \alpha)]}.$$

M-Step. Obtain $\alpha^{(l)}$, $\beta^{(l)}$, and $\theta^{(l)}$ by separately maximizing $l_1(\alpha; \boldsymbol{\eta}^{(l)} | \mathbf{Z})$, $l_2(\beta; \mathbf{Y}, \boldsymbol{\eta}^{(l)} | \mathbf{T}, \mathbf{Z})$, and $l_3(\theta; \mathbf{Y}, \boldsymbol{\eta}^{(l)} | \mathbf{T}, \mathbf{Z})$ in (2.3) with respect to α, β, θ, respectively.

Under mild regularity conditions, the **M-Step** is equivalent to solving each of the estimating equations:

$$\frac{\partial l_1(\alpha; \boldsymbol{\eta}^{(l)} | \mathbf{Z})}{\partial \alpha} = \sum_{i=1}^{n} [\eta_i^{(l)} - p(\mathbf{Z}_i; \alpha)] \frac{\partial p(\mathbf{Z}_i; \alpha)/\partial \alpha}{p(\mathbf{Z}_i; \alpha)[1 - p(\mathbf{Z}_i; \alpha)]} = 0, \tag{2.7}$$

$$\frac{\partial l_2(\beta; \mathbf{Y}, \boldsymbol{\eta}^{(l)} | \mathbf{T}, \mathbf{Z})}{\partial \beta} = \sum_{i=1}^{n} \eta_i^{(l)} [Y_i - \mu_1(T_i, \mathbf{Z}_i; \beta)] \frac{\partial \mu_1(T_i, \mathbf{Z}_i; \beta)/\partial \beta}{\mu_1(T_i, \mathbf{Z}_i; \beta)} = 0, \tag{2.8}$$

and

$$\frac{\partial l_3(\theta; \mathbf{Y}, \boldsymbol{\eta}^{(l)} | \mathbf{T}, \mathbf{Z})}{\partial \theta} = \sum_{i=1}^{n} (1 - \eta_i^{(l)}) [Y_i - \mu_0(T_i, \mathbf{Z}_i; \theta)] \frac{\partial \mu_0(T_i, \mathbf{Z}_i; \theta)/\partial \theta}{\mu_0(T_i, \mathbf{Z}_i; \theta)} = 0. \tag{2.9}$$

We can verify the conditions that ensure that the resulting sequence $\{(\alpha^{(l)}, \beta^{(l)}, \theta^{(l)}) : l = 1, 2, \ldots\}$ converges to the MLE $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ from $L(\alpha, \beta, \theta; \mathcal{P})$ in (2.2). This procedure for the ZIP model coincides with the estimation procedure presented by Hall and Shen (2010). We use their variation of the EM algorithm to adjust for outliers.

2.3.2 Maximum Pseudo-Likelihood Estimation

Suppose that a set of independent observations from the population, denoted $\mathcal{Q} = \{(Y_i, T_i, \mathbf{Z}_i) : i = 1, \ldots, m\}$, is available in addition to the data from the cohort $\mathcal{P}$. One may estimate (α, β, θ) with the likelihood function based on the primary data in combination with the supplementary information, which is the product of $L(\alpha, \beta, \theta; \mathcal{P})$ in (2.2) and $L(\theta; \mathcal{Q}) = \prod_{i=1}^{m} [Y_i | \eta_i \equiv 0, T_i, \mathbf{Z}_i; \theta]$. The efficiency of the MLE for the combined data is presumably higher than that of the MLE discussed in Section 2.3.1 based only on the primary data. However, the computational issues remain.

In many practical situations, the sample size m can be large relative to the size n of the primary data, and thus the supplementary data alone can lead to a consistent estimator of θ with sufficient efficiency. For a comparison between the cohort and the population, for example, the CAYACS program collected data from the population with a sample size (m) 10 times the size of the primary data (McBride *et al.*, 2011); m could be larger if necessary.

We propose the following pseudo-likelihood for estimating (α, β) using such an estimator of θ from the supplementary data, achieving an easily implementable estimation procedure with reasonably high efficiency, called the pseudo-MLE.

Assume that $\mathcal{Q}$ yields $\tilde{\theta}$, an estimator for the parameters in the regression model associated with the not-at-risk class, and $\sqrt{m}(\tilde{\theta} - \theta)$ converges in distribution to the normal distribution with zero mean and variance $AV_{\tilde{\theta}}(\theta)$ as $m \to \infty$. For example, the MLE of θ from the aforementioned $L(\theta; \mathcal{Q})$ satisfies the assumptions about $\tilde{\theta}$. It yields a pseudo-MLE of (α, β), denoted $(\tilde{\alpha}, \tilde{\beta})$, maximizing the pseudo-likelihood function, which is (2.2) with θ fixed at $\tilde{\theta}$, with respect to (α, β). This estimation procedure is considerably simpler than the procedure for computing the MLE $\hat{\alpha}$ and $\hat{\beta}$ jointly with $\hat{\theta}$ given in Section 2.3.1. The computational intensity is reduced by roughly one-third in general. The pseudo-MLE can be found by applying the adapted EM algorithm in Section 2.3.1 with $\theta = \tilde{\theta}$ throughout the algorithm.

Following the arguments in Gong and Samaniego (1981), we establish the consistency and asymptotic normality of $(\tilde{\alpha}, \tilde{\beta})$. Specifically, as $n \to \infty$ and $m \to \infty$, and assuming that $n/m \to k > 0$ and $\tilde{\theta}$ is independent of the primary data, $\sqrt{n}(\tilde{\alpha} - \alpha, \tilde{\beta} - \beta)'$ converges to the normal distribution with mean zero and variance

$$\mathrm{AV}_{(\tilde{\alpha},\tilde{\beta})}(\alpha, \beta, \theta) = I_{11}^{-1} + kI_{11}^{-1} I_{12} AV_{\tilde{\theta}}(\theta) I_{21} I_{11}^{-1}. \tag{2.10}$$

Let $FI(\alpha, \beta, \theta)$ be the Fisher information matrix of the likelihood function $L(\alpha, \beta, \theta; \mathcal{P})$ in (2.2). Suppose that $\tilde{\theta}$ is a consistent estimator from a set of supplementary data of size m and $\sqrt{m}(\tilde{\theta} - \theta) \xrightarrow{d} N\big(0, AV_{\tilde{\theta}}(\theta)\big)$ as $m \to \infty$. For example, the MLE of θ based on the supplementary data for the not-at-risk class satisfies the assumptions.

Assuming that the primary and supplementary data are independent, the conventional regularity conditions ensure the limiting joint distribution:

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \frac{\partial \log L(\alpha,\beta,\theta;\mathcal{P})}{\partial(\alpha,\beta,\theta)} \\ \sqrt{m}(\tilde{\theta} - \theta) \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} FI(\alpha, \beta, \theta) & 0 \\ 0 & AV_{\tilde{\theta}}(\theta) \end{pmatrix}\right). \tag{2.11}$$

Partition $FI(\alpha, \beta, \theta)$ as follows:

$$\begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = -E \begin{bmatrix} \dfrac{\partial^2 \log[Y|T, \mathbf{Z}; \alpha, \beta, \theta]}{\partial(\alpha, \beta)^2} & \dfrac{\partial^2 \log[Y|T, \mathbf{Z}; \alpha, \beta, \theta]}{\partial(\alpha, \beta)\partial\theta} \\ \dfrac{\partial^2 \log[Y|T, \mathbf{Z}; \alpha, \beta, \theta]}{\partial\theta\partial(\alpha, \beta)} & \dfrac{\partial^2 \log[Y|T, \mathbf{Z}; \alpha, \beta, \theta]}{\partial\theta^2} \end{bmatrix}. \tag{2.12}$$

Recall that the pseudo-MLE $(\tilde{\alpha}, \tilde{\beta})$ maximizes $L(\alpha, \beta, \tilde{\theta}; \mathcal{P})$ and is the solution to $\partial \log L(\alpha, \beta, \tilde{\theta}; \mathcal{P})/\partial(\alpha, \beta) = 0$ almost surely. Thus, the first-order Taylor expansion of the

partial derivative of the pseudo log-likelihood function yields

$$\frac{\partial \log L(\alpha,\beta,\tilde{\theta};\mathcal{P})}{\partial(\alpha,\beta)} + (\tilde{\alpha}-\alpha,\tilde{\beta}-\beta)\frac{\partial^2 \log L(\alpha,\beta,\tilde{\theta};\mathcal{P})}{\partial(\alpha,\beta)^2} \approx 0.$$

Given the continuity of $\partial \log L(\alpha,\beta,\theta;\mathcal{P})/\partial(\alpha,\beta)$ and $\partial^2 \log L(\alpha,\beta,\theta;\mathcal{P})/\partial(\alpha,\beta)^2$ with respect to θ, as $n \to \infty$ and $m \to \infty$, we can show that $-n^{-1}\partial^2 \log L(\alpha,\beta,\tilde{\theta};\mathcal{P})/\partial(\alpha,\beta)^2$ converges to I_{11} a.s. and $n^{-1/2}\partial \log L(\alpha,\beta,\tilde{\theta};\mathcal{P})/\partial(\alpha,\beta)$ converges to $N(0, I_{11}^*)$ in distribution. Here $I_{11}^* = I_{11} + kI_{12}AV_{\tilde{\theta}}(\theta)I_{21}$, if $n/m \to k$. Following the standard arguments for the consistency and asymptotic normality of an MLE, we can then establish the consistency of the pseudo-MLE $(\tilde{\alpha},\tilde{\beta})$ and its asymptotic normality with the variance given in (2.10).

The expression for the asymptotic variance in (2.10) shows that the efficiency of the pseudo-MLE $(\tilde{\alpha},\tilde{\beta})$ can be close to that of the MLE of (α,β) with a known θ when either k or $AV_{\tilde{\theta}}$ is small. This indicates that the efficiency of the pseudo-MLE $(\tilde{\alpha},\tilde{\beta})$ may exceed the efficiency of the MLE of (α,β) jointly obtained with the MLE of θ using the primary data only.

Note that the corresponding blocks of $-n^{-1}\partial^2 \log L(\alpha,\beta,\theta;\mathcal{P})/\partial(\alpha,\beta,\theta)^2$ are consistent estimators for the matrices I_{11}, I_{12}, and I_{21} with the pseudo-MLE plugged in. They, together with a consistent estimator of $AV_{\tilde{\theta}}(\theta)$, naturally form a consistent estimator of $AV_{(\tilde{\alpha},\tilde{\beta})}(\alpha,\beta,\theta)$. The derivation of (2.10) and the aforementioned consistent variance estimator require the underlying model specification. In practice, a more robust variance estimator is often preferable, just as the Huber sandwich variance estimator for the variance of the MLE is preferred to anticipate possible model misspecification (Huber, 1967). This consideration leads us to estimate I_{11}^{-1}, the first term in (2.10), with the corresponding Huber sandwich estimator, which results in an extended Huber sandwich estimator.

Following the partition of (2.12), denote the blocks of $\widehat{FI}(\alpha,\beta,\theta) = -\frac{1}{n}\partial^2 \log L(\alpha,\beta,\theta;\mathcal{P})/\partial(\alpha,\beta,\theta)^2$ by $\hat{I}_{11}(\alpha,\beta,\theta)$, $\hat{I}_{12}(\alpha,\beta,\theta)$, $\hat{I}_{21}(\alpha,\beta,\theta)$, and $\hat{I}_{22}(\alpha,\beta,\theta)$. The following is a consistent estimator for $AV_{\tilde{\alpha},\tilde{\beta}}(\alpha,\beta,\theta)$ in (2.10):

$$\hat{I}_{11}^{-1}(\alpha,\beta,\theta) + \frac{n}{m}\hat{I}_{11}^{-1}(\alpha,\beta,\theta)\hat{I}_{12}(\alpha,\beta,\theta)\widehat{AV}_{\tilde{\theta}}\hat{I}_{21}(\alpha,\beta,\theta)\hat{I}_{11}^{-1}(\alpha,\beta,\theta) \tag{2.13}$$

with (α,β,θ) substituted by $(\tilde{\alpha},\tilde{\beta},\tilde{\theta})$ and $\widehat{AV}_{\tilde{\theta}}$ a consistent estimator of $AV_{\tilde{\theta}}(\theta)$.

Partition the variance matrix of the score function based on the primary data as follows:

$$\mathrm{Var}\begin{bmatrix} \dfrac{\partial \log[Y|T,\mathbf{Z};\alpha,\beta,\theta]}{\partial(\alpha,\beta)} \\ \dfrac{\partial \log[Y|T,\mathbf{Z};\alpha,\beta,\theta]}{\partial\theta} \end{bmatrix} = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}.$$

Denote the block in the sample moment estimator for the variance matrix corresponding to Π_{11} by $\hat{\Pi}_{11}(\alpha,\beta,\theta)$. It yields a robust variance estimator for $AV_{\tilde{\alpha},\tilde{\beta}}(\alpha,\beta,\theta)$ if $\hat{I}_{11}^{-1}(\alpha,\beta,\theta)$

in (2.13) is replaced by

$$\hat{I}_{11}^{-1}(\alpha,\beta,\theta)\widehat{\Pi}_{11}(\alpha,\beta,\theta)\hat{I}_{11}^{-1}(\alpha,\beta,\theta). \tag{2.14}$$

In fact, (2.14) is the Huber sandwich variance estimator for the MLE of (α,β) with a known θ.

2.4 Simulation Study

We conducted simulation studies to examine the finite-sample properties of the MLE and pseudo-MLE in terms of efficiency and robustness to model misspecification. The numerical studies throughout this dissertation were carried out using the R package for statistical computing (http://www.r-project.org).

This chapter adopts the common parametric specifications for $p(\mathbf{Z})$ and $\mu_\eta(T,\mathbf{Z})$, the logistic and loglinear regression models:

$$\text{logit}\{p(\mathbf{Z};\alpha)\} = \alpha_0 + \alpha_1^{'}\mathbf{Z}, \tag{2.15}$$

and

$$\log\{\mu_1(T,\mathbf{Z};\beta)\} = \beta_0 + \beta_1^{'}\mathbf{Z} + \beta_2\log T; \quad \log\{\mu_0(T,\mathbf{Z};\theta)\} = \theta_0 + \theta_1^{'}\mathbf{Z} + \theta_2\log T. \tag{2.16}$$

With slight modifications, our estimation procedures and discussions are applicable to other parametric specifications.

2.4.1 Description of Data Generation

We simulated n independent individuals from the two latent classes: the at-risk and not-at-risk groups. We used the outcomes reported by McBride *et al.* (2011) to choose the parameter values for the data generation in the simulations. Specifically, we simulated two potential risk factors: a binary variable *sex* as the indicator of a male subject, and a continuous variable (*age*) as the standardized age of an individual at the beginning of the study. These two risk factors together with the latent indicator η of the at-risk class and the individual observation time T were generated as follows. For the ith individual in the study,

(i) $\text{sex}_i \sim \text{Bin}(1,\frac{1}{2})$, the Bernoulli distribution with a success probability of $\frac{1}{2}$;

(ii) $\text{age}_i \sim \text{Beta}(0.7,0.8)$, the Beta distribution with the parameter values chosen to follow the distribution of the standardized age variable in the CAYACS program;

(iii) $\eta_i \sim \text{Bin}(1,p_i)$, where $\text{logit}(p_i) = 1 - \text{sex}_i - 0.8\text{age}_i$;

(iv) $T_i \sim \text{Beta}(2,1)$.

The event counts Y_i were then generated in the following two settings, designed to assess the efficiency and robustness of the estimators.

Simulation Setting 1: Efficiency Study. Conditional on $(\eta_i, T_i, \text{sex}_i, \text{age}_i)$, Y_i was generated from Poisson distributions as follows:

(i) for $\eta_i = 1$, the mean is $\mu_1(T_i, \text{sex}_i, \text{age}_i) = T_i \exp(1.8 - 0.6\text{sex}_i - 0.5\text{age}_i)$;

(ii) for $\eta_i = 0$, the mean is $\mu_0(T_i, \text{sex}_i, \text{age}_i) = T_i \exp(0.5 - 0.3\text{sex}_i - 0.25\text{age}_i)$.

Simulation Setting 2: Robustness Study. For individual i, ξ_i was generated from the gamma distribution with mean 1 and variance γ_i: $\xi_i \sim Gamma(1, \gamma_i)$. Conditional on $(\eta_i, T_i, \text{sex}_i, \text{age}_i, \xi_i)$, Y_i was generated from a Poisson distribution with mean $\xi_i \mu_{\eta_i}(T_i, \text{sex}_i, \text{age}_i)$, where $\mu_{\eta_i}(T_i, \text{sex}_i, \text{age}_i)$ was the same as in the efficiency study for $\eta_i = 1$ or 0. Note that if $\gamma_i > 0$, the variance of the simulated event count Y_i conditional on $(\eta_i, T_i, \text{sex}_i, \text{age}_i)$ is $(1 + \gamma_i)\mu_{\eta_i}(T_i, \text{sex}_i, \text{age}_i)$. Three model-misspecification scenarios were simulated:

Case (i) $\gamma_i = \gamma > 0$ regardless of η_i, i.e., the underlying distributions of both classes were misspecified;

Case (ii) $\gamma_i = \gamma > 0$ if $\eta_i = 1$ and $\gamma_i = 0$ if $\eta_i = 0$, i.e., only the at-risk class was misspecified;

Case (iii) $\gamma_i = 0$ if $\eta_i = 1$ and $\gamma_i = \gamma > 0$ if $\eta_i = 0$, i.e., only the not-at-risk class was misspecified.

We chose the parameter γ to be $\frac{1}{2}$, 1, or 2 to simulate mild, medium, or severe overdispersed counts, respectively.

We formed the observed (primary) data as $\{(Y_i, T_i, \text{sex}_i, \text{age}_i) : i = 1, \ldots, n\}$ in the simulations. The supplementary information was generated independently as realizations of $(Y, T, \text{sex}, \text{age})$ from a group of m independent individuals with the same distribution as the not-at-risk class in each of the simulation settings.

2.4.2 Simulation Outcomes

Each of the experimental settings described in Section 2.4.1 was simulated 250 times. For each simulated data set, we evaluated both the MLE and the pseudo-MLE for the parameters in the LCM, i.e., the mixture Poisson model in Section 2. We also evaluated the standard error estimators of the MLE and the pseudo-MLE based on the conventional variance estimator for the MLE and the Huber sandwich variance estimator, and the two variance estimators for the pseudo-MLE given in Section 2.3.2. We computed the evaluations of $\tilde{\theta}$ used in the pseudo-MLE procedure and the estimates of the parameters θ in

the frequency model for the not-at-risk class based on the supplementary information using the R function *glm*. We implemented both the MLE and pseudo-MLE procedures by (a) maximizing the observed data likelihood and the pseudo-likelihood functions via an R optimization function and (b) applying the EM algorithm described in Section 2.3. The resulting estimates from (a) were close to the estimates from (b). The estimates from the EM algorithm are discussed below.

Table 2.1: Simulation Outcomes: Efficiency Study

(Primary data $n = 500$; 250 repetitions)											
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	1	-1	-0.8	1.8	-0.6	-0.5	1	0.5	-0.3	-0.25	1
	MLE of (α, β, θ)										
$\text{sm}^\dagger$	1.000	-1.026	-0.770	1.790	-0.599	-0.500	1.006	0.487	-0.295	-0.261	1.009
$\text{sse}^\ddagger$	0.231	0.291	0.397	0.081	0.053	0.070	0.056	0.192	0.090	0.142	0.129
$\text{sm}^\dagger_{\text{se}}$	0.243	0.302	0.420	0.084	0.058	0.069	0.060	0.197	0.094	0.141	0.133
$\text{sm}^\dagger_{\text{sw.se}}$	0.248	0.312	0.433	0.083	0.057	0.068	0.058	0.198	0.096	0.142	0.135
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							MLE of θ			
sm	1.003	-1.011	-0.775	1.791	-0.602	-0.499	1.005	0.503	-0.301	-0.251	1.000
sse	0.220	0.266	0.350	0.080	0.048	0.066	0.055	0.030	0.013	0.020	0.021
sm_{se}	0.231	0.252	0.382	0.083	0.051	0.066	0.059	0.029	0.014	0.022	0.021
$\text{sm}_{\text{sw.se}}$	0.232	0.252	0.383	0.081	0.050	0.064	0.058	0.029	0.014	0.022	0.021

†The sample means of the parameter estimates (sm), the conventional standard error estimates ($\text{sm}^\dagger_{\text{se}}$), and the sandwich standard error estimates ($\text{sm}^\dagger_{\text{sw.se}}$).

‡The sample standard errors (sse) of the parameter estimates.

Table 2.1 presents a summary of the parameter estimates and the asymptotic standard error estimates in the efficiency study with $n = 500$ and $m = 5000$ based on 250 replicates. The sample means (sm) of all the parameter estimators are close to the corresponding true values of the parameters: the relative differences range from 0% to 3.7%. This verifies the consistency of both the MLE and the pseudo-MLE. The sample standard errors (sse) of the pseudo-MLE estimators overall appear smaller than those of the MLE estimators. That is, the supplementary information along with the smaller number of parameters to be estimated may compensate for the pseudo-MLE's potential loss of efficiency, leading to better efficiency than that for the evaluable MLE with the primary data. Table 2.1 also presents the sm of the two standard error estimators, the conventional and sandwich estimators, for both the MLE and the pseudo-MLE. We see that sm_{se} (or sm_{pse}) and $\text{sm}_{\text{se.sw}}$ (or $\text{sm}_{\text{pse.sw}}$) are essentially the same. They are close to the corresponding sse of the estimators, with the absolute differences ranging from 0.2% to 3.6%. This shows that the accuracy of both standard error estimators is satisfactory in practice.

We considered additional simulation settings in the efficiency study. For comparison, we evaluated the MLE of (α, β, θ) based on the primary data combined with the realizations of the latent indicator η. The sm and sse were close to those associated with the pseudo-MLE. To further explore the contribution of the supplementary information, we evaluated two other sets of estimators of α and β: the MLEs with θ fixed at the true value and the pseudo-MLEs with θ estimated based on the supplementary information with size $m = 500$. As anticipated, the sse of the MLEs with the true θ were smaller than the sse of the MLEs with θ jointly estimated, and the sse of the pseudo-MLEs for $m = 500$ were slightly larger than those for $m = 5000$, which were close to those for the MLEs with the true θ. We also evaluated the estimators with the size of the observed primary data set to $n = 100$. The findings were the same.

Regardless of the value of the overdispersion parameter γ, the simulation outcomes in the three cases of the robustness study show that the MLE is sensitive to model misspecification overall, but the robustness of the pseudo-MLE varies. The sm for the MLE reveal some serious biases in the simulated situations, especially for the regression coefficients in the risk model. The differences of the sm for the pseudo-MLE from the true parameter values are considerably smaller. Particularly in case (iii), which simulated situations where only the underlying frequency model for the not-at-risk class (i.e., the class where $\eta = 0$) was misspecified, the pseudo-MLE estimates are basically unbiased. In all three cases, the sm of the standard error estimates based on the conventional variance estimator for MLE have discrepancies compared with the sse associated with both the MLE and the pseudo-MLE estimators. The sm of the corresponding sandwich standard error estimator, on the other hand, is close to the sse. This verifies the robustness of the sandwich estimator. We summarize the simulation results of the three cases of the robustness study with $m = 5000$, $\gamma = 1$ in Table 2.2.

We conducted another simulation study to explore the difference in robustness between the MLE and pseudo-MLE in situations similar to case (iii). We substituted the mixed Poisson model with a mixture of two Poisson models for the class $\eta = 0$: the mean of one component was the same as the mean of the class $\eta = 0$ in the efficiency study, and the mean of the second component was close to the mean of the group $\eta = 1$. We varied the proportion of the second component in the mixture from 10% to 80%, and observed that the corresponding bias in the MLE of (α, β) changed from minor to major, while the pseudo-MLE of the parameters remained close to the true values. This further suggests the benefit of using supplementary information. See Table 2.3 for a summary of the simulation outcomes.

Table 2.2: Simulation Outcomes: Robustness Study

(Primary data $n = 500$; 250 repetitions)											
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	1	-1	-0.8	1.8	-0.6	-0.5	1	0.5	-0.3	-0.25	1
	Case (i) $\gamma = 1$										
	MLE of (α, β, θ)										
$\text{sm}^{\dagger}$	-0.864	-0.114	-0.129	2.510	-0.707	-0.528	0.950	0.614	-0.692	-0.537	1.012
$\text{sse}^{\ddagger}$	0.385	0.416	0.648	0.326	0.214	0.376	0.220	0.346	0.213	0.379	0.236
$\text{sm}^{\dagger}_{\text{se}}$	0.216	0.218	0.347	0.041	0.018	0.025	0.029	0.087	0.040	0.055	0.061
$\text{sm}^{\dagger}_{\text{sw.se}}$	0.328	0.411	0.612	0.307	0.207	0.308	0.217	0.339	0.226	0.336	0.232
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							GEE estimate of θ			
sm	-0.704	-0.544	-0.450	2.453	-0.557	-0.425	0.952	0.498	-0.299	-0.252	1.002
sse	0.282	0.299	0.449	0.274	0.165	0.287	0.195	0.062	0.033	0.052	0.044
sm_{se}	0.220	0.234	0.369	0.041	0.020	0.026	0.029	0.067	0.032	0.050	0.047
$\text{sm}_{\text{sw.se}}$	0.266	0.288	0.459	0.259	0.156	0.254	0.191	0.067	0.032	0.050	0.047
	Case (ii) $\gamma = 1$										
	MLE of (α, β, θ)										
sm	-0.877	-0.787	-0.645	2.499	-0.496	-0.389	0.963	0.703	-0.430	-0.343	0.996
sse	0.301	0.353	0.526	0.333	0.189	0.296	0.244	0.199	0.114	0.177	0.139
sm_{se}	0.236	0.262	0.409	0.046	0.024	0.029	0.032	0.081	0.035	0.053	0.057
$\text{sm}_{\text{sw.se}}$	0.301	0.346	0.546	0.297	0.180	0.285	0.219	0.200	0.110	0.174	0.142
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							GEE estimate of θ			
sm	-0.723	-0.880	-0.704	2.430	-0.464	-0.372	0.973	0.499	-0.301	-0.249	1.000
sse	0.256	0.303	0.445	0.302	0.170	0.256	0.220	0.030	0.014	0.022	0.020
sm_{se}	0.229	0.256	0.398	0.043	0.023	0.027	0.030	0.029	0.014	0.022	0.021
$\text{sm}_{\text{sw.se}}$	0.267	0.297	0.472	0.272	0.163	0.264	0.201	0.029	0.014	0.022	0.021
	Case (iii) $\gamma = 1$										
	MLE of (α, β, θ)										
sm	1.305	-0.623	-0.462	1.762	-0.606	-0.499	1.003	-0.384	-0.513	-0.418	1.140
sse	0.253	0.247	0.381	0.085	0.051	0.073	0.059	0.575	0.259	0.420	0.422
sm_{se}	0.231	0.221	0.351	0.066	0.031	0.047	0.047	0.326	0.122	0.188	0.226
$\text{sm}_{\text{sw.se}}$	0.244	0.250	0.394	0.083	0.051	0.072	0.060	0.526	0.257	0.407	0.370
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							GEE estimate of θ			
sm	1.190	-1.119	-0.845	1.800	-0.507	-0.436	0.980	0.494	-0.297	-0.252	1.003
sse	0.237	0.225	0.358	0.080	0.044	0.066	0.056	0.061	0.035	0.048	0.041
sm_{se}	0.243	0.242	0.382	0.079	0.042	0.059	0.056	0.067	0.032	0.050	0.047
$\text{sm}_{\text{sw.se}}$	0.230	0.229	0.361	0.078	0.045	0.062	0.056	0.067	0.032	0.050	0.047

†The sample means of the parameter estimates (sm), the conventional standard error estimates ($\text{sm}^{\dagger}_{\text{se}}$), and the sandwich standard error estimates ($\text{sm}^{\dagger}_{\text{sw.se}}$).

‡The sample standard errors (sse) of the parameter estimates.

Table 2.3: Simulation Outcomes: Additional Robustness Study

(Primary data $n = 500$; 250 repetitions)											
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	1	-1	-0.8	1.8	-0.6	-0.5	1	0.5	-0.3	-0.25	1
	Case (iii.A1): Two components in $\eta = 0$ in the ratio $9:1$										
	MLE of (α, β, θ)										
$\text{sm}^{\dagger}$	1.039	-0.975	-0.783	1.796	-0.605	-0.498	1.000	0.472	-0.303	-0.242	1.003
$\text{sse}^{\ddagger}$	0.257	0.297	0.438	0.077	0.053	0.062	0.055	0.195	0.099	0.140	0.140
$\text{sm}^{\dagger}_{\text{se}}$	0.245	0.296	0.421	0.083	0.055	0.068	0.059	0.200	0.096	0.143	0.135
$\text{sm}^{\dagger}_{\text{sw.se}}$	0.252	0.311	0.440	0.082	0.057	0.069	0.058	0.206	0.101	0.151	0.139
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							GEE estimate of θ			
sm	1.030	-1.002	-0.795	1.799	-0.601	-0.495	0.998	0.514	-0.306	-0.248	1.000
sse	0.245	0.250	0.406	0.077	0.046	0.060	0.055	0.031	0.014	0.023	0.024
sm_{se}	0.233	0.251	0.383	0.083	0.051	0.066	0.059	0.031	0.015	0.023	0.022
$\text{sm}_{\text{sw.se}}$	0.234	0.251	0.384	0.081	0.049	0.064	0.058	0.031	0.015	0.023	0.022
	Case (iii.A2): Two components in $\eta = 0$ in the ratio $5:5$										
	MLE of (α, β, θ)										
sm	1.217	-0.744	-0.552	1.775	-0.628	-0.524	1.000	0.335	-0.338	-0.318	1.034
sse	0.285	0.353	0.467	0.091	0.055	0.077	0.059	0.316	0.162	0.264	0.218
sm_{se}	0.256	0.288	0.426	0.075	0.046	0.059	0.053	0.240	0.110	0.164	0.163
$\text{sm}_{\text{sw.se}}$	0.299	0.362	0.526	0.083	0.059	0.077	0.058	0.309	0.175	0.268	0.208
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							GEE estimate of θ			
sm	1.147	-1.020	-0.827	1.792	-0.585	-0.490	0.993	0.621	-0.339	-0.269	0.997
sse	0.244	0.272	0.362	0.089	0.045	0.065	0.060	0.034	0.018	0.028	0.024
sm_{se}	0.245	0.261	0.401	0.081	0.050	0.065	0.058	0.037	0.018	0.028	0.026
$\text{sm}_{\text{sw.se}}$	0.245	0.257	0.398	0.081	0.046	0.064	0.058	0.037	0.018	0.028	0.026
	Case (iii.A3): Two components in $\eta = 0$ in the ratio $2:8$										
	MLE of (α, β, θ)										
sm	1.408	-0.006	0.142	1.759	-0.705	-0.586	1.002	-0.034	-0.919	-0.941	1.309
sse	0.516	0.573	0.842	0.098	0.068	0.093	0.062	1.161	0.659	0.968	0.778
sm_{se}	0.289	0.304	0.474	0.065	0.032	0.048	0.046	0.384	0.199	0.252	0.268
$\text{sm}_{\text{sw.se}}$	0.484	0.621	0.975	0.101	0.073	0.113	0.064	1.084	0.622	0.970	0.674
								Supplementary data $m = 5000$			
	Pseudo-MLE of (α, β)							GEE estimate of θ			
sm	1.170	-0.990	-0.773	1.792	-0.591	-0.497	0.992	0.795	-0.385	-0.292	1.000
sse	0.271	0.277	0.430	0.088	0.051	0.071	0.060	0.036	0.017	0.029	0.024
sm_{se}	0.267	0.295	0.447	0.083	0.053	0.067	0.059	0.037	0.018	0.027	0.026
$\text{sm}_{\text{sw.se}}$	0.265	0.288	0.443	0.081	0.048	0.067	0.058	0.037	0.018	0.027	0.026

†The sample means of the parameter estimates (sm), the conventional standard error estimates ($\text{sm}^{\dagger}_{\text{se}}$), and the sandwich standard error estimates ($\text{sm}^{\dagger}_{\text{sw.se}}$).

‡The sample standard errors (sse) of the parameter estimates.

2.5 Analysis I of CAYACS Physician Claims

This section presents an analysis of the CAYACS visit records summarized in cross-sectional counts using the methodology described in this chapter. We analyze the cohort data $\mathcal{P}$ and the population data $\mathcal{Q}$. The pseudo-MLE result based on $\mathcal{P}$ and $\mathcal{Q}_{\text{orig}}$ from Wang *et al.* (2014) is listed in Table 2.6 for comparison and discussed in Section 2.6. The population data $\mathcal{Q}_{\text{orig}}$ that does not match the start time to the survivor cohort is not used elsewhere in this thesis.

2.5.1 Preliminary Analysis

To avoid potential collinearity in the regression analysis, we chose the following six variables as covariates from the list of the potential risk factors identified by the study team: *sex* (male vs. female), *age at entry* (subjects entered the study five years after the cancer diagnosis), *socioeconomic status* (SES, high vs. low based on the income in the neighbourhood at start of follow-up), *relapse or second cancer* (RSC, yes vs. no relapse or second cancer at start of follow-up), *diagnosis period* (1990s vs. 1980s), and *treatment* (chemo but no radiation, radiation but no chemo, both chemo and radiation, or others). A standardized age value $(age - 5)/20$ was used in the regression models.

In the population sample $\mathcal{Q}$, the covariates sex, SES, and age at entry are available, but the cancer-related covariates (RSC, diagnosis period, and treatment) are not. The counts of visits to both GPs and specialists during the entire follow-up period were calculated for both the cohort and the population. In the population, an individual's start time of follow-up was chosen according to that of a random survivor with the same sex and birth-year. The data $\mathcal{Q}$ has distributions that match the survivor cohort for age at entry and observation length; see Table 2.4.

We excluded individuals who were missing information for the covariates or had an observation period of zero length. We also excluded those in the population who were older than (20+5) = 25 years in 1986, at that time the oldest possible age of the survivors. We also excluded a few outliers. This reduced the size of $\mathcal{P}$ to $n = 1,609$ and $\mathcal{Q}$ to $m = 13,793$. Table 2.4 gives the summary statistics of the covariates, response variable, and observation length associated with $\mathcal{P}$ and $\mathcal{Q}$. To avoid confidentiality issues, we report 5th and 95th percentiles instead of minimum and maximum values for the continuous variables and counts.

Table 2.5 summarizes the quasi-Poisson regression analyses conducted with the visit counts from $\mathcal{P}$ and $\mathcal{Q}$ separately. Adjusted for the independent variables, the frequency of physician visits appears significantly higher in the cohort. In both data sets, the male subjects had fewer physician visits than did the females. This is in agreement with the results reported by McBride *et al.* (2011). In addition, the analysis found that there is no significant difference between the two SES groups in both the population and the cohort. The analysis indicates that the visit counts are highly overdispersed: the estimates for the

Table 2.4: Characteristics Summary: Survivor Cohort vs. General Population in CAYACS

Risk factor	Sex		SES		RSC		Diagnosis period		Treatment			
	Male	Female	High	Low	Yes	No	1990s	1980s	Chemo no Rad	Rad no Chemo	Both	Others
Cohort ($\mathcal{P}$: $n = 1609$)	901	708	659	950	168	1441	960	649	660	139	402	408
Population ($\mathcal{Q}$: $m = 13793$)	7784	6009	5124	8669	-	-	-	-	-	-	-	-

Explanatory variable	Age at entry (in years)		Observation length (in years)	
	The five numbers†	Mean (SD)	The five numbers†	Mean (SD)
Cohort ($\mathcal{P}$: $n = 1609$)	5.8, 8.5, 13.7, 20.3, 24.2	14.4 (6.3)	2.19, 5.05, 9.22, 14.00, 19.09	9.70 (5.43)
Population ($\mathcal{Q}$: $m = 13793$)	5.7, 8.4, 13.1, 20.0, 23.9	14.1 (6.2)	1.83, 4.29, 7.80, 12.16, 17.89	8.49 (5.08)

Response variable	Physician-visit counts	
	The five numbers†	Mean (SD)
Cohort ($\mathcal{P}$: $n = 1609$)	<5††, 17, 39, 79, 169	56.39 (52.05)
Population ($\mathcal{Q}$: $m = 13793$)	<5††, 6, 18, 42, 89	27.88 (27.93)

†The sample 5th percentile, first quartile, median, third quartile, and 95th percentile values.

††Due to confidentiality concerns, the BC Ministry of Health does not allow counts lower than 5 to be displayed.

Table 2.5: Quasi-Poisson Regression for General Population and Survivor Cohort

	Population $\mathcal{Q}$			Cohort $\mathcal{P}$		
Factor	*estimate*	*se*	*p-value*	*estimate*	*se*	*p-value*
intercept	1.032	0.031	$< .001$	2.176	0.095	$< .001$
male (vs. female)	-0.376	0.012	$< .001$	-0.360	0.038	$< .001$
age at entry	0.343	0.020	$< .001$	0.152	0.058	0.009
SES high (vs. low)	-0.023	0.013	0.070	-0.008	0.038	0.832
ln(time period)	1.089	0.012	$< .001$	0.876	0.035	$< .001$
dispersion parameter	14.67	$0.24^{\dagger}$		31.78	$1.84^{\dagger}$	

†Standard errors of dispersion parameters estimated by Bootstrap with BS sample size=1000.

overdispersion parameter are 14.67 and 31.78 for the population and the cohort, respectively. The much larger overdispersion for the cohort, along with the higher overall visit frequency, signals potential strata of physician visits in the cohort.

2.5.2 Likelihood-Based Analysis of Visit Counts Under a Latent Class Model

We used the LCM of Section 2.2 to formulate the visit counts of the cohort. We evaluated the MLE and pseudo-MLE presented in Section 2.3. Table 2.6 summarizes the analysis in the first two panels. Both the MLE and pseudo-MLE analyses identified several significant risk factors for later effects: (i) RSC, (ii) diagnosis in 1980s rather than 1990s, and (iii) treatment with radiation but no chemo or both radiation and chemo rather than other treatments. The pseudo-MLE also found a significantly higher risk for female survivors.

For illustration, we present in Figure 2.1 the estimated at-risk probability functions $p(\hat{\alpha}; \mathbf{Z})$ of age at entry together with pointwise approximate 95% confidence intervals (CIs) from the MLE and pseudo-MLE for three typical groups. Group A contains females diagnosed in the 1980s, with RSC, who received radiation treatment; Group B contains females diagnosed in the 1980s, without RSC, who received radiation treatment; and Group C contains males diagnosed in the 1990s, without RSC, with treatment other than chemo/radiation. The risk of later effects for the three groups is found by both the MLE and pseudo-MLE to be significantly different. People in Group A seem likely to suffer such effects, and those in Group C have a low risk.

The MLE and pseudo-MLE analyses are consistent with the findings of a significantly lower visit rate associated with male survivors across the two risk classes and a similar association with the length of the observation period in the not-at-risk class. The results indicate slightly different magnitudes of the visit frequency in the not-at-risk class. See Figure 2.2 for the estimated means of the visit counts over time from the MLE and pseudo-MLE for the two risk classes. The MLE analysis showed that the visit frequency was not

significantly associated with either age at entry or SES across the two risk classes. This disagrees only with the quasi-Poisson regression for age at entry in the not-at-risk class. The pseudo-MLE analysis, using the quasi-Poisson estimates for the general population for the not-at-risk class, yielded results for the visit frequency of the at-risk class and for the risk probability that were similar to those of the MLE analysis. Figure 2.2 presents the estimated means of the cumulative visit counts of the two risk classes over time, along with the pointwise approximate 95% CIs, from the MLE and pseudo-MLE for female and male subjects with low SES and average age at entry.

To verify the findings of the pseudo-MLE and further assess its efficiency, we also evaluated the MLE with the data from the cohort in combination with the population, as described at the beginning of Section 2.3.2. See Section E of the Supplementary Materials in Wang *et al.* (2014). The resulting parameter estimates and estimated standard errors are almost identical to the corresponding estimates from the pseudo-MLE.

We remark that, under the mixture Poisson model assumed in the MLE and pseudo-MLE analyses, the variance of the counts conditional on T, $\mathbf{Z}$ is the mean $E(Y|T,\mathbf{Z})$ plus $p(\mathbf{Z};\alpha)[1-p(\mathbf{Z};\alpha)][\mu_1(T,\mathbf{Z};\beta)-\mu_0(T,\mathbf{Z};\theta)]^2$. This together with the parameter estimates under the LCM yields estimators for the overall overdispersion parameter for the cohort of 18.11 (for the MLE) and 20.80 (for the pseudo-MLE). Compared with the quasi-Poisson analysis for the cohort, about two-thirds of the large overdispersion of the visit counts can be attributed to the two risk classes by the mixture Poisson model. The unexplained part of the overdispersion indicates a departure of the counts from this model. Caution is necessary in the application of these results.

2.6 Summary and Discussion

Motivated by the visit count analysis of the CAYACS program, we have proposed an LCM to formulate the event counts from a cohort with two unobservable classes: the at-risk class and the not-at-risk class. Under a mixture Poisson model, we have presented two likelihood-based inference procedures, the MLE and pseudo MLE, for cross-sectional counts. The pseudo-MLE procedure employs a consistent estimator of the distribution of the not-at-risk class based on the population data. Compared with the MLE with the primary data, it requires less computational effort, has consistency and asymptotic normality, and has potentially higher efficiency. One may apply the proposed methodology with little modification in situations involving more than two classes.

The population data $\mathcal{Q}_{\text{orig}}$ that does not match the start times to those of the cohort was used by Wang *et al.* (2014). We have seen in Table 2.4 that the population data $\mathcal{Q}$ has a similar distribution to $\mathcal{P}$ in age at entry and observation length, with means 14.1 and 8.49, respectively. For $\mathcal{Q}_{\text{orig}}$ the corresponding values are 10.0 and 12.73 respectively (Wang *et al.*, 2014). The pseudo-MLE result based on $\mathcal{P}$ and $\mathcal{Q}_{\text{orig}}$ is listed in the last panel of Table

Table 2.6: Estimates of Parameters and Standard Errors for CAYACS Data[†]

		MLE from $\mathcal{P}$		pseudo-MLE from $\mathcal{P}$ & $\mathcal{Q}$		pseudo-MLE from $\mathcal{P}$ & $\mathcal{Q}_{\text{orig}}$	
Factor		*estimate*	*se.sw*	*estimate*	*pse.sw*	*estimate*	*pse.sw*
In the Risk Model							
Intercept		-0.332	(0.207)	**-0.353**	(0.187)	**-0.683**	(0.188)
Male (vs. female)		-0.348	(0.217)	**-0.338**	(0.131)	**-0.315**	(0.130)
Age at entry		-0.037	(0.324)	-0.164	(0.217)	-0.111	(0.219)
SES high (vs. low)		-0.103	(0.258)	0.038	(0.139)	0.028	(0.132)
RSC (vs. not)		**1.201**	(0.215)	**1.175**	(0.181)	**1.253**	(0.177)
Diagnosis period 1990s (vs. 1980s)		**-0.960**	(0.160)	**-0.753**	(0.132)	**-0.545**	(0.129)
Treatment (vs. other)	Chemo no rad	0.112	(0.161)	0.119	(0.149)	0.145	(0.151)
	Rad no chemo	**0.673**	(0.231)	**0.547**	(0.221)	**0.515**	(0.224)
	Both	**0.357**	(0.164)	**0.367**	(0.160)	**0.392**	(0.162)
In the Frequency Model for the At-Risk Group							
Intercept		**3.665**	(0.172)	**3.466**	(0.118)	**3.386**	(0.111)
Male (vs. female)		**-0.208**	(0.073)	**-0.196**	(0.050)	**-0.180**	(0.047)
Age at entry		0.094	(0.099)	0.137	(0.077)	0.125	(0.076)
SES high (vs. low)		-0.0005	(0.078)	-0.029	(0.049)	-0.023	(0.046)
ln(time period)		**0.474**	(0.062)	**0.544**	(0.042)	**0.592**	(0.041)
				GEE estimates		GEE estimates	
In the Frequency Model for the Not-At-Risk Group				Based on $\mathcal{Q}$		Based on $\mathcal{Q}_{\text{orig}}$	
Intercept		**1.580**	(0.129)	**1.032**	(0.031)	**0.751**	(0.038)
Male (vs. female)		**-0.358**	(0.072)	**-0.376**	(0.012)	**-0.362**	(0.011)
Age at entry		0.165	(0.102)	**0.343**	(0.020)	**0.306**	(0.019)
SES high (vs. low)		0.030	(0.083)	-0.023	(0.013)	**-0.047**	(0.011)
ln(time period)		**0.892**	(0.051)	**1.089**	(0.012)	**1.263**	(0.013)

[†]Significant effect with p-value ≤ 0.05 in **Boldface**

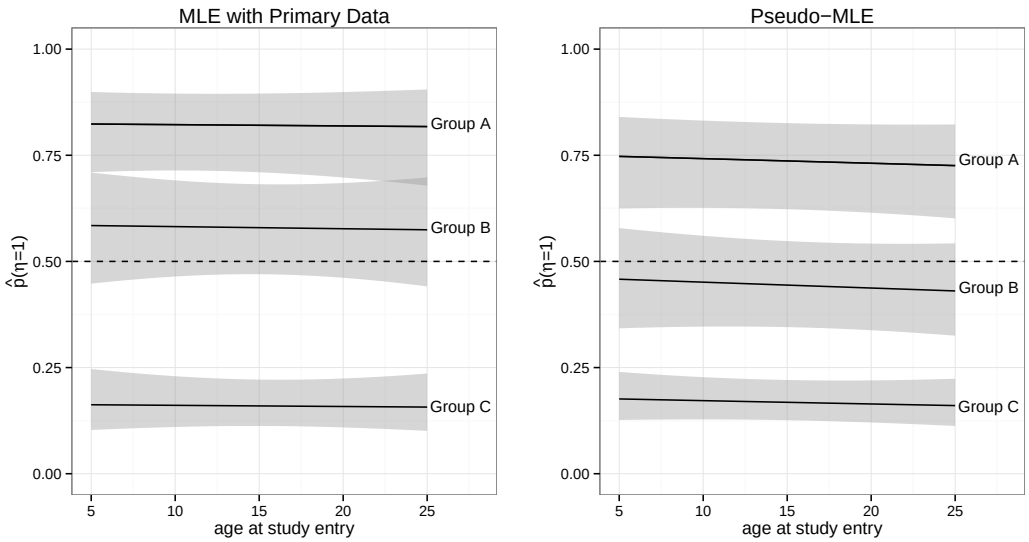

Figure 2.1: Estimated risk probabilities with approximate 95% CIs for three groups. **Group A**: Female, diagnosed in 1980s, with RSC, and treated with radiation. **Group B**: Female, diagnosed in 1980s, no RSC, and treated with radiation. **Group C**: Male, diagnosed in 1990s, no RSC, and treated without chemo/radiation.

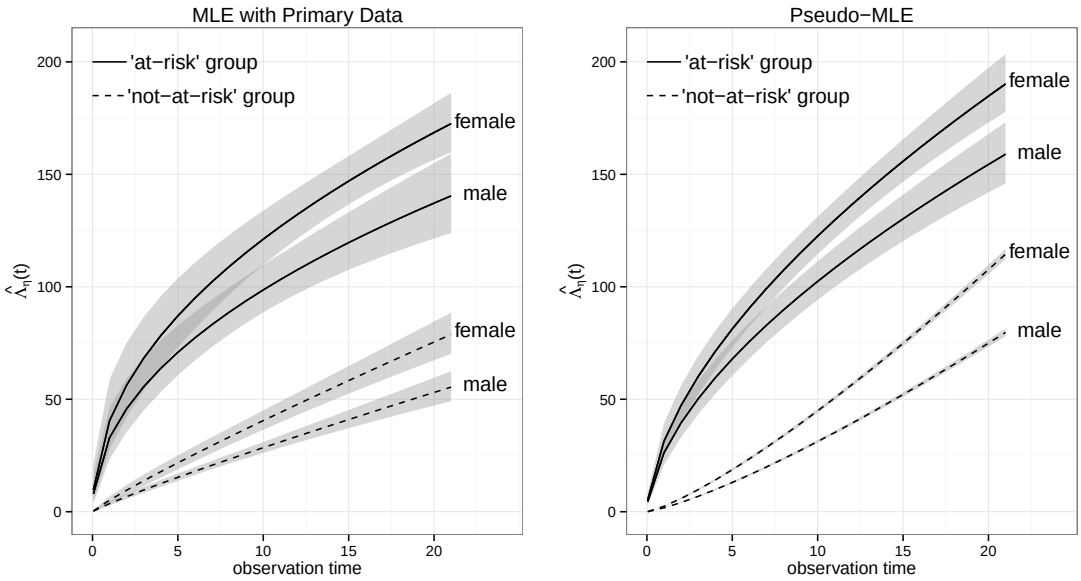

Figure 2.2: Estimated mean function of cumulative visit counts with approximate 95% CIs, for survivors with low SES and average age at entry.

2.6 for comparison. The quasi-Poisson regressions for $\mathcal{Q}$ and $\mathcal{Q}_{\text{orig}}$ respectively are rather different, especially in the magnitudes of intercept and log(time period), which indicates the nonlinear relationship of visit counts and subject aging. Longitudinal studies may be necessary. The estimates for the not-at-risk class based on $\mathcal{Q}$ are much closer to the MLE based only on $\mathcal{P}$ than to the estimates based on $\mathcal{Q}_{\text{orig}}$, so the pseudo-MLE with $\mathcal{Q}$ has similar estimates for the at-risk class and the risk probability models. This similarity supports the use of supplementary information and the pseudo-MLE procedure for the CAYACS data.

The simulation results show that the likelihood-based estimating procedures are quite efficient under the mixture Poisson model, but they lack robustness to model misspecification. Therefore, there is a need for a robust inference procedure. In Chapter 3 we develop an extension of the GEE approaches (Liang and Zeger, 1986). The new approach can be straightforwardly extended to analyze the cost data associated with the physician claims in CAYACS. The model formulation assumes that the time effect for the count of interest is proportional to the length of the observation period on average. The available longitudinal data allow us to consider a semiparametric specification for the mean function of the counts or costs over time, and thus to check this assumption. These longitudinal procedures are discussed in Chapters 4 and 5.

Chapter 3

Extended GEE Procedures with Cross-Sectional Counts

3.1 Introduction

In an analysis with an LCM, one usually needs to specify the underlying probability model in a parametric form for each of the latent classes to avoid nonidentifiability problems in general. In the previous chapter we made good use of the rich information from the general population and developed a so-called pseudo-MLE procedure that did not need a distributional assumption for the not-at-risk class. However, the simulation studies indicated that the pseudo-MLE procedure lacked robustness to overdispersed counts in the at-risk class. Overall, the MLE and pseudo-MLE methods for the LCM in Chapter 2 were quite efficient but lacked robustness to model misspecification, especially for the parameters in the risk probability of the latent indicator. The analysis of the CAYACS data showed that the visit counts in both the cohort and the population were heavily overdispersed. Even when the pseudo-MLE was used, the two latent classes could account for some but not all of the overdispersion. There are always practical concerns about model misspecification. It is often necessary to develop robust inferential procedures, particularly in epidemiological and medical applications. Thus, in this chapter we develop distribution-free estimating procedures for the LCM.

Table 2.6 shows that RSC status has the greatest impact on survivors at risk to later effects; see also Figure 2.1. If we choose the cut-off value of the estimated risk probability $P(\eta = 1|\mathbf{Z}; \hat{\alpha})$ to be around 0.4, almost all the individuals with RSC ($\delta = 1$) are predicted to be in the at-risk class ($\eta = 1$). This supports the assumption made in this chapter. On average, the RSC subgroup in the cohort has more frequent physician visits than the rest of the cohort. They can be considered as representatives of the at-risk class. Based on this assumption, we are able to develop extended GEE inference procedures for the LCM. This overcomes the overdispersion issue of the counts in the mixture Poisson model

and the corresponding likelihood-based methods. The extended GEE approaches can be straightforwardly applied to analyze the cost data of the CAYACS physician claims. They can also be easily extended to analyze longitudinal data by introducing within-subject correlation. The formulation and computation are much easier for these approaches than for likelihood-based approaches for longitudinal data. The longitudinal extensions will be presented in Chapter 5.

3.2 Model Specification

The response $\mathbb{Y} = Y$ is the cross-sectional count, so $[\mathbb{Y}|\cdot] = [Y|\cdot]$, and η is the latent risk indicator. The regression functions for the risk probability and for the counts in the two latent classes are specified as $p(\mathbf{Z};\alpha)$, $\mu_1(T,\mathbf{Z};\beta)$, and $\mu_0(T,\mathbf{Z};\theta)$, respectively.

Conceptually, if a survivor has RSC, he/she is experiencing ongoing problems. Therefore, he/she will have more frequent physician visits compared to the general population and belong to the at-risk class. With this assumption, further descriptive analysis of the CAYACS data indicated that RSC at any time before or after the start of the study greatly increased the overall frequency of physician visits. Hence, we can make additional assumptions. An RSC implies that the survivor belongs to the at-risk class. Furthermore, the overall at-risk class has the same visit patterns regardless of RSC status. Let δ be the RSC indicator. The above assumptions can be formulated as:

Assumption (i) $P(\eta = 1|\delta = 1, \mathbf{Z}) = 1$, i.e., $\delta = 1$ implies $\eta = 1$.

Assumption (ii) $[Y|\eta, \delta, T, \mathbf{Z}] = [Y|\eta, T, \mathbf{Z}]$.

Assumptions (i) and (ii) imply **Assumption (iii)**: $[Y|\eta = 1, T, \mathbf{Z}] = [Y|\delta = 1, T, \mathbf{Z}]$.

To complete the model specification when Assumption (i) is used explicitly, we denote the conditional probability of a survivor with RSC as $P(\delta = 1|\mathbf{Z}) = q(\mathbf{Z})$. In practice, a logistic regression model can be adopted and specified up to a parameter ρ, i.e., $\text{logit}\{q(\mathbf{Z};\rho)\} = \rho_0 + \rho_1^{'}\mathbf{Z}$.

Our goal in this chapter is to develop robust estimating procedures for the parameters α, β, and θ in $p(\mathbf{Z};\alpha)$, $\mu_1(T,\mathbf{Z};\beta)$, and $\mu_0(T,\mathbf{Z};\theta)$ without a distributional assumption for Y, with $\mathcal{P} = \{(Y_i, \delta_i, T_i, \mathbf{Z}_i) : i = 1, \ldots, n\}$ and $\mathcal{Q} = \{(Y_i, T_i, \mathbf{Z}_i) : i = 1, \ldots, m\}$. Clearly, δ is not included in $\mathbf{Z}$, which is the difference between this model and that in Chapter 2.

3.3 Likelihood-Based Estimation with Counts Using Partially Available Membership Information

As in Section 2.3, the event count Y conditional on T and $\mathbf{Z}$ is further specified as a mixture Poisson distribution. The likelihood-based estimations are presented in this section. The

full MLE procedure based on the distribution of δ and the mixture Poisson of Y jointly is computationally expensive. Thus, we also propose three pseudo-MLE procedures. They are computationally less intensive, and the last two are robust to distribution misspecification in the not-at-risk class because they use estimation based on the general population.

The underlying probability models for Y with the two latent classes, $[Y|\eta = 1, T, \mathbf{Z}; \beta]$ and $[Y|\eta = 0, T, \mathbf{Z}; \theta]$, are assumed to be Poisson distributions with conditional expectations $\mu_\eta(T, \mathbf{Z})$ for $\eta = 1, 0$ respectively. The likelihood function for all the parameters $\phi = (\rho, \alpha, \beta, \theta)$ is based on the joint distribution of Y and δ:

$$\begin{aligned}
[Y, \delta|T, \mathbf{Z}; \phi] &= [Y|\delta, T, \mathbf{Z}; \phi][\delta|\mathbf{Z}; \rho] \\
&= [\delta|\mathbf{Z}; \rho] \sum_{\eta=0}^{1} [Y|\eta, \delta, T, \mathbf{Z}; \beta/\theta][\eta|\delta, \mathbf{Z}; \rho, \alpha] \qquad (3.1) \\
&= [\delta|\mathbf{Z}; \rho] \sum_{\eta=0}^{1} [Y|\eta, T, \mathbf{Z}; \beta/\theta][\eta|\delta, \mathbf{Z}; \rho, \alpha]. \qquad (3.2)
\end{aligned}$$

The simplification from (3.1) to (3.2) is based on Assumption (ii). Assumption (i) states that $\delta = 1$ implies $\eta = 1$, but $\eta = 1$ des not necessarily imply $\delta = 1$. The conditional probability $[\eta|\delta, \mathbf{Z}; \rho, \alpha]$ can be derived by decomposing $P(\eta = 1|\mathbf{Z})$:

$$\begin{aligned}
P(\eta = 1|\mathbf{Z}) &= P(\eta = 1, \delta = 1|\mathbf{Z}) + P(\eta = 1, \delta = 0|\mathbf{Z}) \\
&= P(\eta = 1|\delta = 1, \mathbf{Z})P(\delta = 1|\mathbf{Z}) + P(\eta = 1|\delta = 0, \mathbf{Z})P(\delta = 0|\mathbf{Z}) \\
&= P(\delta = 1|\mathbf{Z}) + P(\eta = 1|\delta = 0, \mathbf{Z})P(\delta = 0|\mathbf{Z}).
\end{aligned}$$

Given $P(\eta = 1|\delta = 1, \mathbf{Z}) = 1$ from Assumption (i),

$$P(\eta = 1|\delta = 0, \mathbf{Z}) = \frac{p(\mathbf{Z}; \alpha) - q(\mathbf{Z}; \rho)}{1 - q(\mathbf{Z}; \rho)}. \qquad (3.3)$$

The likelihood function of ϕ with the primary data $\mathcal{P}$ is $L(\phi; \mathcal{P}) \propto \prod_{i=1}^{n} [Y_i, \delta_i|T_i, \mathbf{Z}_i; \phi] = \prod_{i=1}^{n} [Y_i|\delta_i, T_i, \mathbf{Z}_i; \phi] \times \prod_{i=1}^{n} [\delta_i|\mathbf{Z}_i; \rho]$. We write the log-likelihood function as $l(\phi) = \log L(\phi)$. Then

$$l(\phi) = l_1(\phi) + l_2(\rho), \qquad (3.4)$$

where $l_1(\phi) = \sum_{i=1}^{n} \log[Y_i|\delta_i, T_i, \mathbf{Z}_i; \phi]$ and $l_2(\rho) = \sum_{i=1}^{n} \log[\delta_i|\mathbf{Z}_i; \rho]$. We will sometimes use l, l_1, and l_2 for simplicity.

3.3.1 MLE with Primary Data

We can directly maximize $L(\phi)$ or $l(\phi)$ to get the MLE of ϕ, but this is computationally intensive and gives a poor result. We instead use an adaptation of the EM algorithm (Dempster *et al.*, 1977) to compute the MLE of the parameters; see Appendix A. Under the usual regularity conditions, the MLE $(\hat{\rho}, \hat{\alpha}, \hat{\beta}, \hat{\theta})$ has asymptotic consistency and normality:

$$\sqrt{n}\Big((\hat{\rho}, \hat{\alpha}, \hat{\beta}, \hat{\theta})' - (\rho, \alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Pi^{-1}\Sigma\Pi^{-1}\Big)$$

where

$$\frac{1}{\sqrt{n}}\frac{\partial l(\phi)}{\partial \phi} \xrightarrow[n\to\infty]{d} N\Big(0, \Sigma\Big). \tag{3.5}$$

Here Π is the Fisher information matrix

$$-\frac{1}{n}\frac{\partial^2 l(\phi)}{\partial \phi^2} \xrightarrow[n\to\infty]{a.s.} \Pi$$

and $\Sigma = \Pi$ for the MLE. It is important to note that the symbols/notation introduced in the asymptotics in Section 3.3 and Appendix B, i.e., ϕ, Σ, and Π and their extensions, are separate from those used in the rest of the thesis, so there is no inconsistency.

Note that for independent but not identically distributed data the variance of the i^{th} element on the left-hand side of (3.5) depends on $\mathbf{Z}_i$ for the i^{th} observation. Provided the $\mathbf{Z}_i$'s are chosen randomly or have a limiting distribution (Yuan and Jennrich, 1998), property (3.5) holds. Similar properties hold for our three pseudo-likelihood functions.

The variance of the likelihood estimating equations, Σ, can be consistently estimated by $\frac{1}{n}\frac{\partial l}{\partial \phi}\frac{\partial l}{\partial \phi}'$ with the MLE plugged in. The Fisher information matrix can be estimated by $-\frac{1}{n}\frac{\partial^2 l}{\partial \phi^2}$ with the MLE plugged in. Then the asymptotic variance of the MLE can be estimated by either $\frac{1}{n}\hat{\Sigma}^{-1}$ or $\frac{1}{n}\hat{\Pi}^{-1}$.

3.3.2 Type A Pseudo-MLE with Primary Data

The log-likelihood function (3.4) is a sum of two terms, where $l_2(\rho)$ is a function only of ρ. We develop a pseudo-MLE procedure to estimate ρ only from $l_2(\rho)$, denoted $\hat{\rho}_A$, and to estimate α, β, θ from $l_1(\phi)$ with ρ fixed at $\hat{\rho}_A$. The resulting estimator is called the type A pseudo-MLE and denoted $\hat{\alpha}_A, \hat{\beta}_A, \hat{\theta}_A$. The method for finding $\hat{\rho}_A$ is equivalent to fitting a logistic regression model with δ as a response to the covariates $\mathbf{Z}$. The type A pseudo-MLE of (α, β, θ) is obtained by the EM algorithm in Appendix A with ρ fixed at $\hat{\rho}_A$. This pseudo-likelihood procedure works because $l_1(\phi)$ carries little information about ρ.

The asymptotic properties of the type A pseudo-MLE have been derived; see Appendix B. Under the usual regularity conditions, the pseudo-MLE $\hat{\phi}_A = (\hat{\rho}_A, \hat{\alpha}_A, \hat{\beta}_A, \hat{\theta}_A)$ has asymp-

totic consistency and normality:

$$\sqrt{n}\Big((\hat{\rho}_A, \hat{\alpha}_A, \hat{\beta}_A, \hat{\theta}_A)' - (\rho, \alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Pi_A^{-1}\Sigma_A(\Pi_A^{-1})'\Big)$$

where

$$\begin{pmatrix} \frac{1}{\sqrt{n}}\frac{\partial l_2(\rho)}{\partial \rho} \\ \frac{1}{\sqrt{n}}\frac{\partial l_1(\phi)}{\partial(\alpha,\beta,\theta)} \end{pmatrix} \xrightarrow[n\to\infty]{d} N\Big(0, \Sigma_A\Big)$$

and

$$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2 l_2(\rho)}{\partial \rho^2} & 0 \\ \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta,\theta)\partial\rho} & \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta,\theta)^2} \end{pmatrix} \xrightarrow[n\to\infty]{a.s.} \Pi_A.$$

The variance of the type A pseudo-likelihood estimating equations, Σ_A, can be consistently estimated by $\frac{1}{n}\begin{bmatrix} \frac{\partial l_2}{\partial \rho} \\ \frac{\partial l_1}{\partial(\alpha,\beta,\theta)} \end{bmatrix}\begin{bmatrix} \frac{\partial l_2}{\partial \rho} \\ \frac{\partial l_1}{\partial(\alpha,\beta,\theta)} \end{bmatrix}'$ with $\hat{\phi}_A$ plugged in. Π_A can be estimated by $-\frac{1}{n}\begin{bmatrix} \frac{\partial^2 l_2}{\partial \rho^2} & 0 \\ \frac{\partial^2 l_1}{\partial(\alpha,\beta,\theta)\partial\rho} & \frac{\partial^2 l_1}{\partial(\alpha,\beta,\theta)^2} \end{bmatrix}$ with $\hat{\phi}_A$ plugged in. Then the asymptotic variance of the type A pseudo-MLE can be estimated by $\frac{1}{n}\hat{\Pi}_A^{-1}\hat{\Sigma}_A(\hat{\Pi}_A^{-1})'$.

3.3.3 Type B Pseudo-MLE with Primary Data and Independent Supplementary Information

The $\eta = 0$ class has the same visit patterns as the general population. Chapter 2 developed a pseudo-MLE where θ was estimated from an independent supplementary data set, and the pseudo-MLE of the primary parameters (α, β) was efficient, consistent, and robust to distribution misspecification of the $\eta = 0$ class. We can apply this method to the likelihood function (3.4) to develop another type of pseudo-MLE.

Suppose there is an independent supplementary dataset $\mathcal{Q} = \{(Y_i, T_i, \mathbf{Z}_i) : i = 1, \ldots, m\}$ with sample size m, where $\frac{n}{m} \to r > 0$ as $n \to \infty$ and $m \to \infty$. As in Section 2.3.2, a consistent estimator of θ in $\mu_0(T, \mathbf{Z}; \theta)$ can be obtained from $\mathcal{Q}$ without a distributional assumption on Y. Let the estimator be $\tilde{\theta}$; it has asymptotic normality: $\sqrt{m}(\tilde{\theta} - \theta) \xrightarrow[m\to\infty]{d} N\Big(0, AV_{\tilde{\theta}}(\theta)\Big)$. By maximizing the log-likelihood function (3.4) with θ fixed at $\tilde{\theta}$, we get a pseudo-MLE for (ρ, α, β), called the type B pseudo-MLE and denoted $\hat{\rho}_B, \hat{\alpha}_B, \hat{\beta}_B$. The type B pseudo-MLE of (ρ, α, β) can be evaluated by the EM algorithm in Appendix A with θ fixed at $\tilde{\theta}$.

The asymptotic properties of the type B pseudo-MLE have been derived similarly; see Appendix B. Under the usual regularity conditions, the pseudo-MLE $\hat{\phi}_B = (\hat{\rho}_B, \hat{\alpha}_B, \hat{\beta}_B, \tilde{\theta})$ has asymptotic consistency and normality:

$$\sqrt{n}\Big((\hat{\rho}_B, \hat{\alpha}_B, \hat{\beta}_B, \tilde{\theta})' - (\rho, \alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Pi_B^{-1}\Sigma_B(\Pi_B^{-1})'\Big)$$

where

$$\Sigma_B = \begin{pmatrix} \Sigma_B^* & 0 \\ 0 & r^{-1}AV_{\tilde{\theta}}(\theta) \end{pmatrix} \text{ with } \frac{1}{\sqrt{n}}\frac{\partial l(\phi)}{\partial(\rho,\alpha,\beta)} \xrightarrow[n\to\infty]{d} N\Big(0, \Sigma_B^*\Big)$$

and

$$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2 l(\phi)}{\partial(\rho,\alpha,\beta)^2} & \frac{\partial^2 l(\phi)}{\partial(\rho,\alpha,\beta)\partial\theta} \\ 0 & -m \end{pmatrix} \xrightarrow[n\to\infty]{a.s.} \Pi_B.$$

The variance of the type B pseudo-likelihood estimating equations, Σ_B, can be consistently estimated by $\hat{\Sigma}_B = \begin{bmatrix} \hat{\Sigma}_B^* & 0 \\ 0 & \frac{m}{n}\widehat{AV_{\tilde{\theta}}(\theta)} \end{bmatrix}$, where $\hat{\Sigma}_B^* = \frac{1}{n}\frac{\partial l}{\partial(\rho,\alpha,\beta)}\frac{\partial l}{\partial(\rho,\alpha,\beta)}'$ with $\hat{\phi}_B$ plugged in. Π_B can be estimated by $-\frac{1}{n}\begin{bmatrix} \frac{\partial^2 l}{\partial(\rho,\alpha,\beta)^2} & \frac{\partial^2 l}{\partial(\rho,\alpha,\beta)\partial\theta} \\ 0 & -m \end{bmatrix}$ with $\hat{\phi}_B$ plugged in. The asymptotic variance of the type B pseudo-MLE can be evaluated by $\frac{1}{n}\hat{\Pi}_B^{-1}\hat{\Sigma}_B(\hat{\Pi}_B^{-1})'$.

3.3.4 Type AB Pseudo-MLE with Primary Data and Independent Supplementary Information

The third type of pseudo-MLE combines the ideas from the type A and type B pseudo-MLEs. We plug $\hat{\rho}_A$ from Section 3.3.2 and $\tilde{\theta}$ from Section 3.3.3 into the log-likelihood function $l_1(\phi)$ in (3.4). The estimator of the primary parameters (α, β) that maximizes the pseudo-likelihood function $l_1(\hat{\rho}_A, \alpha, \beta, \tilde{\theta})$ is called the type AB pseudo-MLE, denoted $\hat{\alpha}_{AB}, \hat{\beta}_{AB}$. The type AB pseudo-MLE of (α, β) can be evaluated by the EM algorithm in Appendix A with ρ and θ fixed at $\hat{\rho}_A$ and $\tilde{\theta}$.

The asymptotic properties of the type AB pseudo-MLE have also been derived; see Appendix B. Under the usual regularity conditions, the type AB pseudo-MLE $\hat{\phi}_{AB} = (\hat{\rho}_A, \hat{\alpha}_{AB}, \hat{\beta}_{AB}, \tilde{\theta})$ has asymptotic consistency and normality:

$$\sqrt{n}\Big((\hat{\rho}_A, \hat{\alpha}_{AB}, \hat{\beta}_{AB}, \tilde{\theta})' - (\rho, \alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Pi_{AB}^{-1}\Sigma_{AB}(\Pi_{AB}^{-1})'\Big)$$

where

$$\Sigma_{AB} = \begin{pmatrix} \Sigma_{AB}^* & 0 \\ 0 & r^{-1}AV_{\tilde{\theta}}(\theta) \end{pmatrix} \text{ with } \begin{pmatrix} \frac{1}{\sqrt{n}}\frac{\partial l_2(\rho)}{\partial(\rho)} \\ \frac{1}{\sqrt{n}}\frac{\partial l_1(\phi)}{\partial(\alpha,\beta)} \end{pmatrix} \xrightarrow[n\to\infty]{d} N\Big(0, \Sigma_{AB}^*\Big)$$

and

$$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2 l_2(\rho)}{\partial\rho^2} & 0 & 0 \\ \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta)\partial\rho} & \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta)^2} & \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta)\partial\theta} \\ 0 & 0 & -m \end{pmatrix} \xrightarrow[n\to\infty]{a.s.} \Pi_{AB}.$$

The variance of the type AB pseudo-likelihood estimating equations, Σ_{AB}, can be consistently estimated by $\hat{\Sigma}_{AB} = \begin{bmatrix} \hat{\Sigma}_{AB}^* & 0 \\ 0 & \frac{m}{n}\widehat{AV_{\tilde{\theta}}(\theta)} \end{bmatrix}$, where $\hat{\Sigma}_{AB}^* = \frac{1}{n}\begin{bmatrix} \frac{\partial l_2}{\partial\rho} \\ \frac{\partial l_1}{\partial(\alpha,\beta)} \end{bmatrix}\begin{bmatrix} \frac{\partial l_2}{\partial\rho} \\ \frac{\partial l_1}{\partial(\alpha,\beta)} \end{bmatrix}'$ with

$\hat{\phi}_{AB}$ plugged in. Π_{AB} can be estimated by $-\frac{1}{n}\begin{bmatrix} \frac{\partial^2 l_2}{\partial \rho^2} & 0 & 0 \\ \frac{\partial^2 l_1}{\partial(\alpha,\beta)\partial\rho} & \frac{\partial^2 l_1}{\partial(\alpha,\beta)^2} & \frac{\partial^2 l_1}{\partial(\alpha,\beta)\partial\theta} \\ 0 & 0 & -m \end{bmatrix}$ with $\hat{\phi}_{AB}$ plugged in. The asymptotic variance of the type AB pseudo-MLE can be evaluated by $\frac{1}{n}\hat{\Pi}_{AB}^{-1}\hat{\Sigma}_{AB}(\hat{\Pi}_{AB}^{-1})'$.

In the type B and AB pseudo-MLEs, $\tilde{\theta}$ is estimated from an independent supplementary data set, so the asymptotic variances of $(\hat{\rho}_B, \hat{\alpha}_B, \hat{\beta}_B)$ and $(\hat{\alpha}_{AB}, \hat{\beta}_{AB})$ can be decomposed into two terms as for the pseudo-MLE in Chapter 2 (2.10). The first term is the asymptotic variance when the true value of θ is known, and the second term accommodates the extra variation arising because θ is estimated from an independent data set. The decomposition is also valid for the extended GEE estimators, since the supplementary information is again used to estimate θ. When a set of parameters, e.g., $\hat{\rho}_A$ in the type A and AB pseudo-MLEs and the β estimates from the type P extended GEE (discussed below), is estimated from other correlated data, the method that we used to derive the asymptotic variances remains appropriate. If one ignores the estimation of parameters from other sources, dependent or independent, the estimate of the precision of the estimators is likely to be overly optimistic (Reid, 2010). Remarkably, the variance estimations presented in Sections 3.3 and 3.4 are appropriate for the parameters estimated from other dependent and independent sources and are also robust to distribution misspecification.

3.4 Extended GEE Inference Procedures

The likelihood-based approaches for the LCM presented in the previous section are based on the distribution assumption of the counts. However, in practice overdispersion is a common feature of count data with an assumed Poisson distribution. It is also often desirable to have simple alternatives to MLE methods, such as least squares. We develop extended GEE methods for LCMs by specifying only the mean and variance functions of Y in the two classes and not its underlying distribution.

As in Section 3.2, the conditional expectations of Y for the at-risk and not-at-risk groups are $E(Y|\eta, T, \mathbf{Z}) = \mu_\eta(T, \mathbf{Z})$, where $\eta = 0, 1$. Given the specification of the risk probability $p(\mathbf{Z}; \alpha)$, the marginal expectation of Y given T and $\mathbf{Z}$, $E(Y|T, \mathbf{Z}; \alpha, \beta, \theta)$, denoted $\Lambda(T, \mathbf{Z}; \alpha, \beta, \theta)$, can be derived as

$$\Lambda(T, \mathbf{Z}; \alpha, \beta, \theta) = p(\mathbf{Z}; \alpha)\mu_1(T, \mathbf{Z}; \beta) + \{1 - p(\mathbf{Z}; \alpha)\}\mu_0(T, \mathbf{Z}; \theta). \tag{3.6}$$

The conditional variance functions of Y given T and $\mathbf{Z}$ for each class are $\mathrm{Var}(Y|\eta, T, \mathbf{Z}) = \Sigma_\eta(T, \mathbf{Z})$, which may involve other variance-related parameters. Then the marginal variance of Y given T and $\mathbf{Z}$, denoted $\mathrm{Var}(Y|T, \mathbf{Z}) = \Sigma(T, \mathbf{Z})$, which involves the parameters (α, β, θ),

can be derived as

$$\Sigma(T,\mathbf{Z}) = p(\mathbf{Z};\alpha)\Sigma_1(T,\mathbf{Z}) + [1-p(\mathbf{Z};\alpha)]\Sigma_0(T,\mathbf{Z}) + p(\mathbf{Z};\alpha)[1-p(\mathbf{Z};\alpha)]\{\mu_1(T,\mathbf{Z};\beta) - \mu_0(T,\mathbf{Z};\theta)\}^2.$$

A direct GEE application under the weighted least-squares principle can construct estimating equations for the parameters (α,β,θ) based on the mean function $\Lambda(T,\mathbf{Z};\alpha,\beta,\theta)$ from the primary data $\mathcal{P}$. It may be weighted by $\Sigma(T,\mathbf{Z})$.

Let $\psi = (\alpha,\beta,\theta)$. Let d_1, d_2, and d_3 be the dimensions of α, β, and θ, respectively, and $d = d_1 + d_2 + d_3$. The d-dimensional GEEs for the parameters (α,β,θ) based on the marginal expectation $\Lambda(T,\mathbf{Z};\psi)$ in (3.6) are denoted $GEE(\psi) = 0$, where $GEE(\psi) = \sum_{i=1}^n g(Y_i|T_i,\mathbf{Z}_i;\psi)$. The d-dimensional parametric function $g(Y_i|T_i,\mathbf{Z}_i;\psi)$, also written $g_i(\psi)$, is a conditionally unbiased function of Y, $E[g_i(\psi)|T_i,\mathbf{Z}_i] = 0$.

A linear form of $g_i(\psi)$ is $g_i(\psi) = A_i(\psi)[Y_i - \Lambda(T_i,\mathbf{Z}_i;\psi)]$, where $A_i(\psi)$ can be any d-dimensional function such that:

1. $A_i(\psi) \perp Y_i | T_i, \mathbf{Z}_i$.
2. The following quantities exist: $\mathrm{Var}\{A_i(\psi)[Y_i - \Lambda(T_i,\mathbf{Z}_i;\psi)]\}$, $\partial A_i(\psi)/\partial\psi$, and $E\{\frac{\partial A_i(\psi)}{\partial\psi}[Y_i - \Lambda(T_i,\mathbf{Z}_i;\psi)] - A_i(\psi)\frac{\partial\Lambda(T_i,\mathbf{Z}_i;\psi)}{\partial\psi}\}$.

For example, $A_i(\psi)$ could be $\frac{\partial\Lambda(T_i,\mathbf{Z}_i;\psi)}{\partial\psi}$, $\frac{\partial\Lambda(T_i,\mathbf{Z}_i;\psi)}{\partial\psi}\Lambda(T_i,\mathbf{Z}_i;\psi)^{-1}$, or $\frac{\partial\Lambda(T_i,\mathbf{Z}_i;\psi)}{\partial\psi}\Sigma(T_i,\mathbf{Z}_i)^{-1}$. Godambe (1991) used the term *modified least square estimating equations* for $GEE(\psi) = 0$ with the third choice above for $A_i(\psi)$. The solutions of these equations, under some regularity conditions, converge in probability to the true value, and they are the linear estimating functions with minimum variance.

The $GEE(\psi)$ consists of three parts: d_1-dimensional, d_2-dimensional, and d_3-dimensional estimating functions for α, β, and θ, respectively, i.e., $GEE(\psi) = (GEE^\alpha(\psi)', GEE^\beta(\psi)', GEE^\theta(\psi)')'$. We write $g_i(\psi) = (g_i^\alpha(\psi)', g_i^\beta(\psi)', g_i^\theta(\psi)')'$.

3.4.1 Type J Extended GEE Estimator

The pseudo-MLE in Chapter 2 and the type B and AB pseudo-MLEs above have shown the advantages of using a consistent estimator of θ, $\tilde{\theta}$, from the supplementary data $\mathcal{Q}$. Procedures using $\tilde{\theta}$ are computationally less intensive and more efficient. For example, quasi-Poisson regression can be used to estimate θ in the general population without a distributional assumption on Y. We also use the idea of $\tilde{\theta}$ estimated from $\mathcal{Q}$ for the GEE procedures in this chapter.

We construct extended GEEs $GEE\text{-J}(\psi) = 0$ for the parameters α, β, and θ, where $GEE\text{-J}(\psi) = (GEE^\alpha(\psi)', GEE^\beta(\psi)', m(\tilde{\theta}-\theta)')'$. We refer to these as type J extended GEEs. Their joint solution is called the type J extended GEE estimator and denoted $\tilde{\psi}_J = (\tilde{\alpha}_J, \tilde{\beta}_J, \tilde{\theta})$. The type J extended GEE estimator $\tilde{\psi}_J$ satisfies $GEE\text{-J}(\tilde{\psi}_J) = 0$. Rearrangement of the first-order Taylor expansion of the functions $GEE\text{-J}(\tilde{\psi}_J)$, whose values

are equal to zero, yields the following equations:

$$\frac{1}{\sqrt{n}}\begin{bmatrix} GEE^{\alpha}(\psi) \\ GEE^{\beta}(\psi) \\ m(\tilde{\theta}-\theta) \end{bmatrix} \doteq -\frac{1}{n}\begin{bmatrix} \frac{\partial GEE^{\alpha}(\psi)}{\partial\alpha} & \frac{\partial GEE^{\alpha}(\psi)}{\partial\beta} & \frac{\partial GEE^{\alpha}(\psi)}{\partial\theta} \\ \frac{\partial GEE^{\beta}(\psi)}{\partial\alpha} & \frac{\partial GEE^{\beta}(\psi)}{\partial\beta} & \frac{\partial GEE^{\beta}(\psi)}{\partial\theta} \\ 0 & 0 & -m \end{bmatrix} \cdot \sqrt{n}\begin{bmatrix} \tilde{\alpha}_J-\alpha \\ \tilde{\beta}_J-\beta \\ \tilde{\theta}-\theta \end{bmatrix}. \quad (3.7)$$

Given the above conditions on $A_i(\psi)$, the estimating function $GEE(\psi)$ has asymptotic normality, and its first-order derivatives converge to a constant matrix asymptotically surely. Moreover, because the consistent estimator $\tilde{\theta}$ is estimated from an independent sample, the left-hand side of (3.7) has asymptotic normality:

$$\frac{1}{\sqrt{n}}\begin{bmatrix} GEE^{\alpha}(\psi) \\ GEE^{\beta}(\psi) \\ m(\tilde{\theta}-\theta) \end{bmatrix} \xrightarrow[n\to\infty]{d} N(0,\Phi_J), \text{ where } \Phi_J = \begin{pmatrix} \Phi_J^* & 0 \\ 0 & r^{-1}AV_{\tilde{\theta}}(\theta) \end{pmatrix}$$

and the first term on the right-hand side of (3.7) converges to a constant matrix Ψ_J asymptotically surely:

$$-\frac{1}{n}\begin{bmatrix} \frac{\partial GEE^{\alpha}(\psi)}{\partial\alpha} & \frac{\partial GEE^{\alpha}(\psi)}{\partial\beta} & \frac{\partial GEE^{\alpha}(\psi)}{\partial\theta} \\ \frac{\partial GEE^{\beta}(\psi)}{\partial\alpha} & \frac{\partial GEE^{\beta}(\psi)}{\partial\beta} & \frac{\partial GEE^{\beta}(\psi)}{\partial\theta} \\ 0 & 0 & -m \end{bmatrix} \xrightarrow[n\to\infty]{a.s.} \Psi_J. \quad (3.8)$$

Therefore, it is easy to derive the asymptotic distribution for the type J extended GEE estimator $\tilde{\psi}_J$ as:

$$\sqrt{n}\Big((\tilde{\alpha}_J,\tilde{\beta}_J,\tilde{\theta})'-(\alpha,\beta,\theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0,\Psi_J^{-1}\Phi_J(\Psi_J^{-1})'\Big).$$

The variance of $(GEE^{\alpha}(\psi), GEE^{\beta}(\psi))$, Φ_J^*, can be consistently estimated by $\hat{\Phi}_J^* = \frac{1}{n}\sum_{i=1}^{n}[g_i^{\alpha}(\psi), g_i^{\beta}(\psi)][g_i^{\alpha}(\psi), g_i^{\beta}(\psi)]'$ with $\tilde{\psi}_J$ plugged in; then $\hat{\Phi}_J = \begin{bmatrix} \hat{\Phi}_J^* & 0 \\ 0 & \frac{m}{n}\widehat{AV_{\tilde{\theta}}(\theta)} \end{bmatrix}$. Moreover, Ψ_J can be estimated by the left-hand side of (3.8) with $\tilde{\psi}_J$ plugged in. Then the asymptotic variance of the type J extended GEE estimator can be evaluated by $\frac{1}{n}\hat{\Psi}_J^{-1}\hat{\Phi}_J(\hat{\Psi}_J^{-1})'$.

The type J extended GEE estimator $\tilde{\psi}_J$ does not use the additional assumptions in Section 3.2, which basically ignore the information on δ. Although we can establish their asymptotic properties, the estimating equations may not be sensitive and efficient, since they rely only on the structure of $\Lambda(T,\mathbf{Z};\psi)$, which does not provide as much information as likelihood functions provide. We want to make use of the δ information in GEE, so we propose the following type P extended GEE.

3.4.2 Type P Extended GEE Estimator

The type P extended GEE estimator uses Assumption (iii) in Section 3.2 explicitly. The subgroup $\delta = 1$ is representative of the $\eta = 1$ class with the same mean function $\mu_1(T, \mathbf{Z}; \beta)$ and variance function $\Sigma_1(T, \mathbf{Z})$ as those of the response variable. Therefore, β can be estimated from the $\delta = 1$ subgroup alone. Let s be the sample size of the $\delta = 1$ subgroup, and assume that $s/n \nrightarrow 0$ as $n \to \infty$.

Let $GEE^{*\beta}(\beta) = 0$ be a d_2-dimensional estimating equation about only β, where $GEE^{*\beta}(\beta) = \sum_{i=1}^{n} k(Y_i, \delta_i | T_i, \mathbf{Z}_i; \beta) = \sum_{i=1}^{n} k_i(\beta)$. The d_2-dimensional parametric function $k_i(\beta)$ is a conditionally unbiased function of Y, i.e., $E[k_i(\beta)|\delta_i, T_i, \mathbf{Z}_i] = 0$ under our assumption. A linear form of $k_i(\beta)$ is $k_i(\beta) = B_i(\beta)\delta_i[Y_i - \mu_1(T_i, \mathbf{Z}_i; \beta)]$. $B_i(\beta)$ can be any d_2-dimensional function satisfying conditions similar to those on $A_i(\psi)$ to ensure that the resulting estimator has good asymptotic properties. For example, $B_i(\beta)$ could be $\frac{\partial \mu_1(T_i, \mathbf{Z}_i; \beta)}{\partial \beta}$, $\frac{\partial \mu_1(T_i, \mathbf{Z}_i; \beta)}{\partial \beta} \mu_1(T_i, \mathbf{Z}_i; \beta)^{-1}$, or $\frac{\partial \mu_1(T_i, \mathbf{Z}_i; \beta)}{\partial \beta} \Sigma_1(T_i, \mathbf{Z}_i)^{-1}$. A conventional way to estimate β by $GEE^{*\beta}(\beta) = 0$ is to fit a quasi-Poisson regression or negative binomial regression for the $\delta = 1$ subgroup as a sample.

Let the type P extended GEE for the parameters (α, β, θ) be $GEE\text{-P}(\psi) = 0$, where $GEE\text{-P}(\psi) = (GEE^{\alpha}(\psi)', GEE^{*\beta}(\beta)', m(\tilde{\theta} - \theta)')'$. We call the solution of $GEE\text{-P}(\psi) = 0$ the type P extended GEE estimator, denoted $\tilde{\psi}_P = (\tilde{\alpha}_P, \tilde{\beta}_P, \tilde{\theta})$. The type P extended GEE estimator $\tilde{\psi}_P$ satisfies $GEE\text{-P}(\tilde{\psi}_P) = 0$. Rearrangement of the first-order Taylor expansion of the functions $GEE\text{-P}(\tilde{\psi}_P)$, whose values are equal to zero, yields the following equations:

$$\frac{1}{\sqrt{n}} \begin{bmatrix} GEE^{\alpha}(\psi) \\ GEE^{*\beta}(\beta) \\ m(\tilde{\theta} - \theta) \end{bmatrix} \doteq -\frac{1}{n} \begin{bmatrix} \frac{\partial GEE^{\alpha}(\psi)}{\partial \alpha} & \frac{\partial GEE^{\alpha}(\psi)}{\partial \beta} & \frac{\partial GEE^{\alpha}(\psi)}{\partial \theta} \\ 0 & \frac{\partial GEE^{*\beta}(\beta)}{\partial \beta} & 0 \\ 0 & 0 & -m \end{bmatrix} \cdot \sqrt{n} \begin{bmatrix} \tilde{\alpha}_P - \alpha \\ \tilde{\beta}_P - \beta \\ \tilde{\theta} - \theta \end{bmatrix}. \quad (3.9)$$

Under the conditions on $A_i(\psi)$ and $B_i(\beta)$, the estimating function $GEE\text{-P}(\psi)$ has asymptotic normality, and its first-order derivatives converge to a constant matrix asymptotically surely. Moreover, because $\tilde{\theta}$ is estimated from an independent sample, the left-hand side of (3.9) has asymptotic normality:

$$\frac{1}{\sqrt{n}} \begin{bmatrix} GEE^{\alpha}(\psi) \\ GEE^{*\beta}(\beta) \\ m(\tilde{\theta} - \theta) \end{bmatrix} \xrightarrow[n \to \infty]{d} N(0, \Phi_P), \text{ where } \Phi_P = \begin{pmatrix} \Phi_P^* & 0 \\ 0 & r^{-1} A V_{\tilde{\theta}}(\theta) \end{pmatrix},$$

and the first term on the right-hand side of (3.9) converges to a constant matrix Ψ_P asymptotically surely:

$$-\frac{1}{n} \begin{bmatrix} \frac{\partial GEE^{\alpha}(\psi)}{\partial \alpha} & \frac{\partial GEE^{\alpha}(\psi)}{\partial \beta} & \frac{\partial GEE^{\alpha}(\psi)}{\partial \theta} \\ 0 & \frac{\partial GEE^{*\beta}(\beta)}{\partial \beta} & 0 \\ 0 & 0 & -m \end{bmatrix} \xrightarrow[n \to \infty]{a.s.} \Psi_P. \quad (3.10)$$

Therefore, it is easy to derive the asymptotic distribution for the type P extended GEE estimator $\tilde{\psi}_P$ as:

$$\sqrt{n}\Big((\tilde{\alpha}_P, \tilde{\beta}_P, \tilde{\theta})' - (\alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Psi_P^{-1}\Phi_P(\Psi_P^{-1})'\Big).$$

The variance of $(GEE^{\alpha}(\psi), GEE^{*\beta}(\beta))$, Φ_P^*, can be consistently estimated by $\hat{\Phi}_P^* = \frac{1}{n}\sum_{i=1}^n [g_i^{\alpha}(\psi), k_i(\beta)][g_i^{\alpha}(\psi), k_i(\beta)]'$ with $\tilde{\psi}_P$ plugged in; then $\hat{\Phi}_P = \begin{bmatrix} \hat{\Phi}_P^* & 0 \\ 0 & \frac{m}{n}\widehat{AV_{\tilde{\theta}}(\theta)} \end{bmatrix}$. Moreover, Ψ_P can be estimated by the left-hand side of (3.10) with $\tilde{\psi}_P$ plugged in. The asymptotic variance of the type P extended GEE estimator can be evaluated by $\frac{1}{n}\hat{\Psi}_P^{-1}\hat{\Phi}_P(\hat{\Psi}_P^{-1})'$.

3.5 Another Class of Extended GEE Estimation

Another way to use the information on δ in GEE is to derive the mean function of Y given δ explicitly using Assumptions (i) and (ii) in Section 3.2 and the conditional probability $P(\delta = 1|\mathbf{Z}) = q(\mathbf{Z};\rho)$. The expectation of Y conditional on δ, T, and $\mathbf{Z}$, $E(Y|\delta, T, \mathbf{Z}; \rho, \alpha, \beta, \theta)$, denoted $\Lambda^*(\delta, T, \mathbf{Z}; \rho, \psi)$, can be derived as

$$\Lambda^*(\delta, T, \mathbf{Z}; \rho, \psi) = \mu_1(T, \mathbf{Z}; \beta)[\frac{p(\mathbf{Z};\alpha) - q(\mathbf{Z};\rho)}{1 - q(\mathbf{Z};\rho)}]^{1-\delta} + (1-\delta)\mu_0(T, \mathbf{Z}; \theta)\frac{1 - p(\mathbf{Z};\alpha)}{1 - q(\mathbf{Z};\rho)}. \tag{3.11}$$

Let the variance function of Y given δ, T, and $\mathbf{Z}$ be $\Sigma^*(\delta, T, \mathbf{Z})$. Another set of GEEs for the parameters (α, β, θ) is based on the mean function of Y given δ, T, and $\mathbf{Z}$, $\Lambda^*(\delta, T, \mathbf{Z}; \rho, \psi)$, called GEE_2. As in Section 3.3, we plug $\hat{\rho}_A$ into the mean function $\Lambda^*(\delta, T, \mathbf{Z}; \rho, \psi)$ in the GEE_2 procedures. Evaluated at $\hat{\rho}_A$, $\Lambda^*(\delta, T, \mathbf{Z}; \rho, \psi)$ remains a function of ψ only. The second set of d-dimensional GEEs for the parameters (α, β, θ) based on the mean function $\Lambda^*(\delta, T, \mathbf{Z}; \hat{\rho}_A, \psi)$ in (3.11) are denoted $GEE_2(\psi) = 0$, where $GEE_2(\psi) = \sum_{i=1}^n h(Y_i|\delta_i, T_i, \mathbf{Z}_i; \psi)$. The d-dimensional parametric function $h(Y_i|\delta_i, T_i, \mathbf{Z}_i; \psi)$, also written $h_i(\psi)$, is a conditionally unbiased function of Y, i.e., $E[h_i(\psi)|\delta_i, T_i, \mathbf{Z}_i] = 0$. A linear form of $h_i(\psi)$ is $h_i(\psi) = C_i(\psi)[Y_i - \Lambda^*(\delta_i, T_i, \mathbf{Z}_i; \hat{\rho}_A, \psi)]$, where $C_i(\psi)$ can be any d-dimensional function satisfying conditions similar to those on $A_i(\psi)$. For example, $C_i(\psi)$ could be $\frac{\partial\Lambda^*(\delta_i, T_i, \mathbf{Z}_i; \hat{\rho}_A, \psi)}{\partial\psi}$, $\frac{\partial\Lambda^*(\delta_i, T_i, \mathbf{Z}_i; \hat{\rho}_A, \psi)}{\partial\psi}\Lambda^*(\delta_i, T_i, \mathbf{Z}_i; \hat{\rho}_A, \psi)^{-1}$, or $\frac{\partial\Lambda^*(\delta_i, T_i, \mathbf{Z}_i; \hat{\rho}_A, \psi)}{\partial\psi}\Sigma^*(\delta_i, T_i, \mathbf{Z}_i)^{-1}$.

The $GEE_2(\psi)$ also consists of three parts: d_1-dimensional, d_2-dimensional, and d_3-dimensional estimating functions for α, β, and θ, respectively, i.e., $GEE_2(\psi) = (GEE_2^{\alpha}(\psi)', GEE_2^{\beta}(\psi)', GEE_2^{\theta}(\psi)')'$. We write $h_i(\psi) = (h_i^{\alpha}(\psi)', h_i^{\beta}(\psi)', h_i^{\theta}(\psi)')'$. Similarly to the type J and P extended GEE estimators in the last section, we propose two extended GEE_2 estimators below.

3.5.1 Type J Extended GEE$_2$ Estimator

The d-dimensional type J extended GEE$_2$ for the parameters (α, β, θ) is GEE_2-J$(\psi) = 0$, where GEE_2-J$(\psi) = (GEE_2^{\alpha}(\psi)', GEE_2^{\beta}(\psi)', m(\tilde{\theta} - \theta)')'$. The joint solution is the type J extended GEE$_2$ estimator, denoted $\tilde{\psi}_{2J} = (\tilde{\alpha}_{2J}, \tilde{\beta}_{2J}, \tilde{\theta})$. The type J extended GEE$_2$ estimator $\tilde{\psi}_{2J}$ satisfies GEE_2-J$(\tilde{\psi}_{2J}) = 0$. Rearrangement of the first-order Taylor expansion of the functions GEE_2-J$(\tilde{\psi}_{2J})$, whose values are equal to zero, yields the following equations:

$$\frac{1}{\sqrt{n}} \begin{bmatrix} GEE_2^{\alpha}(\psi) \\ GEE_2^{\beta}(\psi) \\ m(\tilde{\theta} - \theta) \end{bmatrix} \doteq -\frac{1}{n} \begin{bmatrix} \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \alpha} & \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \beta} & \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \theta} \\ \frac{\partial GEE_2^{\beta}(\psi)}{\partial \alpha} & \frac{\partial GEE_2^{\beta}(\psi)}{\partial \beta} & \frac{\partial GEE_2^{\beta}(\psi)}{\partial \theta} \\ 0 & 0 & -m \end{bmatrix} \cdot \sqrt{n} \begin{bmatrix} \tilde{\alpha}_{2J} - \alpha \\ \tilde{\beta}_{2J} - \beta \\ \tilde{\theta} - \theta \end{bmatrix}. \quad (3.12)$$

Given the conditions on $C_i(\psi)$, the estimating function $GEE_2(\psi)$ has asymptotic normality, and its first-order derivatives converge to a constant matrix asymptotically surely. Moreover, because the consistent estimator $\tilde{\theta}$ is estimated from an independent sample, the left-hand side of (3.12) has asymptotic normality:

$$\frac{1}{\sqrt{n}} \begin{bmatrix} GEE_2^{\alpha}(\psi) \\ GEE_2^{\beta}(\psi) \\ m(\tilde{\theta} - \theta) \end{bmatrix} \xrightarrow[n\to\infty]{d} N(0, \Phi_{2J}), \text{ where } \Phi_{2J} = \begin{pmatrix} \Phi_{2J}^{*} & 0 \\ 0 & r^{-1} AV_{\tilde{\theta}}(\theta) \end{pmatrix},$$

and the first term on the right-hand side of (3.12) converges to a constant matrix Ψ_{2J} asymptotically surely:

$$-\frac{1}{n} \begin{bmatrix} \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \alpha} & \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \beta} & \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \theta} \\ \frac{\partial GEE_2^{\beta}(\psi)}{\partial \alpha} & \frac{\partial GEE_2^{\beta}(\psi)}{\partial \beta} & \frac{\partial GEE_2^{\beta}(\psi)}{\partial \theta} \\ 0 & 0 & -m \end{bmatrix} \xrightarrow[n\to\infty]{a.s.} \Psi_{2J}. \quad (3.13)$$

Therefore, it is easy to derive the asymptotic distribution for the type J extended GEE$_2$ estimator $\tilde{\psi}_{2J}$ as:

$$\sqrt{n}\Big((\tilde{\alpha}_{2J}, \tilde{\beta}_{2J}, \tilde{\theta})' - (\alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Psi_{2J}^{-1} \Phi_{2J} (\Psi_{2J}^{-1})'\Big).$$

The variance of $(GEE_2^{\alpha}(\psi), GEE_2^{\beta}(\psi))$, Φ_{2J}^{*}, can be consistently estimated by $\hat{\Phi}_{2J}^{*} = \frac{1}{n}\sum_{i=1}^{n} [h_i^{\alpha}(\psi), h_i^{\beta}(\psi)][h_i^{\alpha}(\psi), h_i^{\beta}(\psi)]'$ with $\tilde{\psi}_{2J}$ plugged in; then $\hat{\Phi}_{2J} = \begin{bmatrix} \hat{\Phi}_{2J}^{*} & 0 \\ 0 & \frac{m}{n}\widehat{AV_s(\theta)} \end{bmatrix}$. Ψ_{2J} can be estimated by the left-hand side of (3.13) with $\tilde{\psi}_{2J}$ plugged in. The asymptotic variance of the type J extended GEE$_2$ estimator can be evaluated by $\frac{1}{n}\hat{\Psi}_{2J}^{-1}\hat{\Phi}_{2J}(\hat{\Psi}_{2J}^{-1})'$.

3.5.2 Type P Extended GEE$_2$ Estimator

As for the type P extended GEE in Section 3.4.2, the d_2-dimensional estimating equation $GEE_2^{\beta}(\psi) = 0$ in the type J extended GEE$_2$ is replaced by $GEE^{*\beta}(\beta) = 0$. The resulting estimating equations are called the type P extended GEE$_2$, GEE_2-P$(\psi) = 0$, where GEE_2-P$(\psi) = (GEE_2^{\alpha}(\psi)', GEE^{*\beta}(\beta)', m(\tilde{\theta}-\theta)')'$. Solving GEE_2-P$(\psi) = 0$ gives the type P extended GEE$_2$ estimator, denoted $\tilde{\psi}_{2P} = (\tilde{\alpha}_{2P}, \tilde{\beta}_P, \tilde{\theta})$. The type P extended GEE$_2$ estimator $\tilde{\psi}_{2P}$ satisfies GEE_2-P$(\tilde{\psi}_{2P}) = 0$. Rearrangement of the first-order Taylor expansion of the functions GEE_2-P$(\tilde{\psi}_{2P})$ yields the following equations:

$$\frac{1}{\sqrt{n}} \begin{bmatrix} GEE_2^{\alpha}(\psi) \\ GEE^{*\beta}(\beta) \\ m(\tilde{\theta}-\theta) \end{bmatrix} \doteq -\frac{1}{n} \begin{bmatrix} \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \alpha} & \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \beta} & \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \theta} \\ 0 & \frac{\partial GEE^{*\beta}(\beta)}{\partial \beta} & 0 \\ 0 & 0 & -m \end{bmatrix} \cdot \sqrt{n} \begin{bmatrix} \tilde{\alpha}_{2P}-\alpha \\ \tilde{\beta}_P-\beta \\ \tilde{\theta}-\theta \end{bmatrix}. \tag{3.14}$$

Under the conditions on $C_i(\psi)$ and $B_i(\beta)$, the estimating function $GEE_2^P(\psi)$ has asymptotic normality, and its first-order derivatives converge to a constant matrix asymptotically surely. Moreover, because $\tilde{\theta}$ is estimated from an independent sample, the left-hand side of (3.14) has asymptotic normality:

$$\frac{1}{\sqrt{n}} \begin{bmatrix} GEE_2^{\alpha}(\psi) \\ GEE^{*\beta}(\beta) \\ m(\tilde{\theta}-\theta) \end{bmatrix} \xrightarrow[n\to\infty]{d} N(0, \Phi_{2P}), \text{ where } \Phi_{2P} = \begin{pmatrix} \Phi_{2P}^* & 0 \\ 0 & r^{-1}AV_s(\theta) \end{pmatrix};$$

and the first term on the right-hand side of (3.14) converges to a constant matrix Ψ_{2P} asymptotically surely:

$$-\frac{1}{n} \begin{bmatrix} \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \alpha} & \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \beta} & \frac{\partial GEE_2^{\alpha}(\psi)}{\partial \theta} \\ 0 & \frac{\partial GEE^{*\beta}(\beta)}{\partial \beta} & 0 \\ 0 & 0 & -m \end{bmatrix} \xrightarrow[n\to\infty]{a.s.} \Psi_{2P}. \tag{3.15}$$

Therefore, it is easy to derive the asymptotic distribution for the type P extended GEE$_2$ estimator $\tilde{\psi}_{2P}$ as:

$$\sqrt{n}\Big((\tilde{\alpha}_{2P}, \tilde{\beta}_P, \tilde{\theta})' - (\alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Psi_{2P}^{-1}\Phi_{2P}(\Psi_{2P}^{-1})'\Big).$$

The variance of $(GEE_2^{\alpha}(\psi), GEE^{*\beta}(\beta))$, Φ_{2P}^*, can be consistently estimated by $\hat{\Phi}_{2P}^* = \frac{1}{n}\sum_{i=1}^n [h_i^{\alpha}(\psi), k_i(\beta)][h_i^{\alpha}(\psi), k_i(\beta)]'$ with $\tilde{\psi}_{2P}$ plugged in; then $\hat{\Phi}_{2P} = \begin{bmatrix} \hat{\Phi}_{2P}^* & 0 \\ 0 & \frac{m}{n}\widehat{AV_{\tilde{\theta}}(\theta)} \end{bmatrix}$. Moreover, Ψ_{2P} can be estimated by the left-hand side of (3.15) with $\tilde{\psi}_{2P}$ plugged in. The asymptotic variance of the type P extended GEE$_2$ estimator can be evaluated by $\frac{1}{n}\hat{\Psi}_{2P}^{-1}\hat{\Phi}_{2P}(\hat{\Psi}_{2P}^{-1})'$.

3.6 Simulation Study

We conduct Monte Carlo simulation studies to examine the finite-sample properties of the four likelihood-based approaches and the four extended GEE methods in terms of efficiency (*Simulation setting 1*) and robustness to distribution misspecification (*Simulation setting 2*) and the model Assumption (ii) in Section 3.2 (*Simulation setting 3*). The parametric specifications for $p(\mathbf{Z})$ and $\mu_\eta(T, \mathbf{Z})$ are the same logistic and loglinear regression models as in Chapter 2, (2.15) and (2.16).

3.6.1 Description of Data Generation

We simulated subject i in the primary data with sex_i, age_i, and T_i as in Section 2.4.1. The only difference is that the method for generating the latent indicator η_i is related to the indicator δ_i. For $i = 1, \ldots, n$,

- $\delta_i \overset{\text{indep}}{\sim} \text{Bin}(1, q_i)$, where $\text{logit}(q_i) = \rho_0 + \rho_1 sex_i + \rho_2 age_i$,
  $\rho = (-0.2, -1.5, -1)$, about 22% $\delta = 1$

- $\eta_i|\delta_i = 1 \equiv 1$
  $\eta_i|\delta_i = 0 \overset{\text{indep}}{\sim} \text{Bin}(1, \pi_i)$, where $\pi_i = \frac{p_i - q_i}{1 - q_i}$ and $\text{logit}(p_i) = \alpha_0 + \alpha_1 sex_i + \alpha_2 age_i$,
  $\alpha = (1, -1, -0.8)$, about 53% $\eta = 1$

The event counts Y_i were generated in the following three settings, designed to assess the efficiency, robustness to the Poisson distribution assumption, and robustness to the model Assumption (ii), respectively, of the estimators. The simulation data are generated as in Section 2.4.1 for simulation settings 1 and 2. Simulation setting 3 is new.

Simulation setting 3: Robustness to model Assumption (ii).

This simulation assesses the robustness of the extended GEE methods to the model Assumption (ii). Under that assumption, given $\eta = 1$, the value of δ does not affect the distribution of the response variable. Thus, $[Y|\eta = 1, T, \mathbf{Z}] = [Y|\eta = 1, \delta = 1, T, \mathbf{Z}] = [Y|\eta = 1, \delta = 0, T, \mathbf{Z}]$. In this setting, the mean function of the $\eta = 1$ class remains $\mu_1(T, \mathbf{Z}; \beta)$. However, within the $\eta = 1$ class, the subgroups $\delta = 1$ and $\delta = 0$ have different distributions of Y. Given $\eta = 1$, the mean function of the $\delta = 1$ subgroup differs from $\mu_1(T, \mathbf{Z}; \beta)$ with a random effect γ_{1i}, and the mean function of the $\delta = 0$ subgroup is determined. The random effect γ_{1i} is generated by $\gamma_{1i} \overset{iid}{\sim} N(0, 1)$ within the range $\big(\log(q_i/p_i), \log(p_i/q_i)\big)$ to ensure that γ_{0i} exists, since γ_{0i} is determined via $\gamma_{0i} = \log[\frac{p_i - \exp(\gamma_{1i})q_i}{p_i - q_i}]$. The event counts Y_i in the two latent classes were generated independently from Poisson distributions as follows:

For $\eta_i = 1$**:** $Y_i|\eta_i = 1, \delta_i = 1 \overset{indep}{\sim} \text{Poisson}\Big(\exp(\gamma_{1i})\mu_1(T_i, sex_i, age_i; \beta)\Big)$
$Y_i|\eta_i = 1, \delta_i = 0 \overset{indep}{\sim} \text{Poisson}\Big(\exp(\gamma_{0i})\mu_1(T_i, sex_i, age_i; \beta)\Big)$

For $\eta_i = 0$: $Y_i|\eta_i = 0 \overset{indep}{\sim} \text{Poisson}\Big(\mu_0(T_i, sex_i, age_i; \theta)\Big)$

The simulated general population data was also as in Section 2.4.1. Although the $\delta = 1$ group may no longer be a good representative for the $\eta = 1$ class, which may lead to biased estimation for β in the type P extended GEE estimators, we expect that the α estimates will still be consistent and efficient.

3.6.2 Simulation Outcomes

Tables 3.1 to 3.9 present summaries of the estimators and their corresponding robust standard error estimators for 250 Monte Carlo datasets generated under the experimental settings described above. We evaluated the MLE and the three pseudo-MLEs for the parameters in the LCM presented in Section 3.3 under simulation settings 1 and 2 to study their efficiency and robustness to distribution misspecification. We also evaluated the extended GEE-based estimators under simulation setting 3 to assess their robustness to model Assumption (ii). We computed $\tilde{\theta}$ and its asymptotic variance used in the type B and AB pseudo-MLEs and the extended GEE estimators using the R function *glm* following a quasi-Poisson regression based on the supplementary information. The evaluation of $\hat{\rho}_A$ via $l_2(\rho)$ alone was used in the type A and AB pseudo-MLEs; we also computed this using the R function *glm* to fit a quasi-binomial regression of the data δ in the cohort.

Tables 3.1 to 3.4 present the results for the MLE and the three pseudo-MLEs under simulation setting 1 and three versions of simulation setting 2, respectively. We implemented the likelihood-based procedures by (a) maximizing the observed data likelihood or pseudo-likelihood via an R optimization function and (b) applying the EM algorithm described in Appendix A. Because of the complication of the likelihood functions, direct maximization by the R optimization function did not give reliable results. The estimates presented in the tables and discussed below were evaluated via the EM algorithm.

Table 3.1 presents the efficiency study of the MLE and the three pseudo-MLEs. The asymptotic variance of the MLE was estimated by the conventional method, the inverse of the Fisher information matrix, since the model was correctly specified. The asymptotic variances of the pseudo-MLEs were estimated by the sandwich variance estimator presented in Section 3.3. The complexity of the implementation reduced as we moved from the MLE to the type A, type B, and type AB pseudo-MLEs, and the computation became less expensive too. All four estimators and their variance estimators were consistent, and the pseudo-MLEs had an efficiency similar to that of the standard MLE. With the estimates of ρ, the type A pseudo-MLE did not lose any efficiency compared to the MLE. With the estimates of θ from the supplementary information, the type B and AB pseudo-MLEs were more efficient in estimating the primary parameters (α, β) compared to the MLE and the type A pseudo-MLE, respectively.

Table 3.2 summarizes the performance of the MLE and the three pseudo-MLEs when only Y in the $\eta = 0$ class is misspecified. We can see the advantage of using the supplementary information to estimate θ. The type B and AB pseudo-MLEs were robust to this type of overdispersion. In contrast, when Y in the $\eta = 1$ class is misspecified, none of the likelihood-based estimators remain consistent. The estimates are especially biased in the α parameters; see Tables 3.3 and 3.4. Under simulation setting 2, the sandwich variance estimator for MLE is better; we use it in Tables 3.2 to 3.4. In the distribution misspecification cases, the robust variance estimation performs well for the MLE and 3 pseudo-MLEs. Overall, the likelihood-based approaches lack robustness to the distributional assumption, at least for the $\eta = 1$ class.

In the numerical analyses in this chapter, we set $A_i(\psi)$ to $\frac{\partial \Lambda(T_i, \mathbf{Z}_i; \psi)}{\partial \psi}$ and $C_i(\psi)$ to $\frac{\partial \Lambda^*(\delta_i, T_i, \mathbf{Z}_i; \rho, \psi)}{\partial \psi}$. We use the estimation of β based on the $\delta = 1$ subgroup data in the type P extended GEE and GEE_2 estimators. We compute it using the R function *glm* following a quasi-Poisson regression. We find the roots of the estimating equations $GEE(\psi) = 0$ and $GEE_2(\psi) = 0$ numerically using the R function *dfsane* in the package BB.

Tables 3.5 to 3.8 present the results for the extended GEE methods under simulation setting 1 and three cases of simulation setting 2, respectively. In these tables, $\tilde{\theta}$ estimated from the supplementary information corresponded to the simulation settings in Tables 3.1 to 3.4. These tables summarize the four extended GEE estimators and their asymptotic standard errors (see Sections 3.4 and 3.5). The type J extended GEE estimator had poor numerical convergence even when the algorithm started from the true parameter values. The estimates were especially biased when the $\eta = 1$ class was misspecified. The asymptotic standard error estimation was not consistent with or without overdispersion. The type J extended GEE estimator for the LCMs without a distributional assumption for each class and without extra model assumptions did not perform well numerically. The type J extended GEE_2 estimator was consistent and had similar performance to that of the type P extended GEE and GEE_2 estimators in Tables 3.5 and 3.6. However, it lost consistency when the $\eta = 1$ class was overdispersed (Tables 3.7 and 3.8). The starting values for the type P extended GEE and GEE_2 estimators did not affect the results. Without loss of generality, the results in Tables 3.5 to 3.9 start from all zeros. On the other hand, the type P extended GEE and GEE_2 estimators that use the δ information had similar performance. They were consistent and also robust to any distribution misspecification in simulation setting 2, although they were less efficient than the likelihood-based approaches.

Simulation setting 3 evaluates the robustness of the extended GEE methods to the model Assumption (ii) in Section 3.2. We estimated β from only the $\eta = 1$ class data and verified that the mean of the $\eta = 1$ class still followed $\mu_1(T, \mathbf{Z}; \beta)$ under simulation setting 3. Since the $\delta = 1$ subgroup data randomly varied from the mean function $\mu_1(T, \mathbf{Z}; \beta)$, Assumption (iii) was invalid. This resulted in biased β estimation in the type P extended GEE estimators, which applied Assumption (iii) explicitly. Therefore, we were concerned

only with the consistency and robustness of the α estimation. Table 3.9 presents the type P extended GEE and GEE_2 and the type J extended GEE_2 estimators. These three estimators had similar performance. The α estimates were slightly inaccurate. Given the relatively large standard errors in the extended GEE estimators, the true values can easily fall in the CIs. Therefore, provided the mean of the $\delta = 1$ subgroup is not too different from that of the $\eta = 1$ class, we can still get reasonable estimates for α. Other methods, e.g., the generalized method of moments (GMM). may improve on the distribution-free GEE approaches for LCMs. This will discuss at the end of Chapter 6.

3.7 Analysis II of CAYACS Physician Claims

We now analyze the count data presented in Section 2.5. The characteristics of the data in the cohort ($\mathcal{P}$: $n = 1609$) and the population ($\mathcal{Q}$: $m = 13793$) were presented in Table 2.4. Our descriptive analysis compared the visit patterns for the survivors with RSC before follow-up, those with RSC during follow-up, and the rest of the cohort. The results showed that the survivors with RSC have a much higher overall visit frequency than the rest of the cohort regardless of the timing of the RSC, and the rest of the cohort still has a higher visit frequency than the general population. These findings motivated our additional model assumptions in Section 3.2.

In principle, survivors with RSC at any time are at risk of ongoing problems. Let δ_0 and δ_T indicate the RSC status at the beginning and the end of the follow-up, respectively. We first used $\delta = \delta_0$ as the RSC status to fit the LCM on the CAYACS data using the MLE, pseudo-MLEs, and extended GEE estimators developed in this chapter; the results are presented in Section 3.7.1. In Section 3.7.2, we apply the models from Section 3.7.1 to predict the at-risk status of the survivors who experienced RSC during the follow-up, i.e., $\delta_0 = 0$ and $\delta_T = 1$. Their δ information was not 1 when the models were fitted, but they should be predicted as $\eta = 1$ according to our model assumption. This can to some extent validate our model and estimation procedures.

Let $\mathcal{P}^\delta$ with $\delta = 0, 1$, respectively, be the two subsets in the cohort with the corresponding δ values, i.e., $\mathcal{P}^1 \cup \mathcal{P}^0 = \mathcal{P}$. In the cohort ($\mathcal{P}$: $n = 1609$), $\mathcal{P}^1$ has sample size $s = 168$ and $\mathcal{P}^0$ has sample size $n - s = 1441$.

We wish to compare the visit counts for $\mathcal{Q}$, $\mathcal{P}$, and the two subsets $\mathcal{P}^0$ and $\mathcal{P}^1$. A plot of visit counts vs. observation length showed a roughly linear relationship. Therefore, visit counts/observation length = visit counts per year can also be used for the comparisons. Table 3.10 presents the contingency tables of the categorical covariates; and Table 3.11 presents the summary statistics of *age at entry*, *observation length*, *visit counts*, and *visit counts per year* for the four datasets. The descriptive statistics show that the distributions of visit counts and visit counts per year decrease from high to low in the order $\mathcal{P}^1$, $\mathcal{P}$, $\mathcal{P}^0$, and $\mathcal{Q}$. The $\delta = 1$ subset has much higher values than the $\delta = 0$ subset, although the $\delta = 1$

subset is only about 10% of the survivors. Taking the population as the reference for the not-at-risk class, the $\delta = 0$ subset contained some individuals from the at-risk class with higher visit frequencies. This confirmed our conjecture that the at-risk class in the cohort consists of the whole $\delta = 1$ subset and part of the $\delta = 0$ subset.

Table 3.10 presents the quasi-Poisson regression analysis for $\mathcal{Q}$, $\mathcal{P}$, $\mathcal{P}^0$, and $\mathcal{P}^1$. After we adjusted for the independent variables, the trend of visit counts in the four datasets was the same as that seen in the descriptive analysis. The subset $\mathcal{P}^1$, with about 10% of the sample, has about 44% more visits than $\mathcal{P}$ on average. The analysis also revealed a varying pattern of the frequency over time in the $\delta = 1$ subset $\mathcal{P}^1$: the estimated coefficient of $\log(T)$ was much smaller than 1, while the coefficients were about 1 for the other three. Thus, the mean of the visit counts did not increase proportionally to observation length in the $\delta = 1$ subset. It may be necessary to study the visit trend over time. The analysis also indicated that the visit counts were highly overdispersed, so the distribution-free extended GEE estimators may perform better for the CAYACS data than MLEs.

3.7.1 Extended GEE Analysis of Visit Counts Under a Latent Class Model

We have seen in simulation setting 2 that the likelihood-based methods lost consistency and provided biased estimation for the parameters in the LCM, especially those in the risk model, when counts in the $\eta = 1$ class were overdispersed. However, the likelihood-based estimates would not have large biases in the regression models for both latent classes. The CAYACS counts were highly overdispersed; see Table 3.12. The likelihood-based approaches may not be valid, but they can still be good reference sets. We analyzed the visit counts under the LCM using the type P extended GEE and GEE_2 estimators, as well as the MLE and the three pseudo-MLEs as a comparison.

Table 3.13 presents the MLE and the three pseudo-MLEs and their corresponding estimated standard errors. All the likelihood-based estimates are similar. They all identify *diagnosis in the 80s* and *radiation but no chemo* as significant factors in the risk model for η. Table 3.14 lists the type P extended GEE and GEE_2 estimators and their standard errors. In this table, β in the regression model of the at-risk class was estimated from the $\delta = 1$ subset data, and the estimates were close to the likelihood-based estimates for β. This confirmed our choice of the $\delta = 1$ subset as representatives to estimate the parameters in the regression model of the at-risk class. The two GEE estimators for α in the risk model were similar and identified the same set of significant risk factors, although they differed from the likelihood-based results. Both GEE methods identified *sex*, *diagnosis in the 80s*, *radiation but no chemo*, and *both chemo and radiation therapy* as the significant risk factors for the risk model. For example, female survivors diagnosed with cancer in the 80s and treated with radiation were much more likely to be in the at-risk class. An interesting finding was the opposite direction of the *age at entry* effects between the risk model for η and the two

regression models in each class. In both of the latent classes, survivors diagnosed at a later age tend to have more frequent visits, but the older survivors are less likely to be in the at-risk class.

In conclusion, the MLE, pseudo-MLEs, and extended GEE estimators gave similar β and θ estimates but differed in the α estimates. However, in general, they all identified the most significant factors for the risk model. The extended GEE estimates of the real data are more reliable, since the counts are overdispersed.

3.7.2 Application: Risk Classification and Prediction in the Survivor Cohort

The LCM can also be used for risk classification and prediction. This section uses the type AB pseudo MLE and type P extended GEE from Section 3.7.1 to classify the cohort into two categories, the at-risk and not-at-risk categories. In Section 3.7.1, we used $\delta = \delta_0$ for the model fitting. In the cohort, the group with $\delta_0 = 0$ and $\delta_T = 1$ (sample size $v = 69$) should be in the at-risk class, according to our model. We can compare the predictions for this group from the two estimators. This also provides a way to validate our model and estimation procedures.

After estimating the parameters (α, β, θ) in the LCM and the standard error of the estimator, we can evaluate consistent estimators of the conditional probability of η and their confidence intervals. For example, a consistent estimator of the probability of $\eta = 1$ given covariates $\mathbf{Z}$ can be calculated by plugging the estimates of α into the risk model, $\hat{P}(\eta = 1|\mathbf{Z}) = p(\mathbf{Z}; \hat{\alpha})$. This quantity estimates the probability of individuals with initial characteristics $\mathbf{Z}$ being in the at-risk class. Using Bayes' theorem, we can also get a consistent estimator for $P(\eta = 1|\mathbf{Z}, T, Y)$ by plugging the estimates of (α, β, θ) into the following formula:

$$P(\eta = 1|\mathbf{Z}, T, Y; \alpha, \beta, \theta) = \frac{[Y|\eta = 1, T, \mathbf{Z}; \beta]p(\mathbf{Z}; \alpha)}{[Y|\eta = 1, T, \mathbf{Z}; \beta]p(\mathbf{Z}; \alpha) + [Y|\eta = 0, T, \mathbf{Z}; \theta](1 - p(\mathbf{Z}; \alpha))}. \tag{3.16}$$

This evaluation needs the distribution of Y in both the latent classes. In the likelihood-based methods, the distribution of Y given η, T, and $\mathbf{Z}$ is assumed to be Poisson. In the extended GEE approaches, we use negative binomial regressions to approximate the distributions of Y in the two classes. The quantity $\hat{P}(\eta = 1|\mathbf{Z}, T, Y)$ evaluates the probability that a survivor is at risk with the subject-specific information Y and T observed after the follow-up, as well as initial characteristics $\mathbf{Z}$. For an individual survivor, a closed form of the CI for $\hat{P}(\eta = 1|\mathbf{Z}, T, Y)$ is not easily derived, given the complication of (3.16). However, since we have derived the asymptotic distributions of the parameter estimators, a CI for $\hat{P}(\eta = 1|\mathbf{Z}, T, Y)$ can be evaluated via a parametric bootstrap procedure.

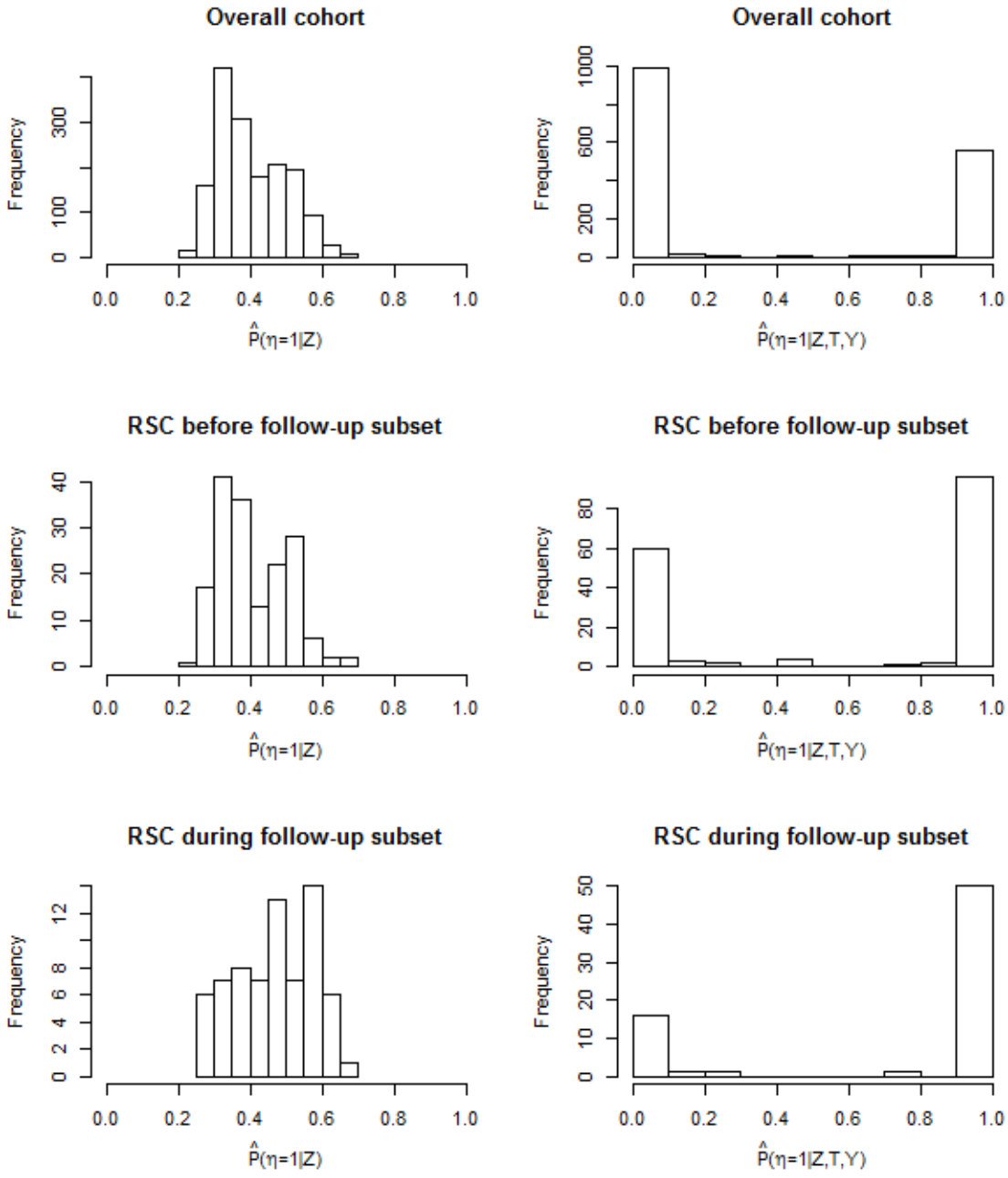

Figure 3.1: Risk probability estimations from the type AB pseudo-MLE.

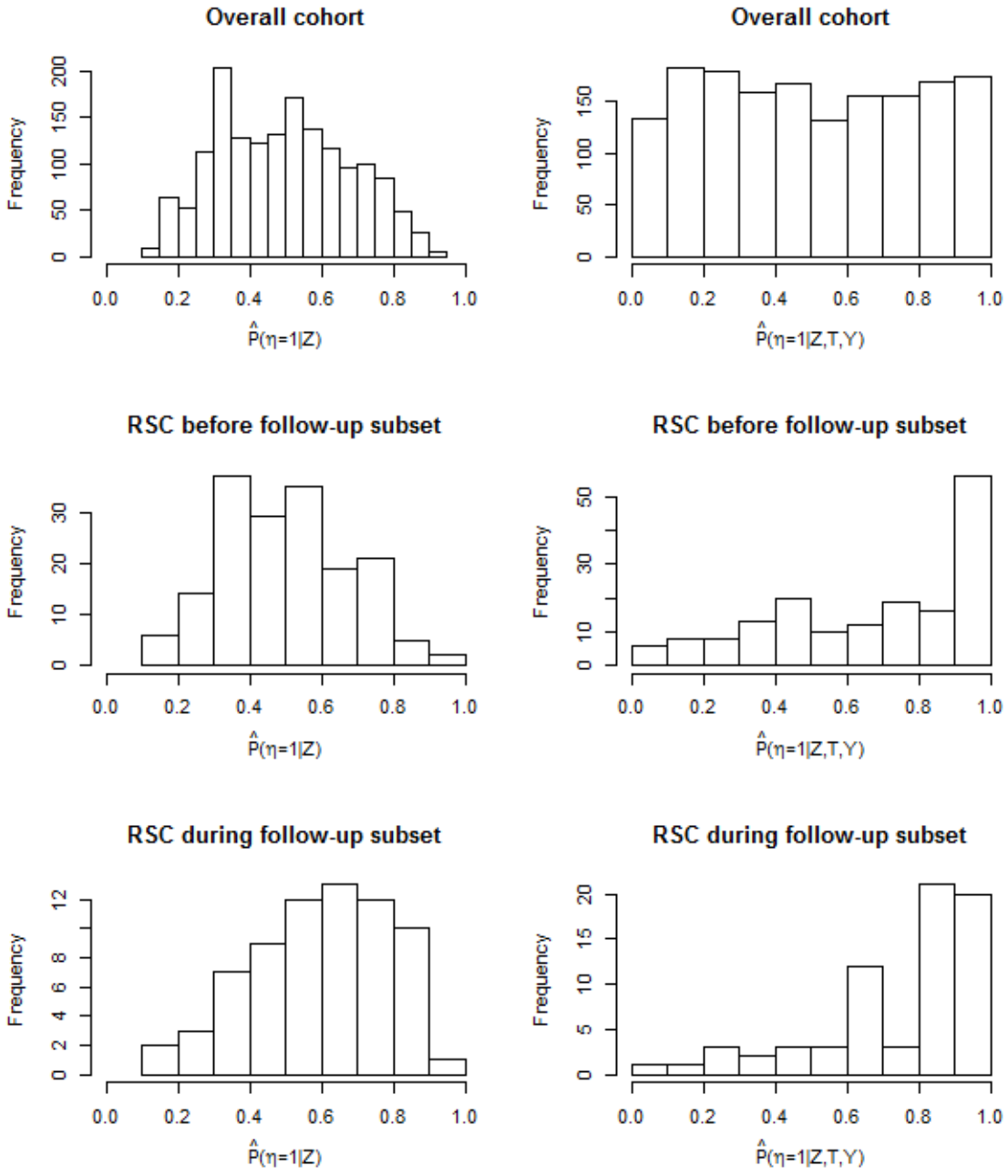

Figure 3.2: Risk probability estimations from the type P extended GEE estimator.

We used the type AB pseudo-MLE in the last column of Table 3.13 and the type P extended GEE in the first column of Table 3.14 to estimate $\hat{P}(\eta_i = 1|\mathbf{Z}_i)$ and $\hat{P}(\eta_i = 1|\mathbf{Z}_i, T_i, Y_i)$ for each individual in the cohort. Figures 3.1 and 3.2 give histograms of the estimated risk probabilities from the pseudo-MLE and the GEE estimator. The histograms on the left are for $\hat{P}(\eta_i = 1|\mathbf{Z}_i)$, and those on the right are for $\hat{P}(\eta_i = 1|\mathbf{Z}_i, T_i, Y_i)$. In the overall cohort, the estimated proportion (see Lin *et al.*, 2000) of subjects in the at-risk class, $\sum_{i=1}^{n} \hat{P}(\eta_i = 1|\mathbf{Z}_i)/n$, was about 40.3% for the pseudo-MLE and 49.5% for the GEE estimator. The first plots of Figures 3.1 and 3.2 also show that the GEE estimator provided a higher risk probability estimation. Given the overdispersed visit counts, the GEE estimation may be more reliable. Comparing the top right plots of Figures 3.1 and 3.2 shows that the estimated risk probabilities, $\hat{P}(\eta_i = 1|\mathbf{Z}_i, T_i, Y_i)$, from the GEE estimator were evenly distributed from 0 to 1, whereas those from the pseudo-MLE were extremely low or high. This is because the distribution of Y was approximated by the negative binomial to overcome the overdispersion in the GEE method. The middle rows of Figures 3.1 and 3.2 show the distributions of the estimated risk probabilities of the RSC before follow-up subset of the cohort ($s = 168$). The final rows of Figures 3.1 and 3.2 show the distributions of the estimated risk probabilities of another subset in the cohort, the RSC during follow-up group ($v = 69$). Compared to the distributions of the overall cohort, both of the estimated risk probabilities, $\hat{P}(\eta_i = 1|\mathbf{Z}_i)$ and $\hat{P}(\eta_i = 1|\mathbf{Z}_i, T_i, Y_i)$, of the subsets had a much higher rate of large values, especially the estimates from the GEE estimator in Figure 3.2.

We classify the cohort by the estimated risk probabilities $\hat{P}(\eta_i = 1|\mathbf{Z}_i)$ and $\hat{P}(\eta_i = 1|\mathbf{Z}_i, T_i, Y_i)$ at a cut-off value. A cut-off value can be chosen subjectively or based on expert opinions. We used the estimated proportions of subjects in the at-risk class, 0.4 and 0.5 as cut-off values, for the estimated risk probabilities from the pseudo-MLE and the GEE estimator, respectively. Table 3.15 summarizes the risk classification (at-risk and not-at-risk) vs. RSC status (no, before follow-up, and during follow-up). The first column uses the criterion that when $\hat{P}(\eta_i = 1|\mathbf{Z}_i) >$ cut-off, subject i is classified into the at-risk group, i.e., $\hat{\eta}_i = 1$; otherwise, $\hat{\eta}_i = 0$. The second column uses the criterion that when $\hat{P}(\eta_i = 1|\mathbf{Z}_i, T_i, Y_i) >$ cut-off, $\hat{\eta}_i = 1$; otherwise, $\hat{\eta}_i = 0$. Survivors with high risk characteristics or a high visit frequency can be at risk of later effects, so the last column used the criterion that when either $\hat{P}(\eta_i = 1|\mathbf{Z}_i)$ or $\hat{P}(\eta_i = 1|\mathbf{Z}_i, T_i, Y_i) >$ cut-off, $\hat{\eta}_i = 1$; otherwise, $\hat{\eta}_i = 0$. Comparing the risk predictions for the RSC during follow-up group shows that the false negative rates from the extended GEE estimator were always better than those from the pseudo-MLE, and the best false negative rate was only about 10% (7/69). This to some extent validated our model, estimating procedures, and the superiority of the extended GEE methods. The survivors with RSC during the follow-up who were not predicted to be at risk by the LCM had infrequent visits and did not have high risk characteristics. We suspect that is because the CAYACS data do not include oncologist visits.

3.8 Summary and Discussion

We have developed robust estimating procedures for the LCM. Our estimators extend the well-established GEE approach and are robust to distribution misspecification.

We introduced a binary variable δ as partial information about the latent risk indicator η, where $\delta = 1$ is a subgroup of $\eta = 1$. We proposed three pseudo-MLEs for the counts and compared them to the MLE under a mixture Poisson distribution. To obtain more robust statistical methods, we proposed two sets of extended GEEs for the parameters in the LCM. We developed two types of extended GEE estimators for each set of extended GEEs by using the supplementary dataset for the estimation of one class alone or together with the partial information about the other class. We derived the asymptotic properties of the pseudo-MLEs and the extended GEE estimators, and we estimated the variances using extended Huber sandwich estimators. We examined the finite-sample properties of the estimators for both efficiency and robustness. Our simulation studies verified that the pseudo-MLEs were efficient but lacked robustness to distribution misspecification, while the extended GEE estimators were robust to distribution misspecification and had satisfactory efficiency. We analyzed the CAYACS counts under the LCM using the estimators. We also performed an application of risk classification and prediction under the fitted LCM. Given the overdispersed counts, the type P GEE methods performed better than likelihood-based estimators in terms of predicting the RSC during follow-up group.

The type P extended GEE and GEE_2 estimators performed equally well in the simulation and the real-data analysis. Because its computation is simpler, we extend the favorable type P extended GEE method to longitudinal settings in Chapter 5 by adjusting for within-subject correlation. This will be referred to as the extended GEE estimator.

Table 3.1: Simulation Outcomes of MLE and Three Pseudo-MLEs. Setting 1: Efficiency Study

	(Primary data $n = 1000$; 250 repetitions)													
Parameter	ρ_0	ρ_1	ρ_2	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	-0.2	-1.5	-1	1	-1	-0.8	1.8	-0.6	-0.5	1	0.5	-0.3	-0.25	1
	MLE of $(\rho, \alpha, \beta, \theta)$													
sm*	-0.200	-1.495	-1.006	1.002	-1.004	-0.812	1.799	-0.597	-0.500	1.001	0.499	-0.296	-0.250	0.998
sse**	0.140	0.179	0.237	0.150	0.170	0.265	0.054	0.033	0.042	0.039	0.128	0.054	0.088	0.086
sm^*_{se}	0.146	0.175	0.257	0.161	0.183	0.270	0.055	0.034	0.044	0.039	0.129	0.060	0.091	0.087
	MLE of ρ from δ only			Type A pseudo-MLE of (α, β, θ)										
sm	-0.200	-1.495	-1.006	1.003	-1.004	-0.813	1.799	-0.597	-0.500	1.001	0.499	-0.296	-0.250	0.998
sse	0.140	0.180	0.238	0.152	0.170	0.268	0.054	0.033	0.042	0.039	0.128	0.055	0.089	0.087
sm_{se}	0.147	0.175	0.259	0.162	0.185	0.273	0.054	0.034	0.044	0.039	0.129	0.060	0.092	0.088
											Supplementary data $m = 5000$			
	Type B pseudo-MLE of (ρ, α, β)										MLE of θ			
sm	-0.200	-1.495	-1.005	1.002	-0.998	-0.808	1.799	-0.598	-0.500	1.001	0.499	-0.300	-0.249	1.001
sse	0.139	0.179	0.236	0.148	0.161	0.256	0.054	0.032	0.042	0.039	0.028	0.014	0.022	0.020
sm_{se}	0.146	0.175	0.256	0.156	0.165	0.255	0.054	0.032	0.043	0.038	0.029	0.014	0.022	0.021
				Type AB pseudo-MLE of (α, β)										
sm				1.003	-0.998	-0.809	1.799	-0.598	-0.500	1.001				
sse				0.149	0.160	0.258	0.054	0.032	0.042	0.039				
sm_{se}				0.157	0.165	0.255	0.054	0.032	0.043	0.038				

*The sample means of the parameter estimates (sm), the standard error estimates by the conventional approach for the MLE, and the robust standard error estimates for the three pseudo-MLEs (sm_{se}).

**The sample standard errors (sse) of the parameter estimates.

Table 3.2: Simulation Outcomes of MLE and Three Pseudo-MLEs. Setting 2: Case 1 with $\kappa = 1$

	(Primary data $n = 1000$; 250 repetitions)													
Parameter	ρ_0	ρ_1	ρ_2	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	-0.2	-1.5	-1	1	-1	-0.8	1.8	-0.6	-0.5	1	0.5	-0.3	-0.25	1
	MLE of $(\rho, \alpha, \beta, \theta)$													
sm*	-0.221	-1.499	-0.983	1.268	-0.609	-0.495	1.776	-0.601	-0.499	0.998	-0.225	-0.538	-0.415	1.083
sse**	0.145	0.173	0.274	0.151	0.170	0.253	0.058	0.034	0.048	0.041	0.328	0.169	0.280	0.221
sm^*_{se}	0.145	0.175	0.256	0.166	0.168	0.264	0.055	0.034	0.048	0.039	0.334	0.169	0.265	0.232
	MLE of ρ from δ only			Type A pseudo-MLE of (α, β, θ)										
sm	-0.210	-1.505	-1.004	1.273	-0.611	-0.502	1.776	-0.601	-0.499	0.998	-0.225	-0.538	-0.414	1.083
sse	0.149	0.174	0.278	0.151	0.170	0.253	0.058	0.034	0.048	0.041	0.328	0.169	0.280	0.221
sm_{se}	0.148	0.176	0.260	0.166	0.168	0.264	0.055	0.034	0.048	0.039	0.334	0.169	0.265	0.232
											Supplementary data $m = 5000$			
	Type B pseudo-MLE of (ρ, α, β)										GEE of θ			
sm	-0.210	-1.505	-1.003	1.189	-1.040	-0.822	1.797	-0.521	-0.449	0.987	0.495	-0.298	-0.249	1.002
sse	0.146	0.173	0.275	0.143	0.149	0.232	0.056	0.030	0.042	0.040	0.056	0.034	0.056	0.038
sm_{se}	0.147	0.176	0.258	0.158	0.155	0.245	0.053	0.030	0.042	0.038	0.067	0.032	0.050	0.047
				Type AB pseudo-MLE of (α, β)										
sm				1.189	-1.040	-0.823	1.797	-0.521	-0.449	0.987				
sse				0.144	0.149	0.234	0.056	0.030	0.042	0.040				
sm_{se}				0.158	0.155	0.246	0.053	0.030	0.042	0.038				

*The sample means of the parameter estimates (sm) and the robust standard error estimates for the MLE and the three pseudo-MLEs (sm_{se}). In the robustness study, the sandwich-form robust standard error estimates are consistent.

**The sample standard errors (sse) of the parameter estimates.

Table 3.3: Simulation Outcomes of MLE and Three Pseudo-MLEs. Setting 2: Case 2 with $\kappa = 1$

	(Primary data $n = 1000$; 250 repetitions)													
Parameter	ρ_0	ρ_1	ρ_2	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	-0.2	-1.5	-1	1	-1	-0.8	1.8	-0.6	-0.5	1	0.5	-0.3	-0.25	1
	MLE of $(\rho, \alpha, \beta, \theta)$													
sm*	-0.208	-1.499	-0.989	0.380	-1.167	-0.820	1.941	-0.399	-0.398	1.008	0.443	-0.318	-0.282	1.039
sse**	0.140	0.177	0.242	0.147	0.170	0.252	0.182	0.106	0.164	0.132	0.120	0.056	0.093	0.081
sm^*_{se}	0.147	0.175	0.258	0.151	0.167	0.257	0.176	0.114	0.175	0.130	0.117	0.058	0.092	0.081
	MLE of ρ from δ only			Type A pseudo-MLE of (α, β, θ)										
sm	-0.206	-1.500	-0.992	0.382	-1.168	-0.822	1.941	-0.399	-0.398	1.008	0.443	-0.318	-0.282	1.039
sse	0.140	0.177	0.245	0.146	0.169	0.252	0.182	0.106	0.164	0.132	0.120	0.055	0.093	0.081
sm_{se}	0.148	0.175	0.260	0.151	0.167	0.256	0.176	0.114	0.175	0.130	0.117	0.058	0.092	0.081
											Supplementary data $m = 5000$			
	Type B pseudo-MLE of (ρ, α, β)										GEE of θ			
sm	-0.206	-1.499	-0.992	0.384	-1.192	-0.846	1.943	-0.393	-0.393	1.006	0.500	-0.299	-0.250	0.999
sse	0.140	0.177	0.241	0.141	0.160	0.236	0.180	0.102	0.161	0.132	0.029	0.014	0.023	0.021
sm_{se}	0.147	0.175	0.258	0.147	0.159	0.247	0.175	0.111	0.172	0.130	0.029	0.014	0.022	0.021
				Type AB pseudo-MLE of (α, β)										
sm				0.385	-1.193	-0.847	1.943	-0.393	-0.393	1.006				
sse				0.142	0.160	0.238	0.180	0.102	0.161	0.132				
sm_{se}				0.148	0.159	0.247	0.175	0.111	0.172	0.130				

*The sample means of the parameter estimates (sm) and the robust standard error estimates for the MLE and the three pseudo-MLEs (sm_{se}). In the robustness study, the sandwich-form robust standard error estimates are consistent.

**The sample standard errors (sse) of the parameter estimates.

Table 3.4: Simulation outcomes of MLE and Three Pseudo-MLEs. Setting 2: Case 3 with $\kappa = 1$

	(Primary data $n = 1000$; 250 repetitions)													
Parameter	ρ_0	ρ_1	ρ_2	α_0	α_1	α_2	β_0	β_1	β_2	β_3	θ_0	θ_1	θ_2	θ_3
True value	-0.2	-1.5	-1	1	-1	-0.8	1.8	-0.6	-0.5	1	0.5	-0.3	-0.25	1
	MLE of $(\rho, \alpha, \beta, \theta)$													
sm*	-0.225	-1.493	-0.958	0.559	-0.798	-0.537	1.919	-0.471	-0.433	1.006	0.035	-0.449	-0.412	1.099
sse**	0.146	0.169	0.251	0.173	0.179	0.296	0.170	0.108	0.173	0.122	0.240	0.152	0.230	0.151
sm^*_{se}	0.145	0.175	0.254	0.169	0.191	0.296	0.168	0.109	0.173	0.125	0.251	0.149	0.236	0.174
	MLE of ρ from δ only			Type A pseudo-MLE of (α, β, θ)										
sm	-0.199	-1.507	-1.007	0.574	-0.805	-0.560	1.919	-0.471	-0.432	1.006	0.035	-0.449	-0.411	1.099
sse	0.150	0.170	0.258	0.172	0.179	0.293	0.170	0.108	0.173	0.122	0.240	0.152	0.230	0.150
sm_{se}	0.148	0.175	0.259	0.167	0.191	0.293	0.168	0.109	0.173	0.125	0.251	0.149	0.236	0.174
											Supplementary data $m = 5000$			
	Type B pseudo-MLE of (ρ, α, β)										GEE of θ			
sm	-0.215	-1.498	-0.977	0.499	-1.030	-0.739	1.932	-0.401	-0.385	1.003	0.491	-0.302	-0.246	1.006
sse	0.146	0.168	0.253	0.149	0.153	0.251	0.168	0.099	0.163	0.124	0.055	0.034	0.053	0.041
sm_{se}	0.145	0.175	0.255	0.149	0.156	0.245	0.165	0.099	0.162	0.124	0.067	0.032	0.050	0.047
				Type AB pseudo-MLE of (α, β)										
sm				0.509	-1.035	-0.756	1.932	-0.401	-0.385	1.003				
sse				0.149	0.154	0.251	0.168	0.099	0.163	0.124				
sm_{se}				0.149	0.156	0.246	0.165	0.099	0.162	0.124				

*The sample means of the parameter estimates (sm) and the robust standard error estimates for the MLE and the three pseudo-MLEs (sm_{se}). In the robustness study, the sandwich-form robust standard error estimates are consistent.

**The sample standard errors (sse) of the parameter estimates.

Table 3.5: Simulation Outcomes of Extended GEEs. Setting 1: Efficiency Study

	(Primary data $n = 1000$; 250 repetitions)						
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3
True value	1	-1	-0.8	1.8	-0.6	-0.5	1
	Type J extended GEE estimator of (α, β)						
sm	0.953	-1.046	-0.669	1.809	-0.572	-0.517	0.998
sse	0.220	0.235	0.232	0.131	0.128	0.116	0.087
sm_{se}	11.049	3.557	1.008	2.370	1.880	0.650	0.132
	Type P extended GEE estimator of (α, β)						
				From $\delta = 1$ subgroup			
sm	1.009	-0.998	-0.821	1.798	-0.603	-0.495	1.000
sse	0.209	0.227	0.341	0.075	0.050	0.060	0.052
sm_{se}	0.210	0.224	0.356	0.072	0.049	0.059	0.052
	Type J extended GEE_2 estimator of (α, β)						
sm	1.008	-0.998	-0.822	1.803	-0.604	-0.505	1.000
sse	0.222	0.226	0.345	0.080	0.054	0.065	0.058
sm_{se}	0.209	0.224	0.362	0.084	0.051	0.063	0.061
	Type P extended GEE_2 estimator of α						
				From $\delta = 1$ subgroup			
sm	1.007	-1.004	-0.823				
sse	0.223	0.225	0.341				
sm_{se}	0.207	0.220	0.354				

Table 3.6: Simulation Outcomes of Extended GEEs. Setting 2: Case 1 with $\kappa = 1$

	(Primary data $n = 1000$; 250 repetitions)						
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3
True value	1	-1	-0.8	1.8	-0.6	-0.5	1
	Type J extended GEE estimator of (α, β)						
sm	0.911	-1.048	-0.634	1.826	-0.562	-0.543	1.000
sse	0.327	0.242	0.301	0.145	0.156	0.129	0.093
sm_{se}	9.895	4.930	1.054	2.204	1.996	0.763	0.145
	Type P extended GEE estimator of (α, β)						
				From $\delta = 1$ subgroup			
sm	1.021	-1.014	-0.842	1.800	-0.602	-0.498	1.000
sse	0.231	0.249	0.429	0.077	0.048	0.059	0.055
sm_{se}	0.231	0.262	0.403	0.073	0.049	0.060	0.052
	Type J extended GEE_2 estimator of (α, β)						
sm	1.008	-1.023	-0.783	1.797	-0.602	-0.504	1.003
sse	0.234	0.258	0.416	0.094	0.051	0.064	0.068
sm_{se}	0.229	0.264	0.411	0.091	0.052	0.064	0.066
	Type P extended GEE_2 estimator of α						
				From $\delta = 1$ subgroup			
sm	1.009	-1.026	-0.792				
sse	0.234	0.255	0.412				
sm_{se}	0.227	0.259	0.400				

Table 3.7: Simulation Outcomes of Extended GEEs. Setting 2: Case 2 with $\kappa = 1$

	(Primary data $n = 1000$; 250 repetitions)						
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3
True value	1	-1	-0.8	1.8	-0.6	-0.5	1
	Type J extended GEE estimator of (α, β)						
sm	1.340	-0.961	-0.501	1.399	-0.587	-0.616	0.773
sse	2.176	0.608	0.713	1.636	0.340	0.511	1.000
sm_{se}	14.968	6.204	1.532	3.281	2.134	0.690	0.440
	Type P extended GEE estimator of (α, β)						
					From $\delta = 1$ subgroup		
sm	1.053	-0.975	-0.810	1.807	-0.602	-0.502	0.985
sse	0.496	0.608	0.865	0.234	0.182	0.250	0.174
sm_{se}	0.482	0.608	0.841	0.283	0.194	0.235	0.204
	Type J extended GEE_2 estimator of (α, β)						
sm	1.739	-0.869	-0.634	0.978	-0.706	-0.680	0.781
sse	2.569	0.878	1.194	2.246	0.368	0.647	0.885
sm_{se}	0.576	0.695	1.004	0.299	0.208	0.301	0.224
	Type P extended GEE_2 estimator of α						
					From $\delta = 1$ subgroup		
sm	1.085	-1.052	-0.838				
sse	0.519	0.610	0.779				
sm_{se}	0.507	0.610	0.860				

Table 3.8: Simulation Outcomes of Extended GEEs. Setting 2: Case 3 with $\kappa = 1$

	(Primary data $n = 1000$; 250 repetitions)						
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3
True value	1	-1	-0.8	1.8	-0.6	-0.5	1
	Type J extended GEE estimator of (α, β)						
sm	1.236	-0.963	-0.527	1.489	-0.585	-0.576	0.790
sse	2.074	0.579	0.694	1.516	0.370	0.485	0.936
sm_{se}	15.081	6.674	1.387	3.545	3.004	1.084	0.520
	Type P extended GEE estimator of (α, β)						
					From $\delta = 1$ subgroup		
sm	1.151	-1.077	-0.927	1.780	-0.575	-0.477	0.994
sse	0.602	0.588	0.915	0.238	0.179	0.237	0.191
sm_{se}	0.542	0.605	0.918	0.238	0.171	0.243	0.177
	Type J extended GEE_2 estimator of (α, β)						
sm	1.517	-0.793	-0.524	1.278	-0.691	-0.674	0.896
sse	2.285	0.860	1.230	2.157	0.314	0.668	0.768
sm_{se}	2.733	3.016	9.224	82.164	2.625	0.933	54.715
	Type P extended GEE_2 estimator of α						
					From $\delta = 1$ subgroup		
sm	1.083	-0.885	-0.761				
sse	0.907	0.884	0.911				
sm_{se}	0.493	0.674	0.891				

Table 3.9: Simulation Outcomes of Extended GEEs. Setting 3

	(Primary data $n = 1000$; 250 repetitions)						
Parameter	α_0	α_1	α_2	β_0	β_1	β_2	β_3
True value	1	-1	-0.8	1.8	-0.6	-0.5	1
	Type P extended GEE estimator of (α, β)						
				From $\delta = 1$ subgroup			
sm	0.819	-1.329	-0.900	1.836	-0.443	-0.437	0.999
sse	0.232	0.278	0.394	0.116	0.118	0.121	0.085
sm_{se}	0.231	0.252	0.394	0.136	0.088	0.111	0.098
	Type J extended GEE_2 estimator of (α, β)						
sm	0.830	-1.326	-0.903	1.831	-0.443	-0.440	1.002
sse	0.245	0.279	0.380	0.133	0.127	0.123	0.095
sm_{se}	0.226	0.257	0.390	0.127	0.119	0.118	0.093
	Type P extended GEE_2 estimator of α						
				From $\delta = 1$ subgroup			
sm	0.820	-1.326	-0.904				
sse	0.230	0.270	0.378				
sm_{se}	0.225	0.247	0.385				

Table 3.10: Summary of Categorical Variables for General Population, Survivor Cohort, and Cohort Subsets

Sex		SES		Diagnosis period		Treatment			
F	M	High	Low	80s	90s	Chemo no Rad	Rad no Chemo	Both	Others
				Survivor cohort $\mathcal{P}$: $n = 1609$					
708	901	659	950	649	960	660	139	402	408
				$\delta = 1$ subset $\mathcal{P}^1$: $s = 168$					
72	96	68	100	63	103	63	19	35	51
				$\delta = 0$ subset $\mathcal{P}^0$: $n - s = 1441$					
636	805	591	850	586	857	597	120	367	357
				General population $\mathcal{Q}$: $m = 13793$					
6009	7784	5124	8669						

Table 3.11: Summary of Continuous Variables for General Population, Survivor Cohort, and Cohort Subsets

			percentile				
	mean	SD	5th	25th	median	75th	95th
Survivor cohort $\mathcal{P}$: $n = 1609$							
age at entry	14.37	6.30	5.78	8.50	13.70	20.26	24.19
observation length (year)	9.70	5.43	2.19	5.05	9.22	14.00	19.09
visit counts	56.39	52.05	$<5^{\dagger}$	17	39	79	169
visit counts/year	6.34	7.28	0.86	2.58	4.62	7.71	16.83
$\delta = 1$ subset $\mathcal{P}^1$: $s = 168$							
age at entry	13.61	6.17	5.47	8.22	12.65	19.24	23.89
observation length (year)	8.37	5.54	0.86	3.29	8.06	12.06	18.23
visit counts	70.77	60.49	8	24	50	106	204.55
visit counts/year	12.26	13.55	1.25	4.49	7.72	14.82	40.68
$\delta = 0$ subset $\mathcal{P}^0$: $n - s = 1441$							
age at entry	14.46	6.31	5.84	8.54	13.93	20.34	24.22
observation length (year)	9.86	5.40	2.35	5.16	9.35	14.16	19.13
visit counts	54.71	50.73	$<5^{\dagger}$	17	37	77	168
visit counts/year	5.65	5.78	0.81	2.46	4.36	7.23	14.55
General population $\mathcal{P}$: $m = 13793$							
age at entry	14.07	6.17	5.71	8.39	13.05	19.96	23.89
observation length (year)	8.49	5.08	1.83	4.29	7.80	12.16	17.89
visit counts	27.88	27.93	$<5^{\dagger}$	6	18	42	89
visit counts/year	3.13	2.85	$<5^{\dagger}$	1.13	2.43	4.33	8.47

†Because of confidentiality concerns, counts below 5 are not displayed.

Table 3.12: Quasi-Poisson Regression with General Population, Survivor Cohort, and Cohort Subsets

	General Population $\mathcal{Q}$: $m = 13793$			Survivor Cohort $\mathcal{P}$: $n = 1609$		
Factor	*estimate*	*se*	*p-value*	*estimate*	*se*	*p-value*
intercept	1.032	0.031	$< .001$	2.176	0.095	$< .001$
male (vs. female)	-0.376	0.012	$< .001$	-0.360	0.038	$< .001$
age at entry	0.343	0.020	$< .001$	0.152	0.058	0.009
SES high (vs. low)	-0.023	0.013	0.07	-0.008	0.038	0.832
ln(time period)	1.089	0.012	$< .001$	0.876	0.035	$< .001$
dispersion parameter	14.67	0.24^a		31.78	1.84^a	

	$\delta = 0$ subset $\mathcal{P}^0$: $n - s = 1441$			$\delta = 1$ subset $\mathcal{P}^1$: $s = 168$		
Factor	*estimate*	*se*	*p-value*	*estimate*	*se*	*p-value*
intercept	1.904	0.101	$< .001$	3.144	0.212	$< .001$
male (vs. female)	-0.388	0.038	$< .001$	-0.200	0.110	0.072
age at entry	0.140	0.059	0.018	0.332	0.174	0.058
SES high (vs. low)	0.010	0.039	0.796	-0.109	0.113	0.339
ln(time period)	0.974	0.038	$< .001$	0.558	0.078	$< .001$
dispersion parameter	28.39	1.89^a		36.08	3.91^a	

[a]Standard error of dispersion parameters estimated by Bootstrap (B=1000).

Table 3.13: Analysis of CAYACS Data by Likelihood-based Approaches[a]

	MLE		type A pseudo-MLE		type B pseudo-MLE		type AB pseudo-MLE	
Factor	*estimate*	*sw.se*	*estimate*	*sw.se*	*estimate*	*sw.se*	*estimate*	*sw.se*
ρ estimates in the δ Model								
			from δ glm fit					
intercept	**-1.698**	(0.260)	**-1.778**	(0.253)	**-1.704**	(0.261)		
male (vs. female)	-0.042	(0.176)	0.037	(0.166)	-0.032	(0.178)		
age at entry	**-0.612**	(0.279)	**-0.554**	(0.272)	**-0.606**	(0.279)		
SES high (vs. low)	-0.029	(0.167)	-0.016	(0.167)	-0.032	(0.167)		
diagnosis in 90s (vs. 80s)	0.135	(0.173)	0.143	(0.175)	0.129	(0.174)		
treatment (vs. other)								
chemo but no rad	**-0.372**	(0.202)	**-0.385**	(0.204)	**-0.368**	(0.203)		
rad but no chemo	0.229	(0.297)	0.217	(0.299)	0.234	(0.298)		
both	**-0.415**	(0.235)	**-0.392**	(0.233)	**-0.417**	(0.235)		
α estimates in the Risk Model								
intercept	0.253	(0.179)	0.234	(0.179)	0.201	(0.174)	0.185	(0.174)
male (vs. female)	-0.169	(0.130)	-0.150	(0.130)	-0.209	(0.116)	**-0.192**	(0.116)
age at entry	-0.349	(0.215)	-0.339	(0.215)	**-0.456**	(0.189)	**-0.448**	(0.189)
SES high (vs. low)	0.131	(0.136)	0.135	(0.137)	0.137	(0.117)	0.141	(0.117)
diagnosis in 90s (vs. 80s)	**-0.690**	(0.121)	**-0.685**	(0.121)	**-0.592**	(0.116)	**-0.586**	(0.116)
treatment (vs. other)								
chemo but no rad	-0.192	(0.145)	-0.197	(0.145)	-0.127	(0.137)	-0.133	(0.137)
rad but no chemo	**0.550**	(0.215)	**0.546**	(0.215)	**0.464**	(0.208)	**0.460**	(0.209)
both	0.077	(0.152)	0.081	(0.151)	0.129	(0.148)	0.133	(0.148)
β estimates in the Regression Model for the At-risk Class								
intercept	**3.267**	(0.123)	**3.266**	(0.123)	**3.181**	(0.117)	**3.182**	(0.117)
male (vs. female)	**-0.252**	(0.051)	**-0.252**	(0.051)	**-0.232**	(0.047)	**-0.233**	(0.047)
age at entry	**0.152**	(0.080)	**0.152**	(0.080)	**0.185**	(0.071)	**0.185**	(0.071)
SES high (vs. low)	-0.050	(0.048)	-0.050	(0.048)	-0.050	(0.044)	-0.050	(0.044)
ln(time period)	**0.595**	(0.044)	**0.595**	(0.044)	**0.625**	(0.042)	**0.625**	(0.042)
					GEE estimates based on			
θ estimates in the Regression Model for the Not-at-risk Class					Supplementary Data: m=13793			
intercept	**1.423**	(0.104)	**1.423**	(0.104)	**1.032**	(0.031)		
male (vs. female)	**-0.415**	(0.051)	**-0.415**	(0.051)	**-0.376**	(0.012)		
age at entry	**0.233**	(0.077)	**0.234**	(0.077)	**0.343**	(0.020)		
SES high (vs. low)	-0.041	(0.054)	-0.041	(0.054)	-0.023	(0.013)		
ln(time period)	**0.938**	(0.041)	**0.938**	(0.041)	**1.089**	(0.012)		

[a]Significant effect with p-value ≤ 0.05 in **Boldface**.

Table 3.14: Analysis of CAYACS Data by Extended GEE methods[a]

	type P extended GEE		type P extended GEE_2	
Factor	*estimate*	*sw.se*	*estimate*	*sw.se*
α estimates in the Risk Model				
intercept	1.051	(0.646)	**1.251**	(0.685)
male (vs. female)	**-0.857**	(0.460)	**-0.806**	(0.449)
age at entry	-1.175	(0.740)	-1.223	(0.759)
SES high (vs. low)	0.332	(0.435)	0.369	(0.438)
diagnosis in 90s (vs. 80s)	**-0.813**	(0.333)	**-0.841**	(0.310)
treatment (vs. other)				
chemo but no rad	0.032	(0.240)	-0.122	(0.234)
rad but no chemo	**1.209**	(0.573)	**0.711**	(0.418)
both	**0.779**	(0.302)	**0.482**	(0.282)
β estimates in the Regression Model for the At-risk Class				
	GEE estimates based on			
	$\delta = 1$ subgroup: $s = 168$			
intercept	**3.144**	(0.212)		
male (vs. female)	**-0.200**	(0.110)		
age at entry	**0.332**	(0.174)		
SES high (vs. low)	-0.109	(0.113)		
ln(time period)	**0.558**	(0.078)		

[a]Significant effect with p-value ≤ 0.05 in **Boldface**.

Table 3.15: Comparison of Risk Classification and RSC status

Criterion	$\hat{P}(\eta_i = 1\|\mathbf{Z}_i) >$ cut-off		$\hat{P}(\eta_i = 1\|\mathbf{Z}_i, T_i, Y_i) >$ cut-off		either $>$ cut-off	
Classification	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$
RSC status	By type AB pseudo-MLE (cut-off = 0.4)					
No	786	586	935	437	594	778
Before follow-up	**95**	73	**65**	103	**39**	129
During follow-up	**21**	48	**18**	51	**8**	61
RSC status	By type P extended GEE (cut-off = 0.5)					
No	718	654	757	615	568	804
Before follow-up	**86**	82	**55**	113	**41**	127
During follow-up	**21**	48	**10**	59	**7**	62

Chapter 4

Analysis III of CAYACS Physician Claims: Conventional Longitudinal Analysis

Why do we choose longitudinal analysis for the CAYACS data? As described in Section 1.1 and depicted in Figure 1.1, the CAYACS program collected physician claims longitudinally over about 20 years for the survivor cohort and the general population. Our cross-sectional analyses have provided insight into the visit frequency for the population and the at-risk class in the cohort, and the risk assessment and classification of the cohort. However, such analyses cannot reveal the visit trend over time for each class. From the quasi-Poisson analysis in Table 3.12, we see that the visit trend over time (the coefficient of $\log T$) in the population is especially different from that of the RSC group in the cohort.

Longitudinal analysis is effective for studying change over time. It can distinguish variation in responses over time for one person from the overall variation in response (Diggle *et al.*, 2002). For the CAYACS data it allows us to distinguish variations in the visit pattern due to an individual's ageing from variations due to differences in individuals. Longitudinal studies on the CAYACS data also offer other benefits. First, they can handle time-varying covariates. For example, the survivors' SES potentially changes over time; in census years this information can be updated. Second, they can evaluate time-dependent effects. The effects of some time-independent covariates may change over time. For example, the sex effect may vary over time because women visit physicians more frequently during pregnancy. Third, they can conduct subject-specific inference with repeated outcomes from each subject. With cross-sectional data, averaging across people to overcome measurement error ignores natural differences between individuals. With repeated values, we can borrow strength across time for the person of interest and the group (Diggle *et al.*, 2002).

In this chapter and the next, we study the CAYACS yearly visit counts and yearly medical costs. This chapter analyzes the longitudinal data by conventional GEE approaches,

first for the cohort and the population separately for a descriptive comparison, and then for the combined data to test their differences formally. In Chapter 5, we will extend the extended GEE estimator for the LCM developed in Chapter 3 to longitudinal data and analyze the CAYACS data under LCMs.

4.1 CAYACS Longitudinal Data Structure and Description

4.1.1 Discrete Time Scales

The main advantage of a longitudinal study is its effectiveness for studying change over time (Diggle *et al.*, 2002). The choice of the time scale depends on the scientific objectives. For the CAYACS analysis, we use a discrete time unit of one year.

Figure 4.1 illustrates the calendar time scale vs. the individual time scale for hypothetical CAYACS individuals. The calendar time scale, based on the calendar year, is appropriate if, for example, we want to study resource usage by calendar year for administrative purposes. When the patients are the focus, the individual time scale is appropriate.

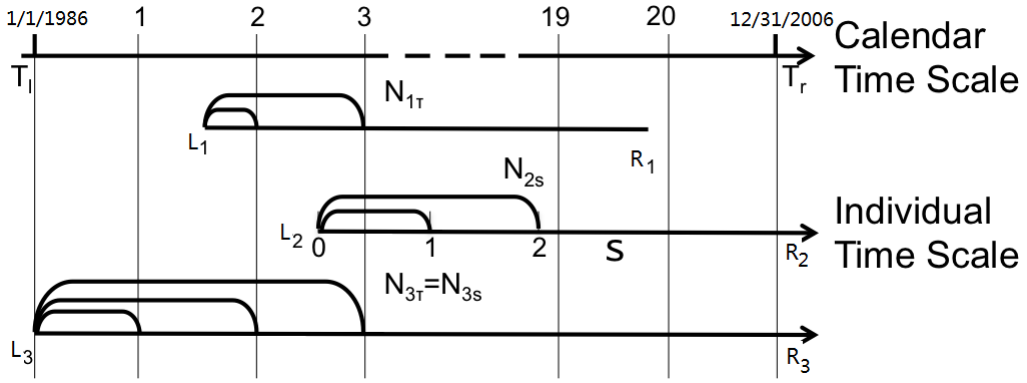

Figure 4.1: Calendar time scale vs. individual time scale for hypothetical individuals.

In the CAYACS data, we have two choices for the starting points of the individual time scale. The first is the individual's birthday, i.e., *age as time scale* according to the individual's age. This is a good choice if we want to compare the visit patterns between the cohort and the population at all ages. It may distinguish the effects of diagnosis age and ageing. As mentioned before, most of the matched population sample was followed from the age of five by the CAYACS program. With this starting point, we do not need to choose the starting points of the population according to the cohort. During the follow-up period, the ages of the subjects range from 5 to 46. However, the physician claims of the cohort are available only five years after diagnosis up to 20 years, and we want to study the patterns of survivors. The second choice of starting point is the start date of the individual's follow-up. We use this option, because it focuses on the visit pattern during follow-up. The starting

point of each individual in the population should be the start of follow-up of a matched individual from the cohort.

For the CAYACS data we assume noninformative censoring for the cohort, since the rate of death or departure from BC was quite low (about 13%). We also assume noninformative censoring for the population, since this rate was again quite low (about 15%).

4.1.2 Clean-up for Yearly Data

In this chapter and the next, the response $\mathbb{Y} = \mathbf{Y}$ is a vector of variables and $[\mathbb{Y}|\cdot] = [\mathbf{Y}|\cdot]$. Let N_{ij} represent subject i's visit count from the $(j-1)^{th}$ year to the j^{th}, where $j = 1, \ldots, J_i$ and J_i varies from individual to individual and $N_{i0} \equiv 0$. J_i is the cluster size of individual i. There is a medical cost corresponding to each physician visit. Let C_{ij} represent subject i's medical cost from the $(j-1)^{th}$ year to the j^{th} year; it is a nonnegative continuous variable, and $C_{ij} = 0$ whenever $N_{ij} = 0$.

We exclude individuals in the cohort with missing information on SES or initial treatment. We calculate full-year visit counts N_{ij} for each individual. We exclude individuals with less than one year of follow-up, and we exclude the last few observations if they cover less than one year, so J_i varies from 1 to 20. The resulting cohort has $n = 1609$ subjects and $N = \sum_{i=1}^{n} J_i = 15354$ yearly observations. Figure 4.2 compares the distributions of diagnosis year and cluster size J_i. The distributions are almost the same and are almost determined. For the population, we exclude individuals with missing SES information, and we calculate full-year visit counts N_{ij}. We exclude individuals with less than one year of follow-up, and we exclude the last few observations if they cover less than one year, so J_i varies from 1 to 20 and has a similar distribution to that of the cohort, after matching with the cohort; see Figure 4.3. The population has $m = 14289$ subjects and $M = \sum_{i=1}^{m} J_i = 124823$ yearly observations.

Each physician visit had a code indicating the fee paid by the government. Only about 1% of the visits had a missing code. We imputed the missing codes by the median values of the codes appearing in the cohort only, or the population only, or both the cohort and the population. We also calculated the full-year costs C_{ij} in Canadian dollars for the cohort and the population.

Let $\mathbf{Z}_i$ be a vector of the p covariates of the i^{th} subject. We are interested in how *sex*, *SES*, and *age at entry* (standardized as values from 0 to 1 in the regression) affect visit counts and costs, so $\mathbf{Z}_i = (\text{sex}_i, \text{SES}_i, \text{age}_i)'$. The cohort data $\mathcal{P} = \{(N_{ij}, C_{ij}, \mathbf{Z}_i) : i = 1, \ldots, n;\ j = 1, \ldots, J_i\}$ and the population data $\mathcal{Q} = \{(N_{ij}, C_{ij}, \mathbf{Z}_i) : i = 1, \ldots, m;\ j = 1, \ldots, J_i\}$ are summarized in Table 4.1.

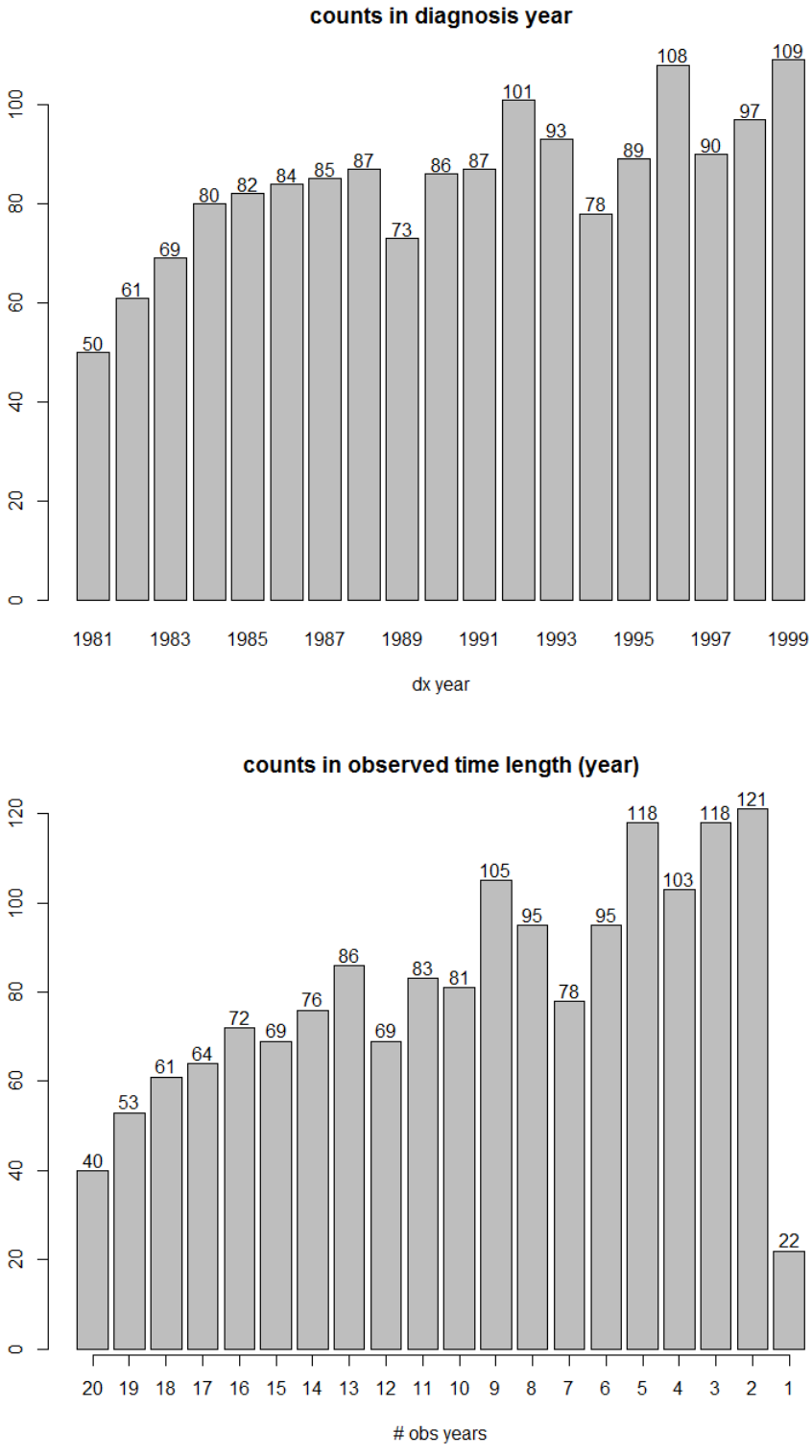

Figure 4.2: Survivor cohort yearly data: Diagnosis year vs. cluster size.

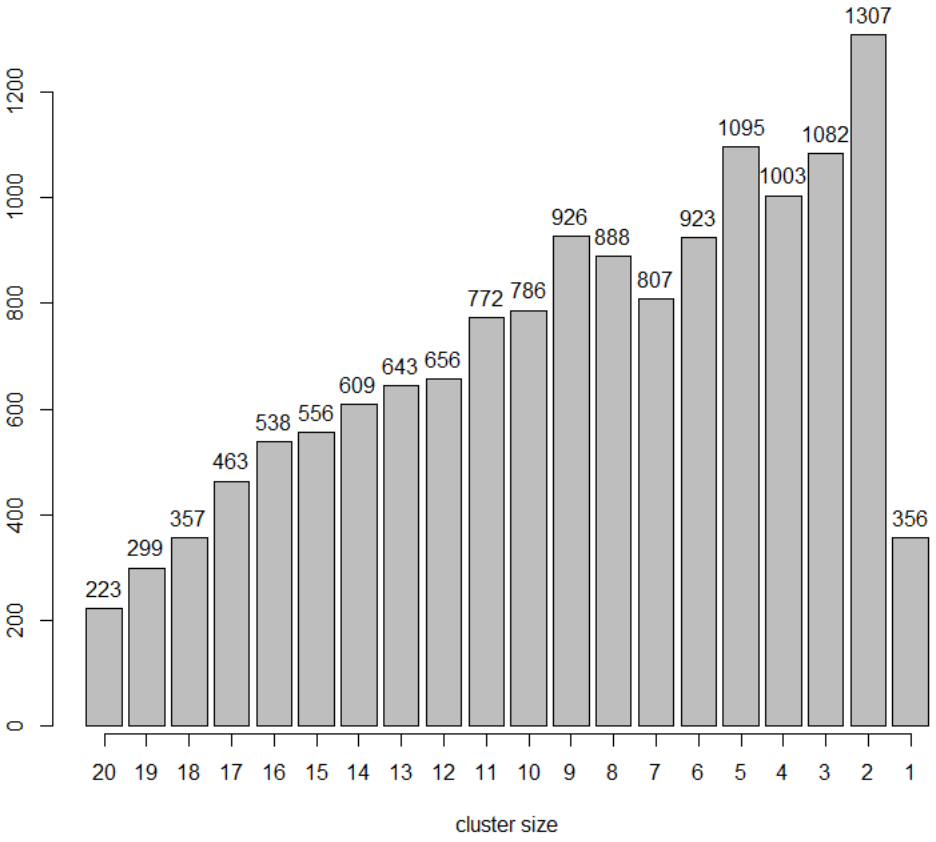

Figure 4.3: General population yearly data: Cluster size.

Table 4.1: Yearly Visit Data: Survivor Cohort vs. General Population

Sample	# subjects	# observations	mean(J_i)	mean(N_{ij})	mean(C_{ij})
Survivor Cohort $\mathcal{P}$	1609	15354	9.5	6.95	378.09
General Population $\mathcal{Q}$	14289	124823	8.7	4.44	194.89

4.1.3 Description of Yearly Visit Counts and Costs

Figure 4.4 shows the means and the corresponding CIs of the yearly visit counts in the twenty years of the follow-up for the cohort and the population. The numbers at the top and bottom of the figure are the sample sizes used to calculate the means and CIs. Over the twenty-year period, the sample sizes decrease and the CIs become wider. On average, the counts are much higher for the cohort than for the population. However, the cohort counts tend to decrease and the population counts tend to increase; the difference between them therefore decreases, especially in the final four years. Figure 4.5 shows the yearly costs during the follow-up period; the trends are similar.

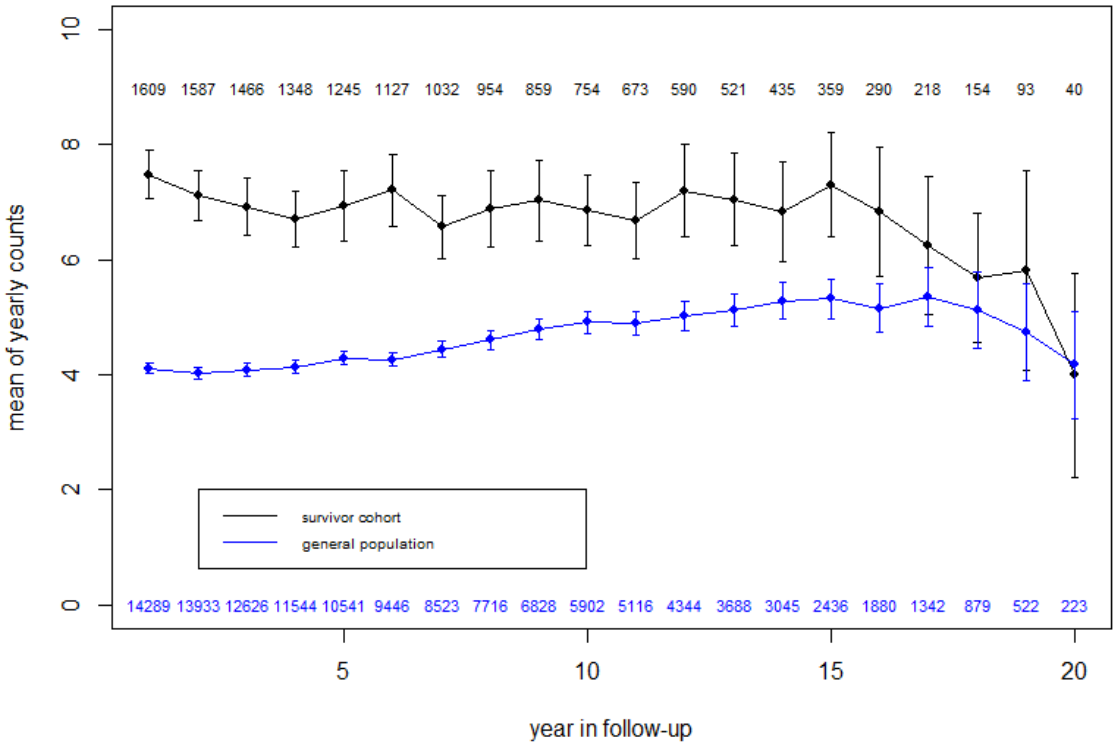

Figure 4.4: Mean and CI of yearly visit counts during follow-up: Survivor cohort vs. general population.

The medical costs are nonnegative continuous measurements. The rates of zero yearly costs are 12% and 23% in the cohort and population, respectively. For the model fitting, we perform a log-transformation of the yearly costs. We set $Y_{ij} = \log(C_{ij} + 5)$ to avoid log-transformation of zeros. The addition of 5 is somewhat arbitrary, but it does not affect the results. Figure 4.6 shows the yearly cost distribution before and after the transformation.

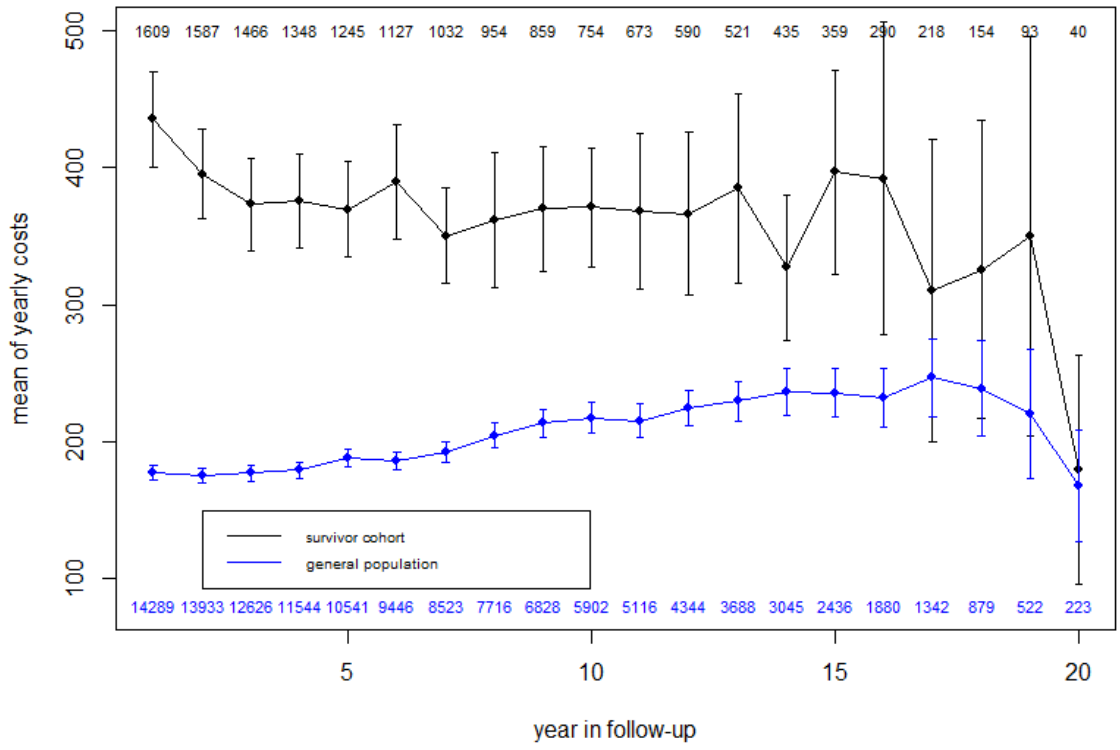

Figure 4.5: Mean and CI of yearly costs during follow-up: Survivor cohort vs. general population.

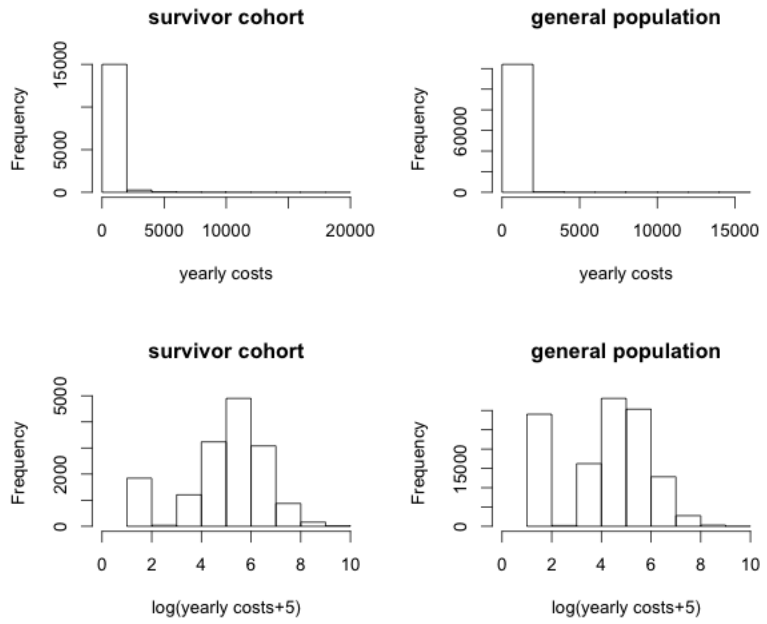

Figure 4.6: Distribution of yearly costs and log-transformed yearly costs.

4.2 Separate Analysis for Cohort and Population

To unify the model specifications, let Y_{ij} represent both the yearly visit counts and the transformed yearly costs. The yearly visit counts are overdispersed and the transformed yearly costs are not normally distributed; see Figure 4.6. Marginal models only need specifications of the mean and variance functions, but not a distributional assumption. Therefore, to study the effects of sex, SES, and age at entry on the longitudinal visit counts and costs, we fit regression models using GEEs to the yearly counts and yearly costs. We focus on inference for the population, not a specific individual. These analyses allow us to descriptively compare the visit patterns of the cohort $\mathcal{P} = \{(Y_{ij}, \mathbf{Z}_i) : i = 1, \ldots, n;\ j = 1, \ldots, J_i\}$ and the population $\mathcal{Q} = \{(Y_{ij}, \mathbf{Z}_i) : i = 1, \ldots, m;\ j = 1, \ldots, J_i\}$.

4.2.1 Mean and Variance Function Specification of Response Variables

Marginal models are designed to separately model the regression of $\mathbf{Y}$ on $\mathbf{Z}$ and the association among repeated observations of $\mathbf{Y}$ for each individual. In addition to modelling the effects of covariates on the expectation, we must specify a model for the association among the observations from each subject. Let the conditional expectation $E(Y_{ij}|\mathbf{Z}_i) = \mu(\mathbf{Z}_i; b_j)$ for the regression. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ_i})'$ be the J_i-dimensional response vector of subject i. The variance function $\mathrm{Var}(\mathbf{Y}_i|\mathbf{Z}_i)$ can be decomposed into $\mathrm{Var}(\mathbf{Y}_i|\mathbf{Z}_i) = T_i^{\frac{1}{2}}(b, \phi)\Gamma_i(\sigma)T_i^{\frac{1}{2}}(b, \phi)$. $T_i(b, \phi)$ is a $J_i \times J_i$ diagonal variance matrix, and the

j^{th} element on the diagonal is $\text{Var}(Y_{ij}|\mathbf{Z}_i)$, the structure of which can vary according to the data type. When Y_{ij} represents the yearly count, the variance function of Y_{ij} can be specified as $\text{Var}(Y_{ij}|\mathbf{Z}_i) = \phi\mu(\mathbf{Z}_i; b_j)$, where ϕ is a dispersion parameter that can also depend on $\mathbf{Z}_i$. When Y_{ij} is the transformed yearly cost, for example, we can specify $\text{Var}(Y_{ij}|\mathbf{Z}_i) = \phi_j$ instead, where the ϕ_j's are scale parameters and can be time-varying. $\Gamma_i(\sigma)$ is a $J_i \times J_i$ correlation matrix, and σ is a vector of correlation parameters, For example, it may be the compound symmetric (CS) structure, $\Gamma_i(\sigma) = \begin{pmatrix} 1 & \sigma & \cdots & \sigma \\ \sigma & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma \\ \sigma & \cdots & \sigma & 1 \end{pmatrix}$. It may also be independent or a time-series structure such as AR(1). Let $l(\cdot)$ be the link function: $l(\cdot) = \text{logit}(\cdot)$ for yearly binaries, $l(\cdot) = \log(\cdot)$ for yearly counts, and $l(\cdot) = I(\cdot)$ for yearly transformed costs.

In general, for longitudinal responses, the effects of time-independent covariates can be time-dependent. Therefore, we specify the mean function as follows:

$$l\{\mu(\mathbf{Z}_i; b_j)\} = b_{0j} + b_{1j}sex_i + b_{2j}SES_i + b_{3j}age_i. \tag{4.1}$$

We label the marginal models $Ma_0a_1a_2a_3.A$ based on the structure of the linear predictor in (4.1) and the structure of $\Gamma_i(\sigma)$. Here a_k indicates whether or not coefficient b_{kj} is time-dependent, for $k = 0, 1, 2, 3$; when $a_k = 0$, $b_{kj} \equiv b_k$. Moreover, A gives the correlation structure of $\Gamma_i(\sigma)$: $A = 0$ is independent correlation, $\Gamma_i(\sigma) = I_i$; $A = 1$ is CS correlation, $\text{Corr}(Y_{ij}, Y_{ij'}) = \sigma$, $j \neq j'$; and $A = 2$ is AR(1) correlation, $\text{Corr}(Y_{ij}, Y_{ij'}) = \sigma^{|j-j'|}$, $j \neq j'$. For example, the label *M1101.1* indicates that in (4.1) $b_{2j} \equiv b_2$; SES has a time-independent effect and intercept, sex, and age all have time-dependent effects on the longitudinal response; and the correlation of observations from the same subject has a CS structure.

There are $2^4 \times 3 = 48$ combinations of models. We investigated *M0000.0/1/2*, *M1000.0/1/2*, *M1100.0/1/2*, *M1001.0/1/2*, *M1101.0/1/2*, and *M1111.0/1/2* for both yearly counts and costs, and we separately fitted these models to the cohort and the population.

4.2.2 Results and Comparison of Yearly Counts

We now separately analyze the yearly visit counts for the cohort and the population and compare them graphically. When $Y_{ij} = N_{ij}$, the link function $l(\cdot) = \log(\cdot)$ and $\text{Var}(Y_{ij}|\mathbf{Z}_i) = \phi\mu(\mathbf{Z}_i; b_j)$ are used in the GEEs. Tables 4.2 and 4.3 present the results for *M0000.0/1/2*, *M1000.0/1/2*, *M1100.0/1/2*, *M1001.0/1/2*, *M1101.0/1/2*, and *M1111.0/1/2* for the cohort and the population, respectively, including the dispersion and correlation parameters. The tables present averages when the effects are time-varying. Both the constant effects and

the averaged time-varying effects are similar in all the models for both the cohort and the population. However, the time-varying effect is more informative when it changes over time, as shown in Figures 4.7 to 4.11.

Figure 4.7 plots the models from the second column of Table 4.2 for the cohort. Under CS correlation, there are time-varying effects in *M1000*, *M1100*, *M1001*, *M1101*, and *M1111* compared with *M0000* (in black) the effects are all constant. Figure 4.8 compares the three correlation structures under the mean model *M1101* (*M0000* in red); these are the models in the fifth row of Table 4.2. Figures 4.9 and 4.10 are the corresponding plots for the population.

Figures 4.7 and 4.9 show that the cohort and the population have similar time trends. The overall time effect is not obvious if only the intercept is time-dependent. However, we can see a clear curved time trend when sex and age at entry also have time-varying effects. Although the sex effect is always negative, i.e., males have fewer visits on average, the effect is especially stronger during the 10 to 17 follow-up years, because this is the pregnancy period for most of the female subjects. For the cohort, age at entry also measures the age of the cancer diagnosis. The time-varying intercept effects can be considered visit trends as a result of ageing for survivors at the youngest diagnosis age, and the time-varying age at entry effects add to the intercept effects over time for older survivors. The trends show that the subjects tend to visit physicians less at first and noticeably more a decade later, and this trend is less noteworthy for older subjects. Although the age at entry of the population was randomly chosen according to the cohort, the time-varying effects of intercept and age at entry reflect visit trends as a result of ageing at different starting ages for people without cancer. The effect of SES does not change significantly over time. Therefore, *M1101* is the preferred model.

The GEE estimator of the regression parameters remains consistent even when the correlation structure is misspecified, and the correct specification of the correlation improves the efficiency of the estimator (Liang and Zeger, 1986; Zeger and Liang, 1986). Figures 4.8 and 4.10 show the time-varying effects under the three different correlation structures for *M1101* for the cohort and the population respectively. The effects over time are similar for different correlation structures, and the CS correlation has the best efficiency.

Figure 4.11 compares the time-varying effects of intercept, sex, and age at entry for the cohort and the population under the chosen model *M1101.1* with CS correlation. The intercept of the cohort is always larger than that of the population, but the difference decreases over time. The time trends of the sex and age at entry effects are similar, although on average males and females are closer in the cohort.

4.2.3 Results and Comparison of Yearly Costs

We now separately analyze the yearly medical costs for the cohort and the population. When $Y_{ij} = \log(C_{ij} + 5)$, the identity link function $l(\cdot) = I(\cdot)$ and $\text{Var}(Y_{ij}|\mathbf{Z}_i) = \phi_j$ are used

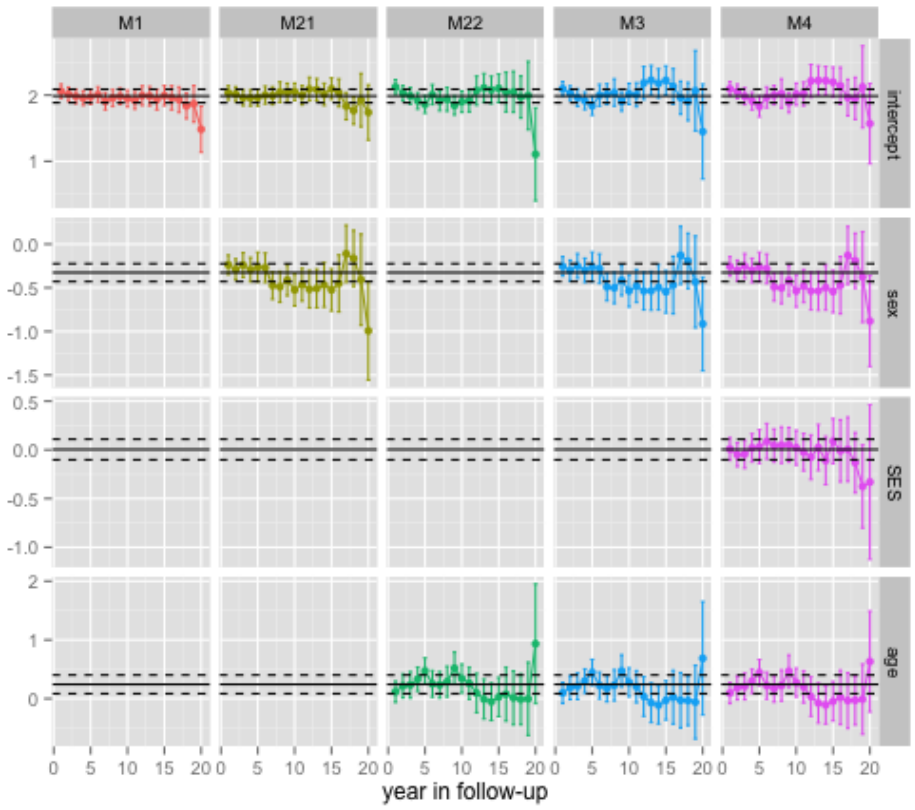

Figure 4.7: Time-dependent coefficients of survivor cohort yearly counts under CS correlation structure.

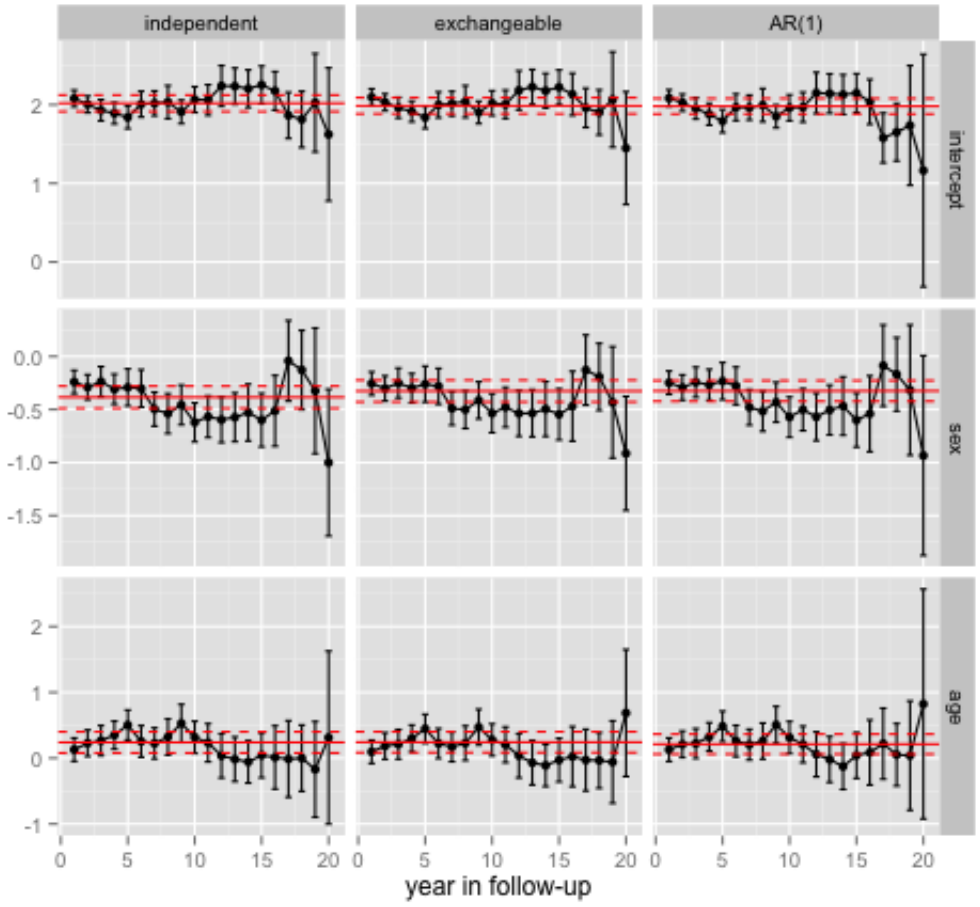

Figure 4.8: Time-dependent coefficients of survivor cohort yearly counts under *M1101*.

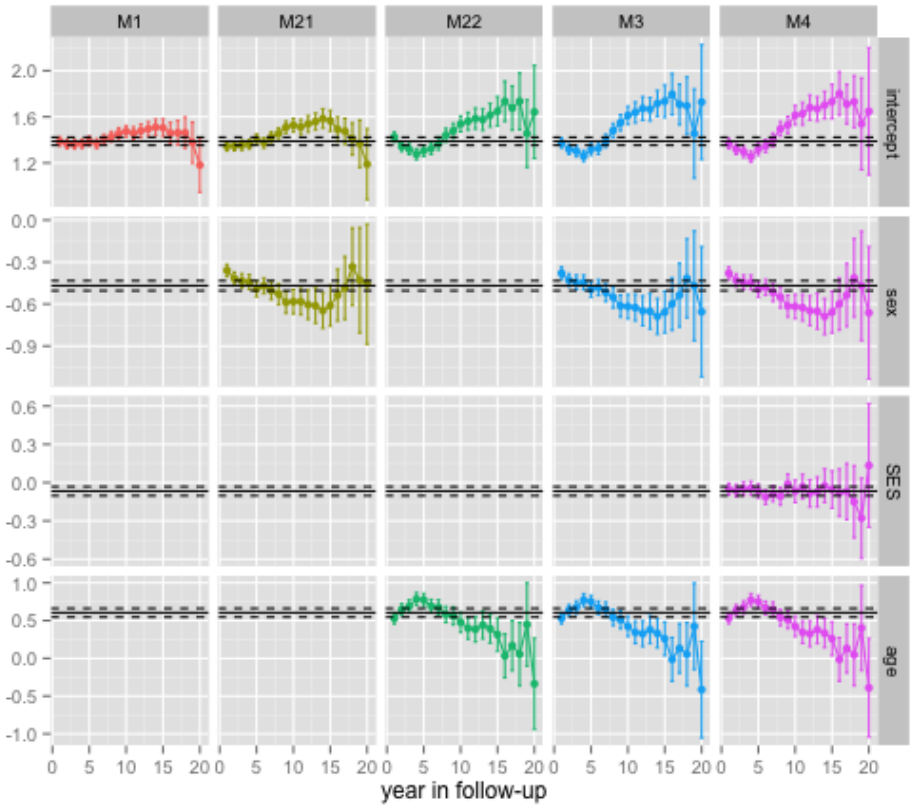

Figure 4.9: Time-dependent coefficients of general population yearly counts under CS correlation structure.

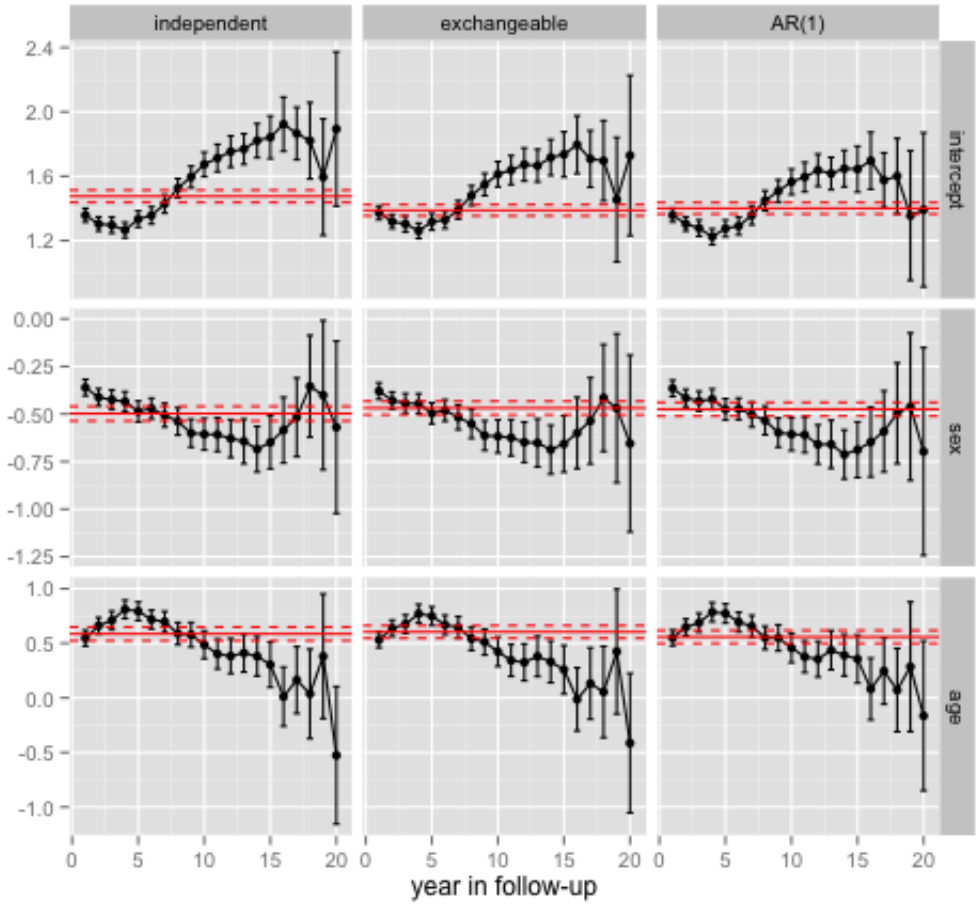

Figure 4.10: Time-dependent coefficients of general population yearly counts under *M1101*.

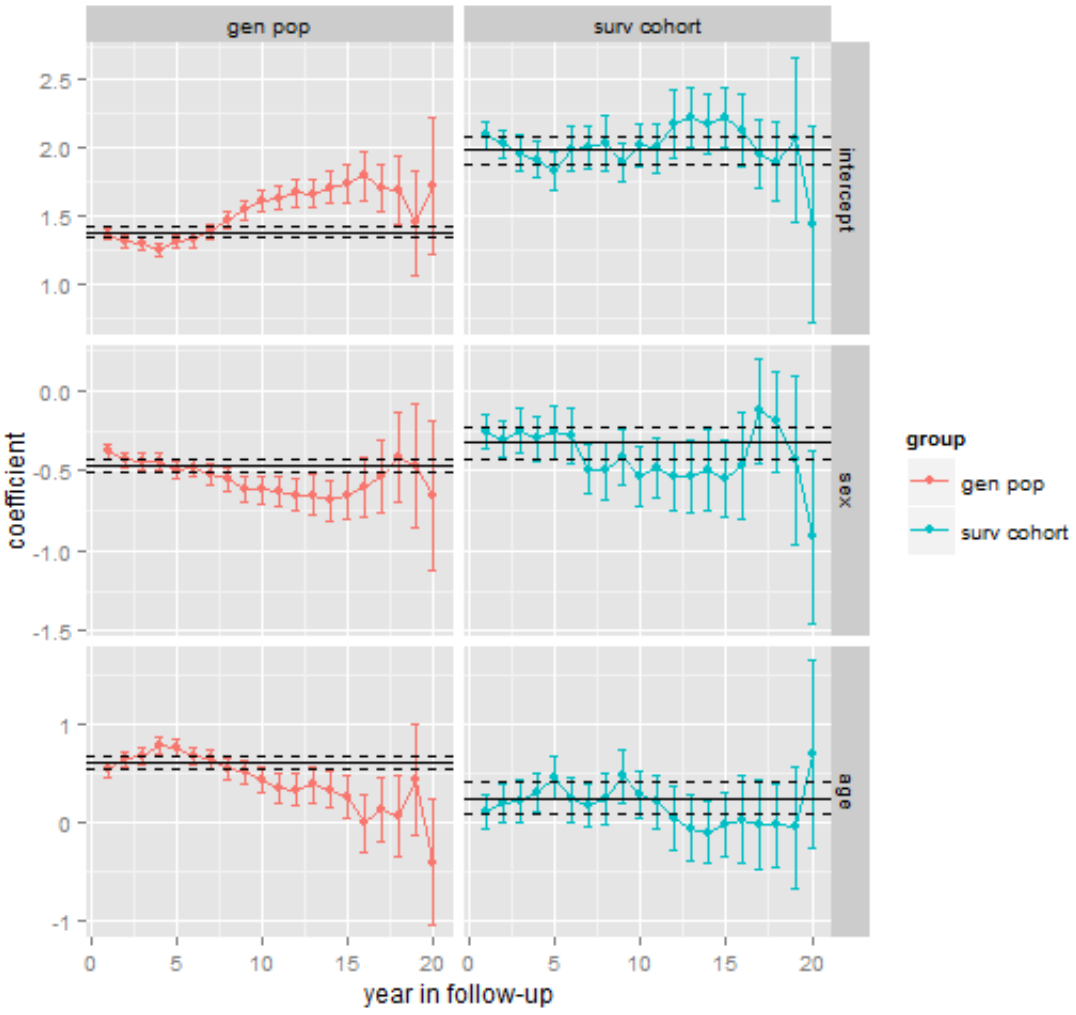

Figure 4.11: Time-dependent coefficients: SC vs. GP comparison for yearly counts under *M1101.1*.

in the GEEs, where the ϕ_j's are scale parameters and change over time. Tables 4.4 and 4.5 give the results for *M0000.0/1/2*, *M1000.0/1/2*, *M1100.0/1/2*, *M1001.0/1/2*, *M1101.0/1/2*, and *M1111.0/1/2* for the cohort and the population, respectively, including the scale and correlation parameters. The tables present averages when the effects are time-varying. Figures 4.12 to 4.16 show the time-varying effects over time. These figures correspond to Figures 4.7 to 4.11 for the yearly costs, with an extra row presenting the estimations of the time-varying scale parameters.

The trends are similar to those for the yearly counts in Section 4.2.2. The main difference is that the time-varying effects of age at entry for the cohort tend to increase from negative to positive over time. This means that older survivors tend to see physicians more frequently but cost less than younger survivors at the beginning, but they catch up the cost later. We can see this by comparing Figures 4.16 and 4.11, where the cohort and the population are contrasted under *M1101.1*.

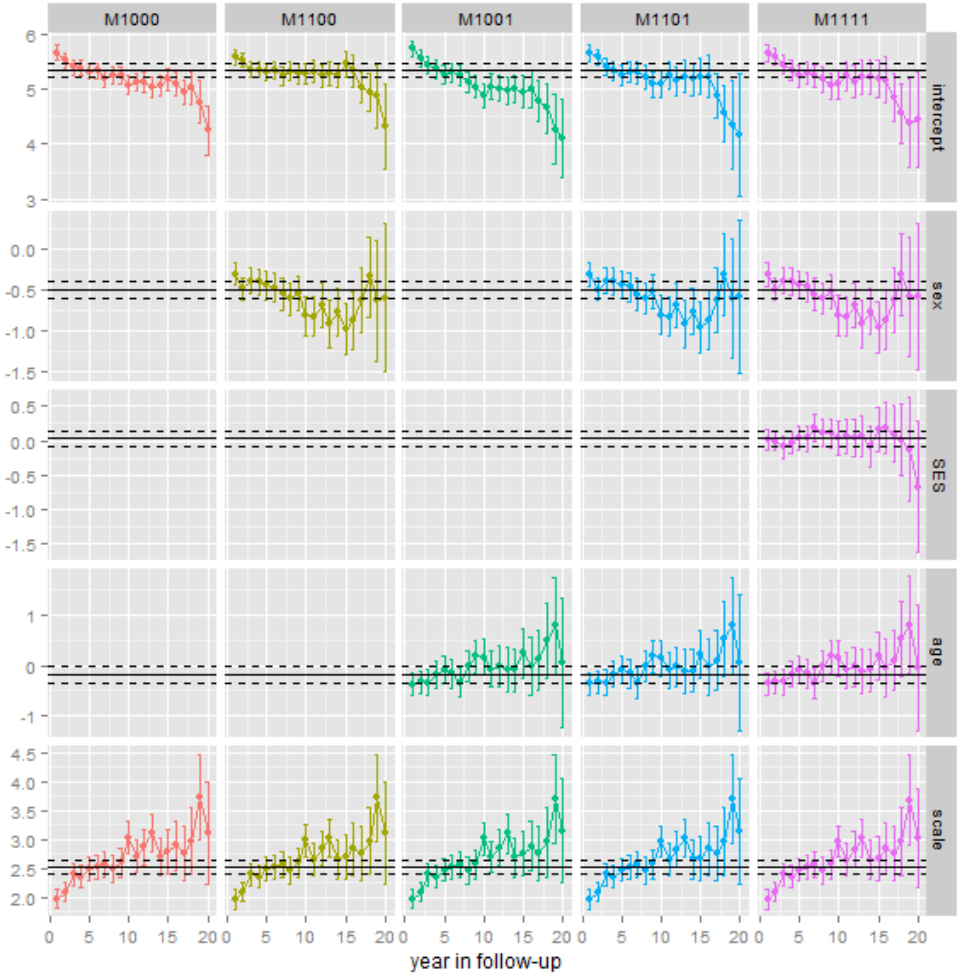

Figure 4.12: Time-dependent coefficients of survivor cohort yearly costs under CS correlation structure.

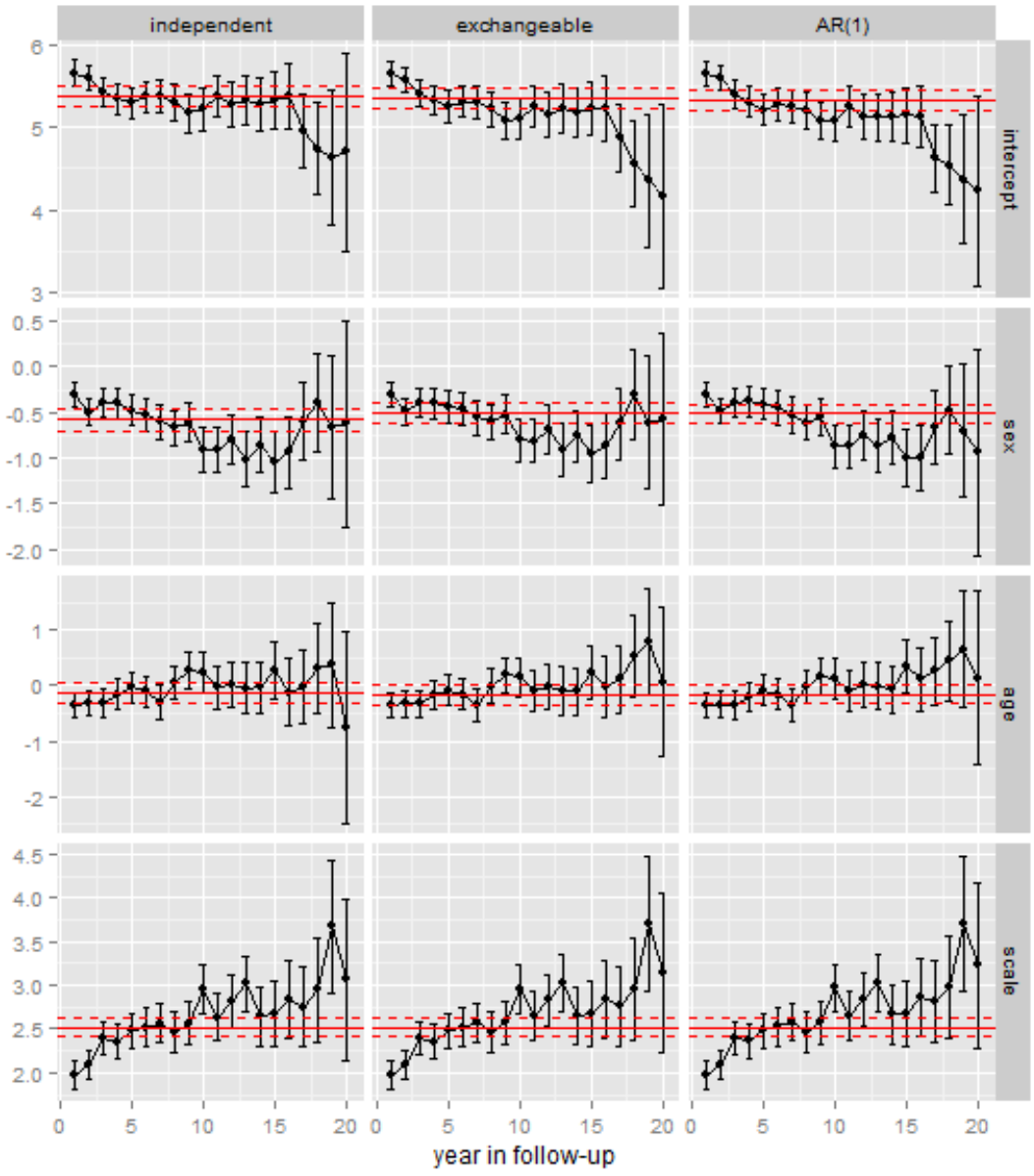

Figure 4.13: Time-dependent coefficients of survivor cohort yearly costs under *M1101*.

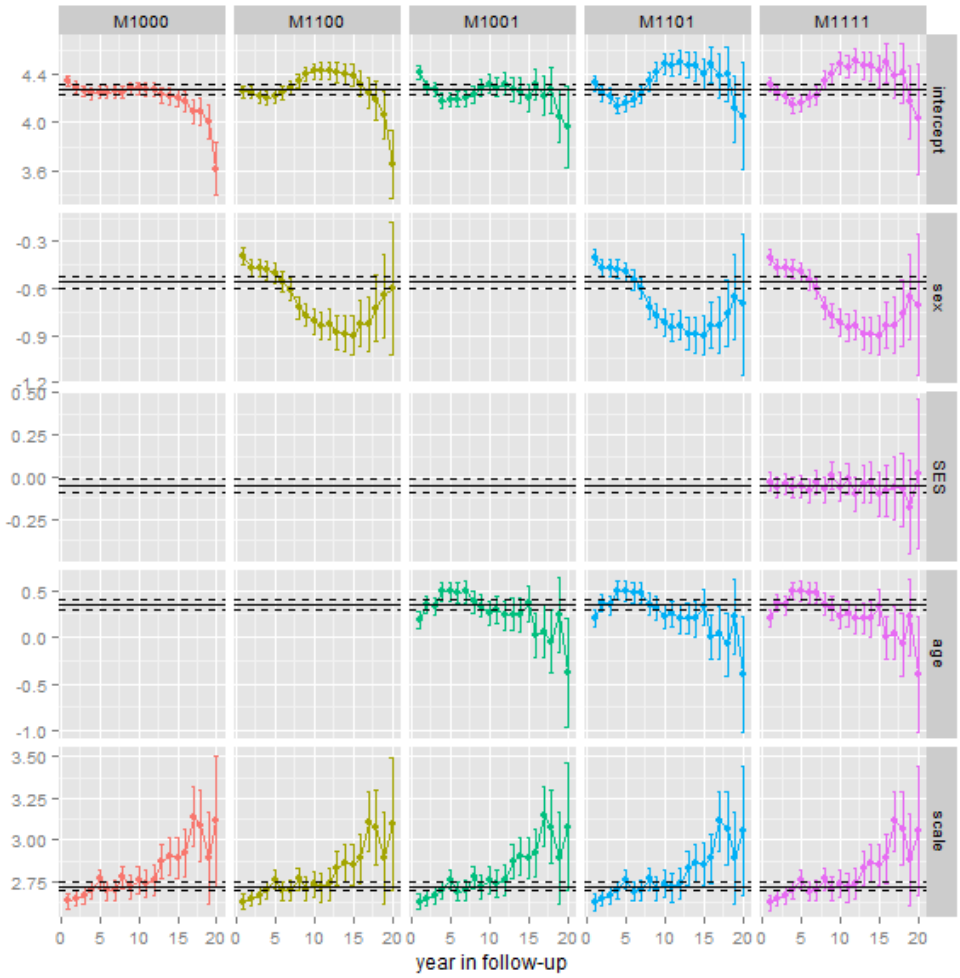

Figure 4.14: Time-dependent coefficients of general population costs under CS correlation structure.

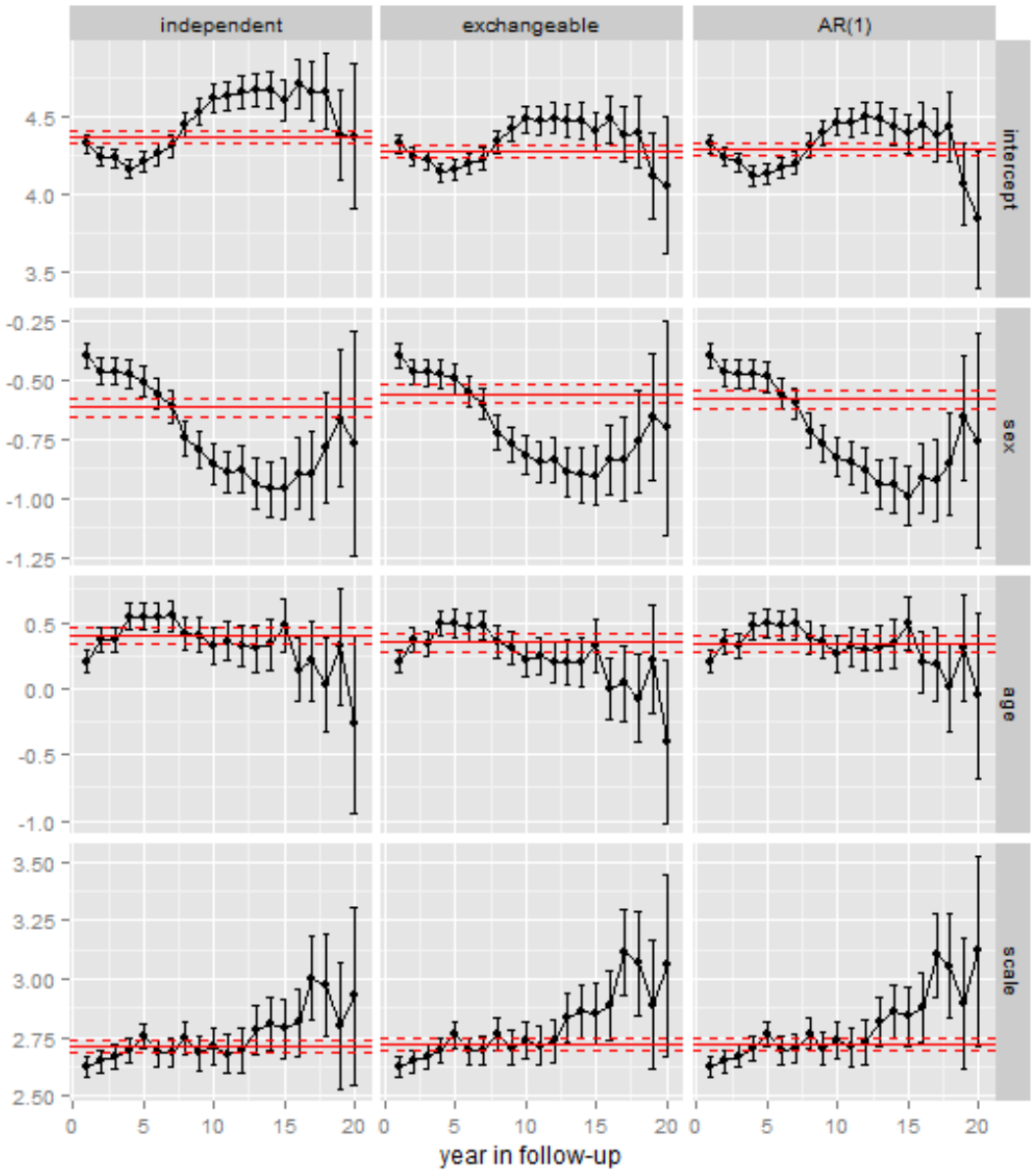

Figure 4.15: Time-dependent coefficients of general population costs under *M1101*.

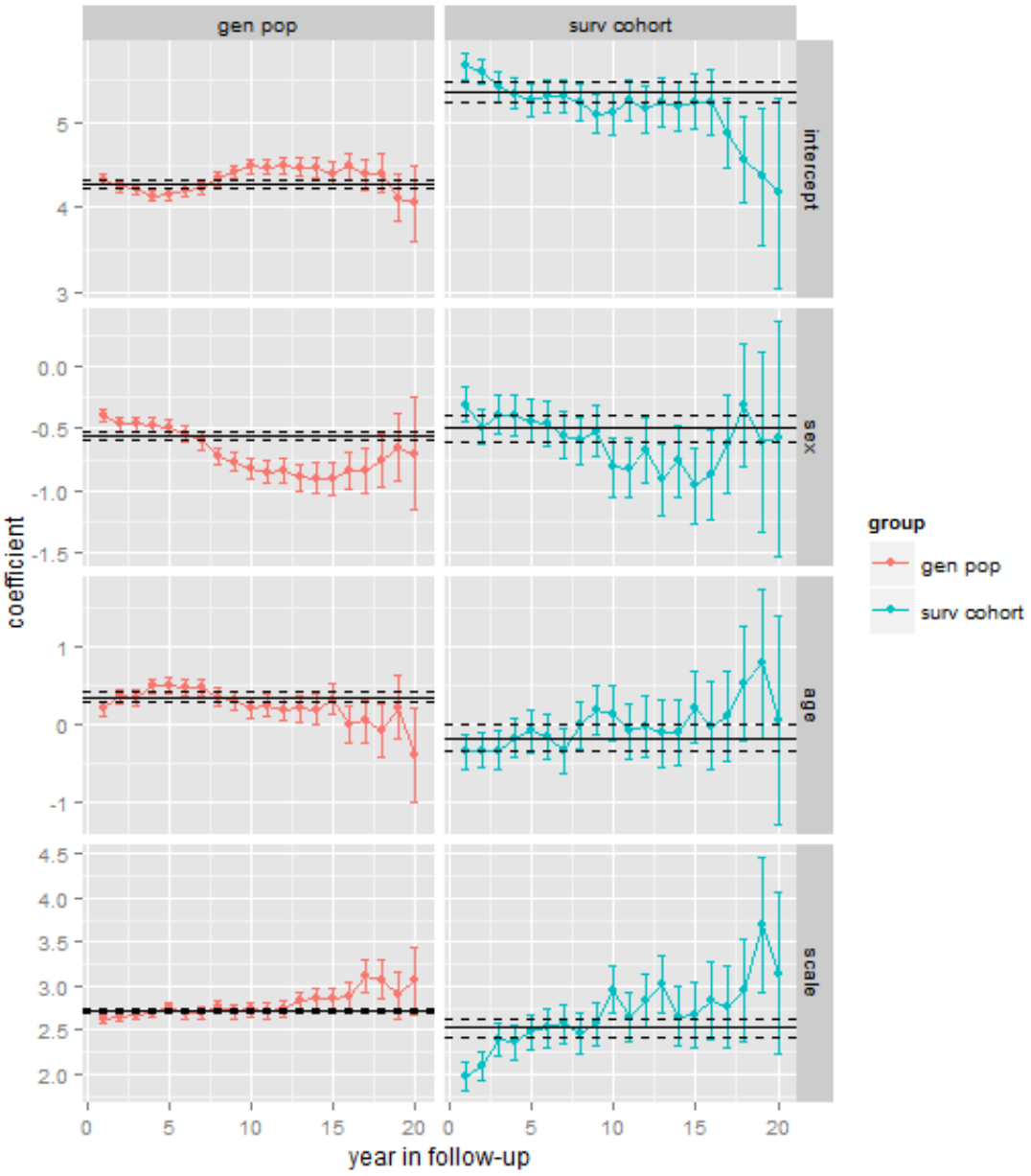

Figure 4.16: Time-dependent coefficients: SC vs. GP comparison for costs under *M1101.1*.

4.3 Analysis of Combined Cohort and Population Data

We now combine the cohort $\mathcal{P}$ and the population $\mathcal{Q}$ into $\mathcal{O}$ and introduce an indicator variable g_i to indicate that subject i is from the cohort:

$$g_i = \begin{cases} 1 & \text{if } i \in \mathcal{P} \\ 0 & \text{if } i \in \mathcal{Q}. \end{cases}$$

We fit marginal models to the combined data $\mathcal{O} = \{(Y_{ij}, \mathbf{Z}_i, g_i) : i = 1, \ldots, n+m;\ j = 1, \ldots, J_i\}$ using GEEs for the yearly visit counts and costs. These analyses allow us to make formal inference on the differences between the the cohort and the population and whether or not the differences are time-varying.

4.3.1 Model Specification for Combined Data

As in Section 4.2.1, we must specify $E(Y_{ij}|\mathbf{Z}_i) = \mu(\mathbf{Z}_i; b_j)$ for the regression model and the variance function $\text{Var}(\mathbf{Y}_i|\mathbf{Z}_i)$. We have seen in Section 4.2 that the intercept, sex, and age have time-varying effects on the longitudinal responses in both the cohort and the population. For the combined data, the effects of *g*, *g*sex*, *g*SES*, and *g*age* numerically measure the differences between the cohort and the population, and they can also be time-dependent. Therefore, the mean function for the combined data is as follows:

$$l\{\mu(\mathbf{Z}_i; b_j)\} = b_{0j} + b_{1j} sex_i + b_{2j} SES_i + b_{3j} age_i + c_{0j} g_i + c_{1j} g_i * sex_i + c_{2j} g_i * SES_i + c_{3j} g_i * age_i. \tag{4.2}$$

We label the models for the combined data $Ma_0a_1a_2a_3.A.B_0B_1B_2B_3$. The a_k's are again based on the structure of the linear predictor in (4.1) indicating whether or not the coefficient b_{kj} is time-dependent, and A is again based on the correlation structure of $\Gamma_i(\sigma)$. The B_k's present the group effects c_{kj} in the last four terms of (4.2), for $k = 0, 1, 2, 3$. $B_k = 0$ indicates that the factor is not in the model; $B_k = 1$ indicates that it is time-independent; and $B_k = 2$ indicates that it is time-dependent. For example, *M1101.1.2101* indicates that in (4.2) $c_{2j} \equiv 0$, *g*SES* is not in the model, the effect of SES is the same for the cohort and the population; the effects of sex and age are different in the two groups but the difference is constant over time ($c_{1j} \equiv c_1$ and $c_{3j} \equiv c_3$); and the intercept is different in the two groups and the difference is also time-varying.

4.3.2 Results and Comparison of Yearly Counts

We now analyze the yearly counts for the combined data and perform inference on the differences between the cohort and the population. When $Y_{ij} = N_{ij}$, the link function $l(\cdot) = \log(\cdot)$ and $\text{Var}(Y_{ij}|\mathbf{Z}_i) = \phi\mu(\mathbf{Z}_i; b_j)$ are used in the GEEs.

As in Section 4.2.2, *M1101.1* is preferred for the yearly counts of both the cohort and the population. We first investigate *M1101.1.1111*. The results show that the SES effect does not differ significantly between the cohort and the population, but the sex and age effects do.

The first panel of Table 4.6 gives the results for *M1101.1.2202*. We conduct Wald-type tests to test whether or not the six time-varying effects are constant; these results are also listed. We see that the tests $c_{1j} \equiv c_1$ for *group*sex* ($p = .544$) and $c_{3j} \equiv c_3$ for *group*age* ($p = .064$), so the effects of sex and age are different in the two groups but the difference is time-independent, and the other effects are time-varying ($p < .0001$). Therefore, we use *M1101.1.2101* to analyze the yearly counts in the combined data. The model estimates are presented in the second panel of Table 4.6. Figure 4.17 shows the time-varying effects over the follow-up period for *M1101.1.2101* and compares the two groups. The intercept effect is different in the two groups, and the difference decreases over time.

4.3.3 Results and Comparison of Yearly Costs

We now analyze the yearly costs for the combined data and perform inference on the differences between the cohort and the population. When $Y_{ij} = \log(C_{ij} + 5)$, the link function $l(\cdot) = I(\cdot)$ and $\text{Var}(Y_{ij}|\mathbf{Z}_i) = \phi_j$ are used in the GEEs, where the ϕ_j's are scale parameters and change over time.

As in Section 4.2.3, *M1101.1* is preferred for the yearly costs of both the cohort and the population. We first investigate *M1101.1.1111*; see the first panel of Table 4.7. The effects of *group*sex* and *group*SES* are not significant, which means that the effects of sex and SES are the same for the two groups. Therefore, we fit *M1101.1.2002* with time-varying group effects on intercept and age and no group effects on sex and SES; see the second panel of Table 4.7. We conduct Wald-type tests to test whether or not the five time-varying effects are constant, e.g., $c_{3j} \equiv c_3$ for *group*age* ($p = .039$). These results are also listed, and they are all time-varying. Figure 4.18 shows the time-varying effects over the follow-up period for *M1101.1.2002* and compares the two groups. The difference of age effect in the two groups over time is consistent with the result analyzed separately in Section 4.2.3.

4.4 Summary and Discussion

We have described the CAYACS longitudinal data, conducted analyses of the yearly counts and costs by the conventional GEE approach, and compared the trends of the cohort and the population through their time-varying effects.

Figures 4.11 and 4.16 to 4.18 show that whether we model these two groups separately or together, the difference over time arises mainly from the time-varying intercept and tends to decrease over time. To see the trends over time, it is necessary to conduct longitudinal analysis. The time-varying effect of age has different trends in the two groups, especially

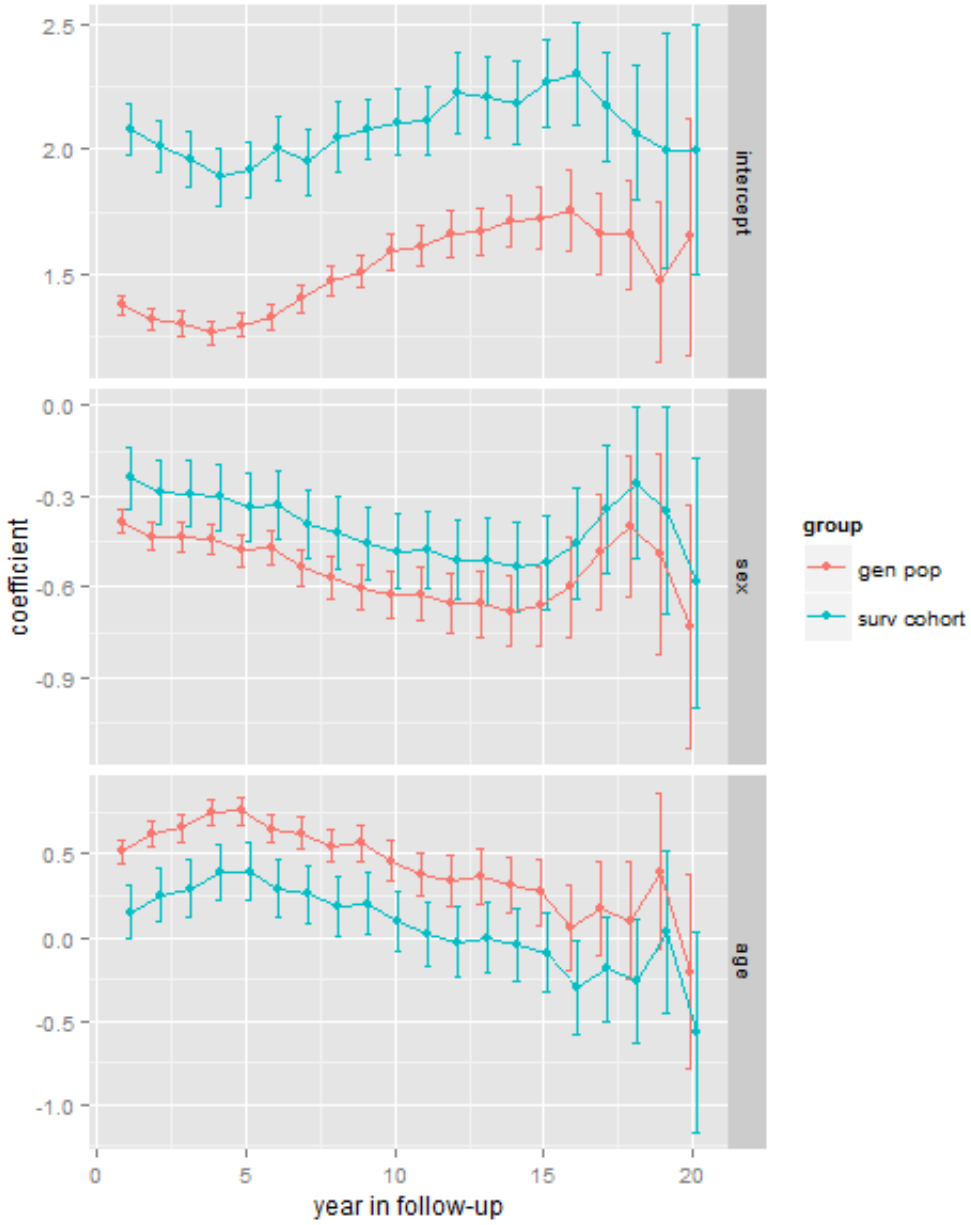

Figure 4.17: Time-dependent coefficients of yearly counts for combined data under *M1101.1.100*.

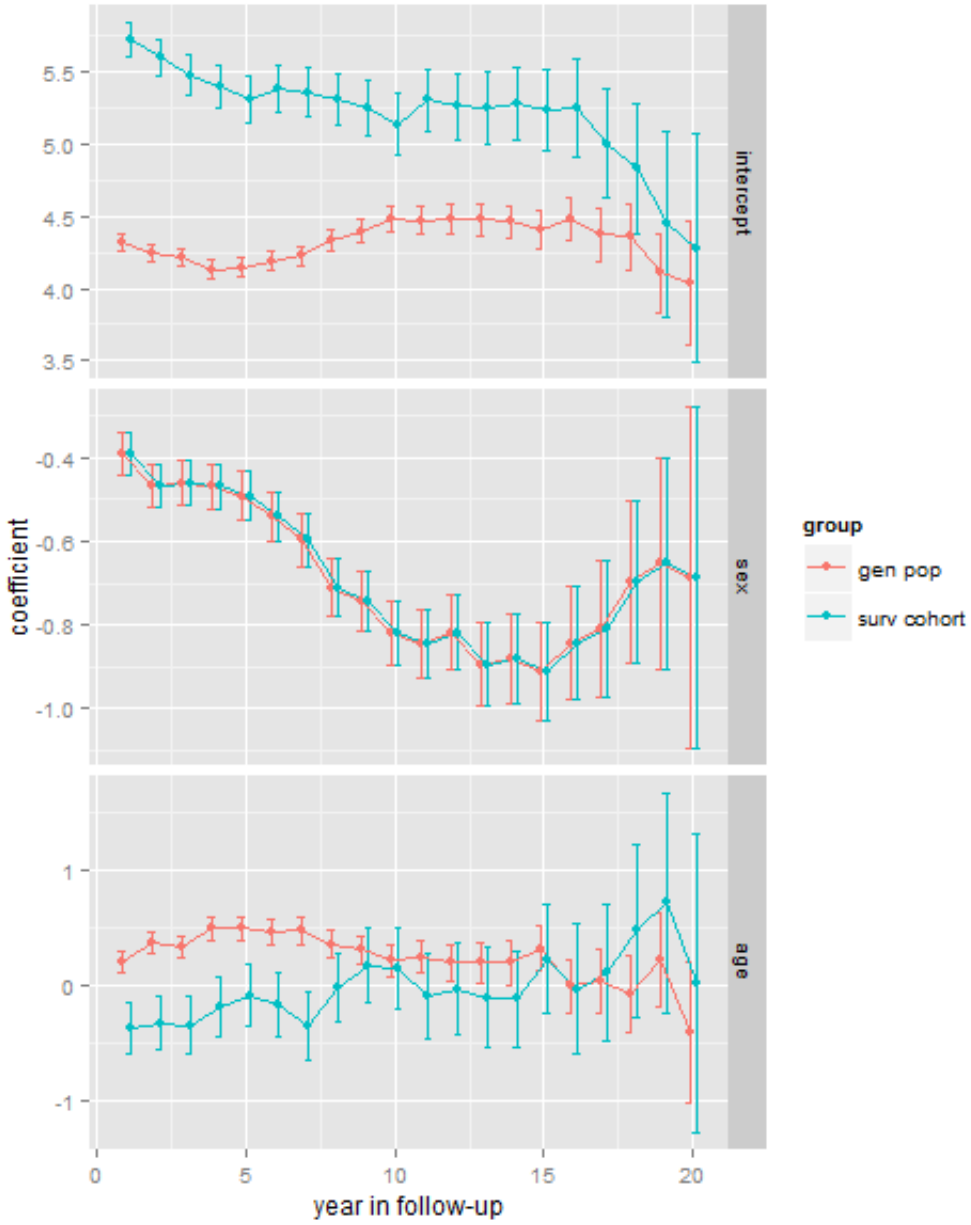

Figure 4.18: Time-dependent coefficients of yearly costs for combined data under *M1101.1.100*.

for the yearly costs, but the sex effect has similar trends. The results imply that females see physicians more frequently and cost more especially during their pregnancy periods, but this is true for both the population and the cohort. Sex may not be a risk factor for later effects of the cancer diagnosis. The LCMs developed in the previous chapters can help to detect the risk factors for the latent at-risk indicator. We will extend the LCM methodology to longitudinal data in the next chapter.

Another benefit of longitudinal analysis is that it permits subject-specific inference by random effect models, which can deal with unexplained variation. This will also be explored in the next chapter.

Table 4.2: Full Survivor Cohort: GEE Analysis of Yearly Visit Counts

	independent			CS			AR(1)		
	est	se	p-value	est	se	p-value	est	se	p-value
	M0000								
intercept	2.018	0.054	0.000	1.987	0.052	0.000	1.981	0.051	0.000
sex	-0.384	0.053	0.000	-0.326	0.052	0.000	-0.325	0.050	0.000
SES	-0.010	0.056	0.856	0.004	0.054	0.947	-0.005	0.051	0.919
age at entry	0.240	0.082	0.004	0.243	0.081	0.003	0.213	0.077	0.006
dispersion	12.632	1.542		12.535	1.529		12.841	1.584	
correlation				0.428	0.028		0.821	0.016	
	M1000								
intercept	1.968[a]	0.059[b]		1.934[a]	0.057[b]		1.877[a]	0.058[b]	
sex	-0.384	0.053	0.000	-0.320	0.052	0.000	-0.320	0.051	0.000
SES	-0.011	0.056	0.841	0.005	0.054	0.923	-0.002	0.052	0.967
age at entry	0.241	0.082	0.003	0.238	0.081	0.003	0.207	0.079	0.009
dispersion	12.617	1.538		12.631	1.537		13.089	1.594	
correlation				0.431	0.027		0.823	0.016	
	M1100								
intercept	1.981[a]	0.059[b]		1.982[a]	0.059[b]		1.912[a]	0.062[b]	
sex	-0.432[a]	0.069[b]		-0.405[a]	0.067[b]		-0.400[a]	0.069[b]	
SES	-0.011	0.056	0.848	0.007	0.054	0.899	-0.006	0.053	0.909
age at entry	0.242	0.082	0.003	0.204	0.081	0.011	0.198	0.079	0.012
dispersion	12.509	1.498		12.573	1.510		13.019	1.563	
correlation				0.431	0.027		0.823	0.015	
	M1001								
intercept	1.994[a]	0.067[b]		1.948[a]	0.067[b]		1.861[a]	0.073[b]	
sex	-0.384	0.053	0.000	-0.330	0.052	0.000	-0.331	0.052	0.000
SES	-0.010	0.056	0.858	0.003	0.054	0.953	-0.008	0.054	0.886
age at entry	0.184[a]	0.104[b]		0.217[a]	0.098[b]		0.249[a]	0.104[b]	
dispersion	12.515	1.511		12.545	1.515		13.011	1.571	
correlation				0.433	0.027		0.824	0.016	
	M1101								
intercept	2.014[a]	0.064[b]		2.007[a]	0.065[b]		1.909[a]	0.077[b]	
sex	-0.433[a]	0.069[b]		-0.416[a]	0.068[b]		-0.413[a]	0.070[b]	
SES	-0.009	0.056	0.867	0.005	0.055	0.926	-0.011	0.055	0.840
age at entry	0.175[a]	0.104[b]		0.162[a]	0.098[b]		0.215[a]	0.105[b]	
dispersion	12.409	1.470		12.495	1.486		12.942	1.537	
correlation				0.433	0.027		0.824	0.015	
	M1111								
intercept	2.024[a]	0.063[b]		2.019[a]	0.065[b]		1.928[a]	0.077[b]	
sex	-0.431[a]	0.069[b]		-0.412[a]	0.068[b]		-0.393[a]	0.069[b]	
SES	-0.052[a]	0.074[b]		-0.036[a]	0.070[b]		-0.081[a]	0.074[b]	
age at entry	0.179[a]	0.103[b]		0.159[a]	0.097[b]		0.205[a]	0.105[b]	
dispersion	12.359	1.451		12.459	1.471		12.881	1.517	
correlation				0.433	0.027		0.825	0.016	

[a]average value over 20 estimates

[b]se of the 20 averaged estimates

Table 4.3: General Population: GEE Analysis of Yearly Visit Counts

	independent			CS			AR(1)		
	est	se	p-value	est	se	p-value	est	se	p-value
	M0000								
intercept	1.475	0.019	0.000	1.389	0.018	0.000	1.401	0.018	0.000
sex	-0.498	0.020	0.000	-0.469	0.018	0.000	-0.475	0.018	0.000
SES	-0.063	0.020	0.002	-0.067	0.018	0.000	-0.064	0.019	0.001
age at entry	0.586	0.032	0.000	0.604	0.030	0.000	0.558	0.030	0.000
dispersion	9.618	0.542		10.262	0.582		10.405	0.588	
correlation				0.384	0.013		0.777	0.009	
	M1000								
intercept	1.519[a]	0.024[b]		1.419[a]	0.024[b]		1.383[a]	0.024[b]	
sex	-0.498	0.019	0.000	-0.476	0.019	0.000	-0.476	0.018	0.000
SES	-0.063	0.020	0.002	-0.067	0.019	0.000	-0.063	0.019	0.001
age at entry	0.587	0.032	0.000	0.592	0.030	0.000	0.560	0.030	0.000
dispersion	9.355	0.486		10.005	0.534		10.309	0.566	
correlation				0.379	0.013		0.774	0.009	
	M1100								
intercept	1.524[a]	0.025[b]		1.440[a]	0.026[b]		1.408[a]	0.025[b]	
sex	-0.511[a]	0.037[b]		-0.506[a]	0.039[b]		-0.539[a]	0.039[b]	
SES	-0.062	0.020	0.002	-0.063	0.019	0.001	-0.062	0.019	0.001
age at entry	0.586	0.032	0.000	0.569	0.031	0.000	0.557	0.030	0.000
dispersion	9.358	0.486		10.034	0.534		10.377	0.577	
correlation				0.379	0.013		0.775	0.009	
	M1001								
intercept	1.596[a]	0.030[b]		1.502[a]	0.033[b]		1.435[a]	0.031[b]	
sex	-0.498	0.019	0.000	-0.494	0.019	0.000	-0.480	0.018	0.000
SES	-0.063	0.020	0.001	-0.067	0.019	0.000	-0.064	0.019	0.001
age at entry	0.430[a]	0.055[b]		0.435[a]	0.059[b]		0.453[a]	0.058[b]	
dispersion	9.301	0.477		9.974	0.529		10.262	0.559	
correlation				0.380	0.013		0.776	0.009	
	M1101								
intercept	1.606[a]	0.034[b]		1.537[a]	0.036[b]		1.468[a]	0.034[b]	
sex	-0.524[a]	0.037[b]		-0.546[a]	0.040[b]		-0.553[a]	0.040[b]	
SES	-0.063	0.020	0.001	-0.062	0.019	0.001	-0.063	0.019	0.001
age at entry	0.427[a]	0.056[b]		0.399[a]	0.060[b]		0.439[a]	0.058[b]	
dispersion	9.307	0.481		10.029	0.537		10.338	0.574	
correlation				0.381	0.013		0.776	0.009	
	M1111								
intercept	1.607[a]	0.034[b]		1.539[a]	0.037[b]		1.471[a]	0.033[b]	
sex	-0.525[a]	0.037[b]		-0.546[a]	0.040[b]		-0.552[a]	0.040[b]	
SES	-0.064[a]	0.035[b]		-0.069[a]	0.037[b]		-0.070[a]	0.038[b]	
age at entry	0.426[a]	0.056[b]		0.399[a]	0.059[b]		0.439[a]	0.058[b]	
dispersion	9.295	0.479		10.019	0.534		10.325	0.570	
correlation				0.382	0.013		0.777	0.009	

[a]average value over 20 estimates

[b]se of the 20 averaged estimates

Table 4.4: Full Survivor Cohort: GEE Analysis of Yearly Visit Costs

	independent			CS			AR(1)		
	est	se	p-value	est	se	p-value	est	se	p-value
	M0000								
intercept	5.369	0.066	0.000	5.354	0.063	0.000	5.326	0.062	0.000
sex	-0.584	0.058	0.000	-0.500	0.055	0.000	-0.516	0.055	0.000
SES	-0.001	0.060	0.982	0.026	0.057	0.648	0.025	0.056	0.650
age at entry	-0.119	0.093	0.199	-0.176	0.087	0.043	-0.159	0.087	0.067
scale	2.520	0.054		2.523	0.054		2.522	0.054	
correlation				0.363	0.015		0.765	0.010	
	M1000								
intercept	5.217[a]	0.070[b]		5.151[a]	0.069[b]		5.087[a]	0.067[b]	
sex	-0.564	0.057	0.000	-0.453	0.055	0.000	-0.482	0.055	0.000
SES	-0.008	0.058	0.895	0.015	0.056	0.785	0.017	0.055	0.765
age at entry	-0.131	0.090	0.146	-0.233	0.087	0.008	-0.186	0.087	0.032
scale	2.692[a]	0.074[b]		2.707[a]	0.073[b]		2.714[a]	0.073[b]	
correlation				0.361	0.014	0.000	0.764	0.010	0.000
	M1100								
intercept	5.269[a]	0.078[b]		5.234[a]	0.074[b]		5.180[a]	0.071[b]	
sex	-0.661[a]	0.085[b]		-0.611[a]	0.077[b]		-0.656[a]	0.077[b]	
SES	-0.007	0.058	0.903	0.017	0.056	0.761	0.018	0.055	0.739
age at entry	-0.128	0.090	0.157	-0.229	0.087	0.008	-0.182	0.086	0.034
scale	2.678[a]	0.075[b]		2.685[a]	0.074[b]		2.696[a]	0.074[b]	
correlation				0.360	0.014	0.000	0.763	0.010	0.000
	M1001								
intercept	5.181[a]	0.080[b]		5.040[a]	0.078[b]		4.988[a]	0.076[b]	
sex	-0.565	0.057	0.000	-0.455	0.055	0.000	-0.484	0.055	0.000
SES	-0.010	0.058	0.869	0.014	0.056	0.796	0.014	0.055	0.806
age at entry	-0.053[a]	0.133[b]		0.000[a]	0.114[b]		0.023[a]	0.118[b]	
scale	2.684[a]	0.075[b]		2.700[a]	0.075[b]		2.710[a]	0.075[b]	
correlation				0.362	0.014	0.000	0.764	0.010	0.000
	M1101								
intercept	5.236[a]	0.086[b]		5.124[a]	0.085[b]		5.086[a]	0.081[b]	
sex	-0.663[a]	0.085[b]		-0.606[a]	0.077[b]		-0.654[a]	0.077[b]	
SES	-0.009	0.058	0.878	0.016	0.056	0.773	0.015	0.055	0.780
age at entry	-0.056[a]	0.132[b]		-0.005[a]	0.114[b]		0.010[a]	0.117[b]	
scale	2.670[a]	0.075[b]		2.680[a]	0.076[b]		2.693[a]	0.076[b]	
correlation				0.361	0.014	0.000	0.763	0.010	0.000
	M1111								
intercept	5.252[a]	0.084[b]		5.129[a]	0.083[b]		5.090[a]	0.081[b]	
sex	-0.661[a]	0.084[b]		-0.604[a]	0.076[b]		-0.652[a]	0.077[b]	
SES	-0.059[a]	0.090[b]		0.007[a]	0.079[b]		0.004[a]	0.079[b]	
age at entry	-0.058[a]	0.132[b]		-0.014[a]	0.113[b]		0.006[a]	0.116[b]	
scale	2.663[a]	0.075[b]		2.673[a]	0.076[b]		2.688[a]	0.076[b]	
correlation				0.361	0.014	0.000	0.763	0.010	0.000

[a] average value over 20 estimates

[b] se of the 20 averaged estimates

Table 4.5: General Population: GEE Analysis of Yearly Visit Costs

	independent			CS			AR(1)		
	est	se	p-value	est	se	p-value	est	se	p-value
	M0000								
intercept	4.370	0.022	0.000	4.277	0.021	0.000	4.291	0.021	0.000
sex	-0.614	0.020	0.000	-0.558	0.019	0.000	-0.582	0.019	0.000
SES	-0.050	0.020	0.014	-0.051	0.020	0.010	-0.055	0.020	0.005
age at entry	0.399	0.034	0.000	0.348	0.032	0.000	0.334	0.032	0.000
scale	2.715	0.014		2.723	0.014		2.724	0.014	
correlation				0.332	0.005		0.724	0.004	
	M1000								
intercept	4.375[a]	0.025[b]		4.193[a]	0.024[b]		4.211	0.024[b]	
sex	-0.612	0.020	0.000	-0.549	0.019	0.000	-0.578	0.019	0.000
SES	-0.050	0.020	0.014	-0.050	0.020	0.012	-0.054	0.020	0.006
age at entry	0.401	0.033	0.000	0.349	0.032	0.000	0.333	0.032	0.000
scale	2.775[a]	0.024[b]		2.820[a]	0.025[b]		2.820[a]	0.025[b]	
correlation				0.333	0.005	0.000	0.725	0.003	0.000
	M1100								
intercept	4.432[a]	0.027[b]		4.269[a]	0.027[b]		4.288[a]	0.026[b]	
sex	-0.718[a]	0.031[b]		-0.687[a]	0.029[b]		-0.721[a]	0.029[b]	
SES	-0.049	0.020	0.015	-0.049	0.020	0.013	-0.053	0.020	0.007
age at entry	0.400	0.033	0.000	0.347	0.032	0.000	0.332	0.032	0.000
scale	2.764[a]	0.025[b]		2.804[a]	0.026[b]		2.806[a]	0.026[b]	
correlation				0.333	0.005	0.000	0.725	0.003	0.000
	M1001								
intercept	4.403[a]	0.030[b]		4.237[a]	0.029[b]		4.214[a]	0.028[b]	
sex	-0.612	0.020	0.000	-0.549	0.019	0.000	-0.579	0.019	0.000
SES	-0.050	0.020	0.013	-0.050	0.020	0.012	-0.054	0.020	0.006
age at entry	0.342[a]	0.050[b]		0.253[a]	0.046[b]		0.332[a]	0.046[b]	
scale	2.771[a]	0.024[b]		2.818[a]	0.025[b]		2.816[a]	0.025[b]	
correlation				0.334	0.005	0.000	0.725	0.003	0.000
	M1101								
intercept	4.469[a]	0.033[b]		4.324[a]	0.032[b]		4.299[a]	0.031[b]	
sex	-0.725[a]	0.031[b]		-0.697[a]	0.030[b]		-0.724[a]	0.029[b]	
SES	-0.049	0.020	0.015	-0.049	0.020	0.013	-0.053	0.020	0.007
age at entry	0.327[a]	0.050[b]		0.235[a]	0.047[b]		0.314	0.046[b]	
scale	2.759[a]	0.025[b]		2.801[a]	0.025[b]		2.802[a]	0.025[b]	
correlation				0.333	0.005	0.000	0.725	0.003	0.000
	M1111								
intercept	4.471[a]	0.034[b]		4.328[a]	0.032[b]		4.307[a]	0.031[b]	
sex	-0.725[a]	0.031[b]		-0.697[a]	0.030[b]		-0.724[a]	0.029[b]	
SES	-0.055[a]	0.032[b]		-0.057[a]	0.030[b]		-0.073[a]	0.029[b]	
age at entry	0.327[a]	0.050[b]		0.234[a]	0.047[b]		0.313[a]	0.046[b]	
scale	2.759[a]	0.025[b]		2.801[a]	0.025[b]		2.802[a]	0.025[b]	
correlation				0.333	0.005	0.000	0.725	0.003	0.000

[a] average value over 20 estimates

[b] se of the 20 averaged estimates

Table 4.6: Combined General Population and Survivor Cohort: Analysis of Yearly Visit Counts

	M1101.1.2202					*M1101.1.2101*		
	est	se	p-value	test	p-value	est	se	p-value
intercept	1.530[a]	0.037[b]		$b_{0j} \equiv b_0$	$< .0001$	1.520[a]	0.034[b]	
sex	-0.547[a]	0.040[b]		$b_{1j} \equiv b_1$	$< .0001$	-0.548[a]	0.035[b]	
SES	-0.051	0.018	0.005			-0.051	0.018	0.005
age	0.398[a]	0.060[b]		$b_{3j} \equiv b_3$	$< .0001$	0.417[a]	0.054[b]	
group (SC vs. GP)	0.493[a]	0.072[b]		$c_{0j} \equiv c_0$	$< .0001$	0.555[a]	0.057[b]	
group*sex (SC vs. GP)	0.131[a]	0.078[b]		$c_{1j} \equiv c_1$	0.544	0.143	0.055	0.009
group*SES (SC vs. GP)	-	-				-	-	
group*age (SC vs. GP)	-0.228[a]	0.114[b]		$c_{3j} \equiv c_3$	0.064	-0.363	0.086	$< .0001$
dispersion	10.3	0.507				10.3	0.508	
correlation	0.39	0.012				0.389	0.012	

[a]average value over 20 estimates

[b]se of the 20 averaged estimates

Table 4.7: Combined General Population and Survivor Cohort: Analysis of Yearly Visit Costs

	M1101.1.1111			*M1101.1.2002*				
	est	se	p-value	est	se	p-value	test	p-value
intercept	4.290[a]	0.031[b]		4.310[a]	0.031[b]		$b_{0j} \equiv b_0$	$< .0001$
sex	-0.604[a]	0.025[b]		-0.687[a]	0.028[b]		$b_{1j} \equiv b_1$	$< .0001$
SES	-0.049	0.020	0.014	-0.042	0.019	0.025		
age	0.247[a]	0.041[b]		0.235[a]	0.047[b]		$b_{3j} \equiv b_3$	$< .0001$
group (SC vs. GP)	1.096	0.066	$< .0001$	0.885[a]	0.071[b]		$c_{0j} \equiv c_0$	$< .0001$
group*sex (SC vs. GP)	0.068	0.058	0.240	-	-			
group*SES (SC vs. GP)	0.072	0.059	0.225	-	-			
group*age (SC vs. GP)	-0.531	0.092	$< .0001$	-0.248[a]	0.121[b]		$c_{3j} \equiv c_3$	0.039
scale	2.800[a]	0.025[b]		2.790[a]	0.024[b]			
correlation	0.336	0.005	$< .0001$	0.336	0.005	$< .0001$		

[a]average value over 20 estimates

[b]se of the 20 averaged estimates

Chapter 5

Extended GEE Procedures for Longitudinal Data

5.1 Introduction

Chapters 2 and 3 formulated an LCM for cross-sectional counts of the CAYACS data. We believe that the survivor cohort ($\mathcal{P}$) consists of two latent classes, indicated by a latent variable η. Let $\mathcal{P}_\eta$ be the two latent classes: the at-risk class ($\mathcal{P}_1$) and the not-at-risk class ($\mathcal{P}_0$). Class $\mathcal{P}_0$ has the same visit patterns as the population ($\mathcal{Q}$), and $\mathcal{P}_1$ has more frequent visits. Let $\mathcal{P}^\delta$ be the subsets of the cohort with or without RSC, indicated by δ. Subjects with RSC in the cohort ($\mathcal{P}^1$) form a part of the at-risk class ($\mathcal{P}^1 \subset \mathcal{P}_1$).

5.1.1 Motivation

The cross-sectional analyses in Chapters 2 and 3 do not reveal the visit trends over time. The longitudinal analysis in Chapter 4 showed that the cohort has different visit trends than the population. It is especially important to distinguish the variation in patterns due to an individual's aging from the variation due to individual differences, given the long follow-up period. The full cohort $\mathcal{P}$ includes a not-at-risk component with the same visit trends as the population ($\mathcal{P}_0 \equiv \mathcal{Q}$). Therefore, we wish to identify the at-risk class and study its visit trends and medical costs over time. LCMs can achieve this. This chapter will extend the extended GEE methodology developed in Chapter 3, in particular, the type P extended GEE estimator, to the longitudinal counts and medical costs.

Some subjects in the cohort experienced RSC. Conceptually, they were suffering consequences of the original diagnoses. Let $\delta = \delta_T$, then $\mathcal{P}^1$ is the subset of survivors with RSC before the end of follow-up. Under additional assumptions in Section 3.2, this group served as representatives of the at-risk class ($\mathcal{P}_1 \equiv \mathcal{P}^1$). Chapter 4 explains how we get the CAYACS yearly data. Table 5.1 summarizes the yearly data from $\mathcal{P}^1$, $\mathcal{P}$, and $\mathcal{Q}$. They have similar means for the cluster size J_i. The overall mean of the yearly counts of $\mathcal{P}^1$ is

more than double that of $\mathcal{Q}$, and the overall mean of the yearly costs of $\mathcal{P}^1$ is about triple that of $\mathcal{Q}$. The overall means of $\mathcal{P}$ are between those of $\mathcal{P}^1$ and $\mathcal{Q}$.

Table 5.1: Yearly Visit Data Summary: $\mathcal{P}^1$ vs. $\mathcal{P}$ vs. $\mathcal{Q}$

Sample	# subjects	# observations	mean(J_i)	mean(N_{ij})	mean(C_{ij})
Subset of SC $\mathcal{P}^1$	237	2283	9.6	9.63	607.45
Survivor Cohort $\mathcal{P}$	1609	15354	9.5	6.95	378.09
General Population $\mathcal{Q}$	14289	124823	8.7	4.44	194.89

Similarly to Figures 4.4 and 4.5, Figures 5.1 and 5.2 compare the means and the corresponding CIs in the twenty years of the follow-up for $\mathcal{P}^1$, $\mathcal{P}$, and $\mathcal{Q}$. The numbers at the top are the sample sizes used to calculate the means and CIs of $\mathcal{P}^1$ in each year. The grey lines and numbers correspond to the black ones in Figures 4.4 and 4.5, which presented the full cohort $\mathcal{P}$. On average, the cohort values started about twice as high as those of the population, and the differences gradually decreased until the values were about the same. We can see that $\mathcal{P}^1$ has much higher visit counts and costs than $\mathcal{P}$ or $\mathcal{Q}$. In general, $\mathcal{P}$ tends to be stable over the years, and $\mathcal{Q}$ tends to increase. However, there is a clear decreasing trend for $\mathcal{P}^1$. The cohort $\mathcal{P}$ is about halfway between $\mathcal{P}^1$ and $\mathcal{Q}$, except in the final four years, when the sample sizes become small and the estimates are less reliable.

The overall mean summary and the yearly mean plots confirm our conjecture that $\mathcal{P}$ is a combination of an at-risk class and a not-at-risk class.

5.1.2 Model Specification

The CAYACS data was described in Section 4.2, and as before Y_{ij} represents the yearly visit counts or the transformed yearly costs; η is the latent risk indicator. We also use Y_{ij} as a longitudinal binary variable to indicate whether or not there are physician visits in the j^{th} year for subject i, i.e., $N_{ij} > 0$, so $C_{ij} > 0$.

Let $\mathbf{Z}_{ij}$ be a vector of p covariates observed in the j^{th} year for the i^{th} subject. As in Section 3.2, we assume that $P(\eta = 1|\delta = 1, \mathbf{Z}) = 1$, i.e., $\delta = 1$ implies $\eta = 1$, and $\delta = 1$ is a representative subgroup of $\eta = 1$ with the same distribution of the responses: $[\mathbf{Y}|\eta = 1, \mathbf{Z}] = [\mathbf{Y}|\delta = 1, \mathbf{Z}]$.

We first focus on inference for the population average within each class, so marginal models are appropriate (Diggle *et al.*, 2002). Marginal models do not specify the joint distribution of the repeated measurements from the same subject, but they specify both the regression model of $\mathbf{Y}$ on $\mathbf{Z}$ and the association among repeated observations of $\mathbf{Y}$ for each individual, separately. Therefore, they are appropriate for the CAYACS data, because the yearly counts are overdispersed and the yearly costs are not normally distributed. Marginal models for LCMs are specified as follows:

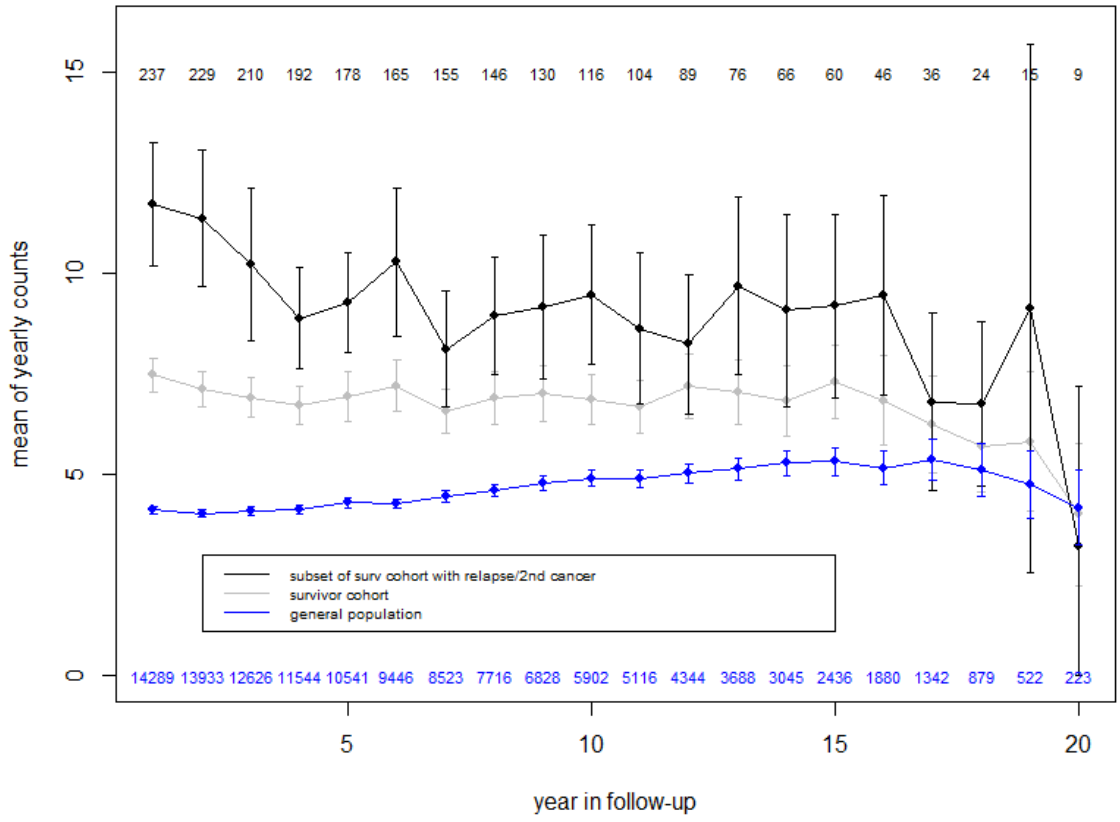

Figure 5.1: Mean and CI of yearly visit counts during follow-up: Subset of survivor cohort with RSC vs. full survivor cohort vs. general population.

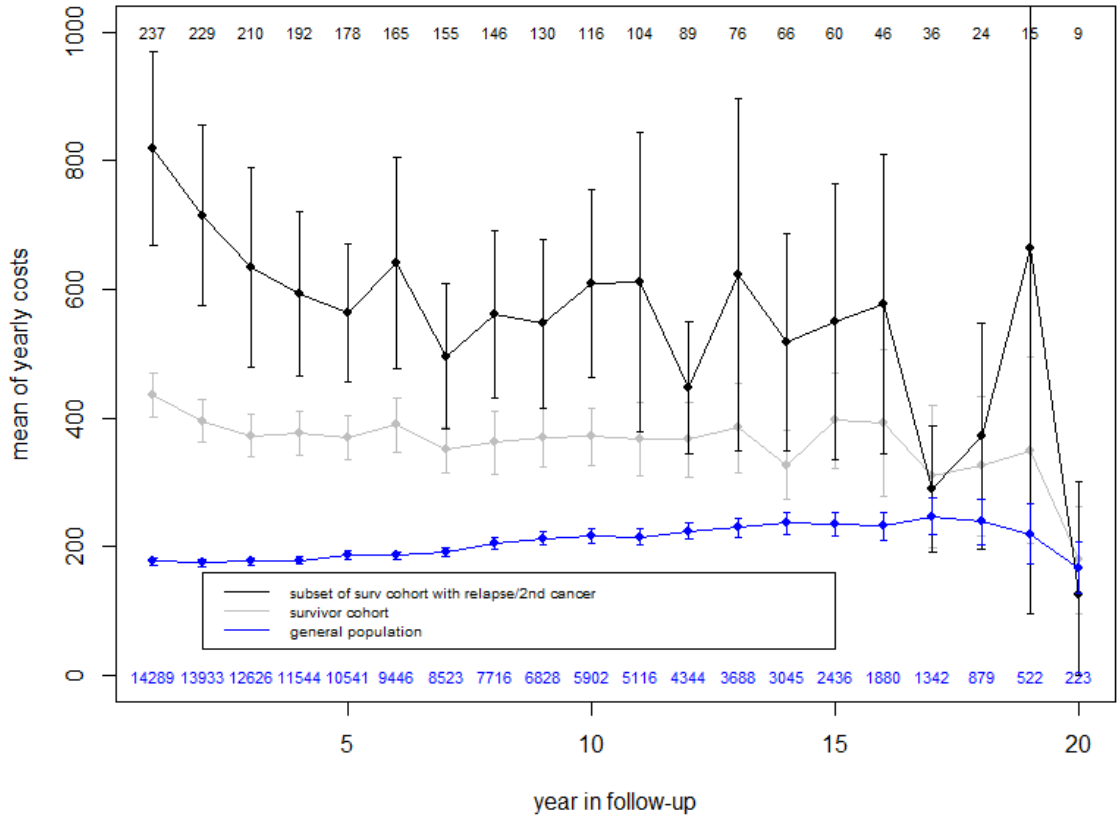

Figure 5.2: Mean and CI of yearly costs during follow-up: Subset of survivor cohort with RSC vs. full survivor cohort vs. general population.

LCMs for longitudinal data

- **Risk model for η**:
  Given a set of time-independent covariates $\mathbf{Z}_i$, let the mean of η_i be $E(\eta_i|\mathbf{Z}_i) = P(\eta_i = 1|\mathbf{Z}_i) = p(\mathbf{Z}_i;\alpha)$, which can be specified up to parameter α. A typical logistic regression form is
  $$\text{logit}\{p(\mathbf{Z}_i;\alpha)\} = \alpha^{'}\mathbf{Z}_i. \tag{5.1}$$

- **Regression model for the at-risk class $\eta = 1$**:
  For the $\eta = 1$ class, the mean of the response Y_{ij} given a p-dimensional time-dependent covariate vector $\mathbf{Z}_{ij}$ can be specified up to parameter β: $E(Y_{ij}|\eta_i = 1, \mathbf{Z}_{ij}) = \mu_1(\mathbf{Z}_{ij};\beta_j)$, where the effects of $\mathbf{Z}_{ij}$ can be time-varying. For each subject i, the J_i-dimensional mean vector of the longitudinal responses $\mathbf{Y}_i$ is $E(\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i) = [\mu_1(\mathbf{Z}_{i1};\beta_1), \ldots, \mu_1(\mathbf{Z}_{iJ_i};\beta_{J_i})]' = \boldsymbol{\mu}_1(\mathbf{Z}_i;\boldsymbol{\beta}) = \boldsymbol{\mu}_{1i}$. Let $l(\cdot)$ be the link function: $l(\cdot) = \text{logit}(\cdot)$ for yearly binaries, $l(\cdot) = \log(\cdot)$ for yearly counts, and $l(\cdot) = I(\cdot)$ for yearly transformed costs. For example, the regression model of the longitudinal responses can be specified in a generalized linear form:
  $$l\{E(Y_{ij}|\eta_i = 1, \mathbf{Z}_{ij})\} = \beta_j^{'}\mathbf{Z}_{ij}. \tag{5.2}$$

  In marginal models, the within-subject association for each individual must also be specified explicitly. Let the $J_i \times J_i$ variance-covariance function of subject i be $\text{Var}(\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i) = \Sigma_1(\mathbf{Z}_i;\boldsymbol{\beta},\psi_1) = \Sigma_{1i}$, where ψ_1 is the additional parameter vector in the variance-covariance function for the $\eta = 1$ class. For example, the variance-covariance function can be decomposed into
  $$\Sigma_1(\mathbf{Z}_i;\boldsymbol{\beta},\psi_1) = T_i^{\frac{1}{2}}(\boldsymbol{\beta},\phi_1)\Gamma_i(\sigma_1)T_i^{\frac{1}{2}}(\boldsymbol{\beta},\phi_1), \tag{5.3}$$
  where $\psi_1 = (\phi_1', \sigma_1')'$. $T_i(\boldsymbol{\beta},\phi_1)$ is a $J_i \times J_i$ diagonal variance matrix, and the j^{th} element on the diagonal is $\text{Var}(Y_{ij}|\eta_i = 1, \mathbf{Z}_{ij})$, the structure of which can vary according to the data type. When Y_{ij} represents the count, the variance function of Y_{ij} can be specified as a quasi-Poisson form $\text{Var}(Y_{ij}|\eta_i = 1, \mathbf{Z}_{ij}) = \phi_1(\mathbf{Z}_{ij})\mu_1(\mathbf{Z}_{ij};\beta_j)$, where ϕ_1 is a dispersion parameter that can also depend on $\mathbf{Z}_{ij}$. When the Y_{ij}'s are continuous responses, we can for example specify $\text{Var}(Y_{ij}|\eta_i = 1, \mathbf{Z}_{ij}) = \phi_1(\mathbf{Z}_{ij})$ instead, where $\phi_1(\mathbf{Z}_{ij})$ is the scale parameter. $\phi_1(\mathbf{Z}_{ij}) \equiv \phi_1$ is a special case. $\Gamma_i(\sigma_1)$ is a $J_i \times J_i$ correlation matrix and σ_1 is a vector of correlation parameters, which can also depend on $\mathbf{Z}_{ij}$, e.g., $\sigma_1(\mathbf{Z}_{ij}) \equiv \sigma_1$. Liang and Zeger (1986) called $\Gamma_i(\cdot)$ the "working" correlation matrix, which makes (5.3) hold if it is indeed the true correlation matrix for $\mathbf{Y}|\eta$. For example, it can be an independent structure, a compound symmetric structure, or a time-series structure such as AR(1). Other choices of the correlation structure include one-dependent correlation, exponential correlation, and Gaussian correlation.

- **Regression model for the not-at-risk class $\eta = 0$:**
  For the $\eta = 0$ class, the mean and variance functions of the longitudinal responses $\mathbf{Y}_i$ are specified in marginal models with different parameters. Let $E(Y_{ij}|\eta_i = 0, \mathbf{Z}_{ij}) = \mu_0(\mathbf{Z}_{ij}; \theta_j)$, and let the J_i-dimensional mean response vector be $E(\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i) = [\mu_0(\mathbf{Z}_{i1}; \theta_1), \ldots, \mu_0(\mathbf{Z}_{iJ_i}; \theta_{J_i})]' = \boldsymbol{\mu}_0(\mathbf{Z}_i; \boldsymbol{\theta}) = \boldsymbol{\mu}_{0i}$. For example, the regression model for the $\eta = 0$ class can also be specified in a generalized linear form:

$$l\{E(Y_{ij}|\eta_i = 0, \mathbf{Z}_{ij})\} = \theta_j^{'}\mathbf{Z}_{ij}. \tag{5.4}$$

  Define the $J_i \times J_i$ variance-covariance function of $\mathbf{Y}_i$ for the $\eta = 0$ class to be $\text{Var}(\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i) = \Sigma_0(\mathbf{Z}_i; \boldsymbol{\theta}, \psi_0) = \Sigma_{0i}$, and

$$\Sigma_0(\mathbf{Z}_i; \boldsymbol{\theta}, \psi_0) = T_i^{\frac{1}{2}}(\boldsymbol{\theta}, \phi_0)\Gamma_i(\sigma_0)T_i^{\frac{1}{2}}(\boldsymbol{\theta}, \phi_0). \tag{5.5}$$

  The structure of Σ_{0i} is similar to that of Σ_{1i}.

Our primary objective in this chapter is to develop extended GEE inference procedures to estimate the parameters α, β, and θ in (5.1), (5.2), and (5.4) based on data from $\mathcal{P} = \{(Y_{ij}, \delta_i, \mathbf{Z}_{ij}) : i = 1, \ldots, n;\ j = 1, \ldots, J_i\}$ and $\mathcal{Q} = \{(Y_{ij}, \mathbf{Z}_{ij}) : i = 1, \ldots, m;\ j = 1, \ldots, J_i\}$. The additional parameters ψ_1 and ψ_0 in the covariance functions are treated as nuisance parameters (Liang and Zeger, 1986).

The regression parameters in marginal models for continuous measurements are equivalent to population-average coefficients in linear mixed effects models, and their variance-covariance functions can have a determined relationship in certain situations (Fitzmaurice *et al.*, 2012). This is not always the case for GLMMs. One exception is for counts when the random effect is only for the intercept in the GLMM (Fitzmaurice *et al.*, 2012). The determined relationships between marginal models and mixed effects models allow us to perform subject-specific inference for the LCMs. Subject-specific inference by the mixed effects model will be discussed in detail in Section 5.4.

5.2 Extended GEE Inference Procedures

The mean and variance functions for the full cohort, $E(\mathbf{Y}_i|\mathbf{Z}_i)$ and $\text{Var}(\mathbf{Y}_i|\mathbf{Z}_i)$, can be derived as follows. Define the J_i-dimensional vector of the mean function to be $E(\mathbf{Y}_i|\mathbf{Z}_i) = \boldsymbol{\Lambda}(\mathbf{Z}_i; \alpha, \beta, \theta) = \boldsymbol{\Lambda}_i$; then

$$\boldsymbol{\Lambda}_i = p(\mathbf{Z}_i; \alpha)\boldsymbol{\mu}_1(\mathbf{Z}_i; \boldsymbol{\beta}) + [1 - p(\mathbf{Z}_i; \alpha)]\boldsymbol{\mu}_0(\mathbf{Z}_i; \boldsymbol{\theta}). \tag{5.6}$$

The corresponding $J_i \times J_i$ variance-covariance function is defined to be $\text{Var}(\mathbf{Y}_i|\mathbf{Z}_i) = \Sigma(\mathbf{Z}_i; \alpha, \beta, \theta, \psi) = \Sigma_i$, where $\psi = (\psi_1', \psi_0')'$. Then

$$\Sigma_i = p(\mathbf{Z}_i;\alpha)\Sigma_{1i} + [1 - p(\mathbf{Z}_i;\alpha)]\Sigma_{0i} + p(\mathbf{Z}_i;\alpha)[1 - p(\mathbf{Z}_i;\alpha)]\{\boldsymbol{\mu}_{1i} - \boldsymbol{\mu}_{0i}\}\{\boldsymbol{\mu}_{1i} - \boldsymbol{\mu}_{0i}\}'. \quad (5.7)$$

Zeger and Liang (1986) extended quasi-likelihood approaches (Wedderburn, 1974; McCullagh, 1983) to longitudinal data by introducing a working correlation matrix for the observations for each subject, giving the GEEs. We follow this idea and introduce an extension of the extended GEE estimator in Chapter 3 for longitudinal data. The GEEs for the parameters $(\alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$ in the longitudinal LCMs are based on the mean and covariance functions (5.6) and (5.7). This method can be applied to not only counts but also continuous measurements and other data types. In the extended GEE method, the estimation of β and θ can be readily obtained from $\mathcal{P}^1$ and $\mathcal{Q}$, respectively. The estimation procedure is as follows:

α estimated by GEE in $\mathcal{P}$ The overall mean and covariance functions of $\mathcal{P}$ have been derived in (5.6) and (5.7). This leads to a set of GEEs for α in LCMs:

$$S_\alpha = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \alpha}' \Sigma_i^{-1}\{\mathbf{Y}_i - \boldsymbol{\Lambda}_i\} = 0. \quad (5.8)$$

In (5.8), both $\boldsymbol{\Lambda}_i$ and Σ_i are functions of α, which results in a complicated nonlinear form in the estimating equations. With starting values of α, we can solve (5.8) with fixed Σ_i with respect to α, then we update Σ_i with the new values of α, and we repeat this process until convergence. This results in a natural iterative estimating algorithm. Our experience shows that the iterative algorithm is more reliable and faster than estimating α by directly solving (5.8).

$\boldsymbol{\beta}$ and ψ_1 estimated by GEE in $\mathcal{P}^1$ As discussed earlier, $\mathcal{P}^1$ is a representative group of the latent at-risk class $\mathcal{P}_1$. Therefore, the parameters of the $\eta = 1$ class can be estimated from the longitudinal data of $\mathcal{P}^1$ alone. For example, the GEE estimator can be used to estimate β in the regression model (5.2) via the following estimating equations:

$$S_\beta = \sum_{i=1}^{n} \delta_i \frac{\partial \boldsymbol{\mu}_{1i}}{\partial \boldsymbol{\beta}}' \Sigma_{1i}^{-1}\{\mathbf{Y}_i - \boldsymbol{\mu}_{1i}\} = 0. \quad (5.9)$$

Liang and Zeger (1986) suggested using consistent moment estimates for ψ_1 in the covariance function. This yields an iterative scheme that switches between estimating β from (5.9) for fixed $\hat{\psi}_1$ and estimating ψ_1 for the moment estimator with fixed values of $\hat{\beta}$.

$\boldsymbol{\theta}$ and ψ_0 estimated by GEE in $\mathcal{Q}$ The latent not-at-risk $\eta = 0$ class has the same visit patterns as the general population. The general population is independent of the survivor cohort, i.e., $\mathcal{P} \perp \mathcal{Q}$. The parameters of the $\eta = 0$ class can be estimated from the longitudinal data $\mathcal{Q}$ alone. For example, the GEE estimator can be used to estimate θ in the regression model (5.4) via the following estimating equations (5.10), and ψ_0 in the covariance function (5.5) can be estimated by a consistent moment estimator.

$$S_\theta = \sum_{i=1}^{m} \frac{\partial \boldsymbol{\mu}_{0i}}{\partial \boldsymbol{\theta}}' \Sigma_{0i}^{-1} \{\mathbf{Y}_i - \boldsymbol{\mu}_{0i}\} = 0. \tag{5.10}$$

Asymptotic Properties Let the extended GEE estimator of $(\alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$ in the longitudinal LCM found by solving (5.8)–(5.10) be $(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\theta}}_L)$. It reduces to the extended GEE estimator in Chapter 3 when the independent working correlation matrix is used in both latent classes, i.e., $\Gamma_i(\cdot) = I_i(\cdot)$. We have established the asymptotic properties of the extended GEE estimator $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ in Chapter 3, i.e., $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ has consistency and asymptotic normality as $n \to \infty$ and $m \to \infty$, with a sandwich-form asymptotic variance. With the within-subject correlation explicitly specified in the extended GEEs (5.8)–(5.10), the asymptotic properties of $(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\theta}}_L)$ follow from those of $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ in Chapter 3, following the extension of Theorem 1 to Theorem 2 in Liang and Zeger (1986). As stated in Liang and Zeger (1986), both the GEE estimator of the regression parameters and its variance are consistent even when the independent working correlation is adopted, given only a correct specification of the regression. Taking the correlation into account increases efficiency while maintaining the other properties; see Theorem 2 in Liang and Zeger (1986). GEE methods for longitudinal data avoid the need for multivariate distributions by assuming a functional form for only the marginal mean and variance. The covariance structure across time is treated as a nuisance. The GEE estimator and its variance are consistent even if the covariance structure is misspecified, which is expected to happen often. Therefore, the extended GEE estimator $(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\theta}}_L)$ has an asymptotic normal distribution following the Chapter 3 results:

$$\sqrt{n}\Big((\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\theta}}_L)' - (\alpha, \boldsymbol{\beta}, \boldsymbol{\theta})'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Psi^{-1}\Phi(\Psi^{-1})'\Big),$$

where Φ is the asymptotic variance of the estimating functions $(S_\alpha, S_\beta, S_\theta)'$,

$$\Phi = \begin{bmatrix} \mathrm{V}(S_\alpha) & \mathrm{Cov}(S_\alpha, S_\beta) & 0 \\ \mathrm{Cov}(S_\alpha, S_\beta)' & \mathrm{V}(S_\beta) & 0 \\ 0 & 0 & \mathrm{V}(S_\theta) \end{bmatrix}, \text{ with zero covariances in } \Phi \text{ since } \mathcal{P} \perp \mathcal{Q}.$$

Here Ψ is the limited constant matrix of the first derivatives of $(S_\alpha, S_\beta, S_\theta)'$ with respect

to $(\alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$, $\Psi = \begin{bmatrix} \mathrm{E}(-\frac{\partial S_\alpha}{\partial \alpha}) & \mathrm{E}(-\frac{\partial S_\alpha}{\partial \boldsymbol{\beta}}) & \mathrm{E}(-\frac{\partial S_\alpha}{\partial \boldsymbol{\theta}}) \\ 0 & \mathrm{E}(-\frac{\partial S_\beta}{\partial \boldsymbol{\beta}}) & 0 \\ 0 & 0 & \mathrm{E}(-\frac{\partial S_\theta}{\partial \boldsymbol{\theta}}) \end{bmatrix}$. Since $\mathcal{P} \perp \mathcal{Q}$, the asymptotic variance of $\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L$, $\mathrm{AV}(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L)$ can be separated from $\mathrm{AV}(\hat{\boldsymbol{\theta}}_L)$ just as in Chapters 2 and 3. Then,

$$\mathrm{AV}(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L) = \frac{1}{n}\Psi_1^{-1}\Phi_1(\Psi_1^{-1})' + \frac{1}{m}\Psi_1^{-1}\Psi_2\mathrm{AV}(\hat{\boldsymbol{\theta}}_L)\Psi_2'(\Psi_1^{-1})', \tag{5.11}$$

where Φ_1 is the 2×2 block in the top left of Φ, Ψ_1 is the 2×2 block in the top left of Ψ, and Ψ_2 is the 2×1 block in the top right of Ψ.

We evaluate Φ and Ψ taking into account the within-subject correlation by following Liang and Zeger (1986). The evaluation borrows strength across the independence of the subjects to estimate a working correlation matrix and hence explicitly account for the time dependence to achieve greater asymptotic efficiency. Following Liang and Zeger (1986) we have

$$\mathrm{E}(\frac{\partial S_\alpha}{\partial \alpha}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \alpha}' \Sigma_i^{-1} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \alpha};$$

$$\mathrm{E}(\frac{\partial S_\alpha}{\partial \boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \alpha}' \Sigma_i^{-1} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \boldsymbol{\beta}};$$

$$\mathrm{E}(\frac{\partial S_\alpha}{\partial \boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \alpha}' \Sigma_i^{-1} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \boldsymbol{\theta}};$$

$$\mathrm{E}(\frac{\partial S_\beta}{\partial \boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\partial \boldsymbol{\mu}_{1i}}{\partial \boldsymbol{\beta}}' \Sigma_{1i}^{-1} \frac{\partial \boldsymbol{\mu}_{1i}}{\partial \boldsymbol{\beta}};$$

$$\mathrm{V}(S_\alpha) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \alpha}' \Sigma_i^{-1} \mathrm{cov}(\mathbf{Y}_i) \Sigma_i^{-1} \frac{\partial \boldsymbol{\Lambda}_i}{\partial \alpha};$$

$$\mathrm{V}(S_\beta) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\partial \boldsymbol{\mu}_{1i}}{\partial \boldsymbol{\beta}}' \Sigma_{1i}^{-1} \mathrm{cov}_1(\mathbf{Y}_i) \Sigma_{1i}^{-1} \frac{\partial \boldsymbol{\mu}_{1i}}{\partial \boldsymbol{\beta}};$$

$$\mathrm{Cov}(S_\alpha, S_\beta) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{\partial \boldsymbol{\Lambda}_i}{\partial \alpha}' \Sigma_i^{-1} \mathrm{cov}_1(\mathbf{Y}_i) \Sigma_{1i}^{-1} \frac{\partial \boldsymbol{\mu}_{1i}}{\partial \boldsymbol{\beta}}.$$

We obtain the estimate of the asymptotic variance $\mathrm{AV}(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L)$, $\widehat{\mathrm{AV}}(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L)$by replacing in (5.11) $\mathrm{cov}(\mathbf{Y}_i)$ by $(\mathbf{Y}_i - \boldsymbol{\Lambda}_i)(\mathbf{Y}_i - \boldsymbol{\Lambda}_i)'$, $\mathrm{cov}_1(\mathbf{Y}_i)$ by $(\mathbf{Y}_i - \boldsymbol{\mu}_{1i})(\mathbf{Y}_i - \boldsymbol{\mu}_{1i})'$, and the parameters by their estimates. The consistency of the extended GEE estimator and its variance depends only on the correct specification of the mean, not on the correct choice of the working correlation matrix. Moreover, the asymptotic variance estimation does not depend on the estimator of the nuisance parameters provided it is consistent.

If we assume that the GEE estimate of $\boldsymbol{\beta}$ from $\mathcal{P}^1$ and the GEE estimate of $\boldsymbol{\theta}$ from $\mathcal{Q}$ are true values, a naive way to estimate the asymptotic variance of $\hat{\alpha}_L$ is

via AV.naive$(\hat{\alpha}_L) = \frac{1}{n}\{\mathrm{E}(-\frac{\partial S_\alpha}{\partial \alpha})\}^{-1}\mathrm{V}(S_\alpha)\{\mathrm{E}(-\frac{\partial S_\alpha}{\partial \alpha})'\}^{-1}$. We evaluate and compare $\widehat{\mathrm{AV}}(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L)$ and $\widehat{\mathrm{AV}}$.naive$(\hat{\alpha}_L)$ in Section 5.3.1; they are called *sw.se* and *sw.se.naive*, respectively.

5.3 Analysis IV.A of CAYACS Physician Claims

Longitudinal analysis allows regression effects to change over time. We have seen in Chapter 4 that there is an obvious nonlinear trend overall and in the sex and diagnosis-age effects for both the yearly counts and costs. The trends may differ between the cohort and the population.

A natural longitudinal yearly binary variable can be constructed to indicate whether or not there are physician visits in the j^{th} year for subject i. There are 29023 observations (23%) in the population with no visits during the corresponding year, and 1844 observations (12%) in the cohort, and only 222 observations (9.7%) in the subset of the cohort with RSC. The longitudinal yearly binaries also carry some information about the later effects of survivors, so we also analyze them, but they are not our focus.

This section analyzes the CAYACS longitudinal yearly binaries, counts, and costs by the LCMs and the corresponding extended GEE inference procedure presented in Section 5.2. We then perform risk assessments of the cohort for later effects, and we classify the cohort into at-risk and not-at-risk classes in the next section.

We consider the following five variables as covariates in the risk model for η in (5.1): sex (male vs. female), SES (high vs. low), age at entry (five years after the diagnosis), diagnosis period (1990s vs. 1980s), and initial cancer treatment (chemotherapy but no radiation, radiation but no chemotherapy, both chemotherapy and radiation, or others). A standardized age value $(age - 5)/20$ was used in all the regression models. Of these covariates, only sex, SES, and age at entry are included in the regression models on Y_{ij}, (5.2) and (5.4), for each latent class. The linear predictor parts of (5.2) and (5.4) are specified as for (4.1), where the effects are time-varying. We choose the working correlation matrix $\Gamma_i(\cdot)$ in (5.3) and (5.5) as in Chapter 4. We adopt the model structure expression $Ma_0a_1a_2a_3.A$ based on the form of the linear predictor and the choice of $\Gamma_i(\cdot)$.

5.3.1 Results Under Latent Class Models

The extended GEE estimating procedure for longitudinal LCMs estimates the parameters in the $\eta = 0$ class from the general population, as in Section 5.2. Therefore, the results from Chapter 4 for longitudinal counts and costs for the population are the estimates for the $\eta = 0$ class of LCMs in this chapter.

We investigated *M0000.0/1/2*, *M1000.0/1/2*, *M1100.0/1/2*, *M1001.0/1/2*, *M1101.0/1/2*, and *M1111.0/1/2* for both yearly counts and costs for $\mathcal{P}^1$, which are

the estimates for the $\eta = 1$ class in the LCMs. Figures 5.3 and 5.4 (yearly counts) and 5.6 and 5.7 (yearly costs) compare the different models, as in Chapter 4. We see similar trends over time for both counts and costs compared to the analysis in Chapter 4. Figure 5.3 compares *M0000*, *M1000*, *M1100*, *M1001*, *M1101*, and *M1111* under a compound symmetric correlation structure. It shows that sex and age at entry have clear time-varying effects on the yearly counts and the intercept. The SES effect is almost time-independent and not significant, except for the final two years, when the sample sizes are small and unreliable. We also study *M1101* under different correlation structures; see Figure 5.4. The results are similar, except that the compound symmetric correlation gives slightly more efficient estimates. Therefore, *M1101.1* is more informative about visit trends over time for both the at-risk and not-at-risk classes. Figure 5.5 compares the time-varying coefficients of the yearly counts in both latent classes under *M1101.1*. The at-risk class has more frequent visits than the not-at-risk class throughout the follow-up period. Both time-varying intercepts first decrease and then increase, but the intercept of the not-at-risk class increases overall, and that of the at-risk class fluctuates around the constant estimate.

Figures 5.6 and 5.7 compare different models for the yearly costs for the $\mathcal{P}^1$ data; the results are similar. Figure 5.8 compares the time-varying coefficients of the yearly costs in both latent classes under the preferable model, *M1101.1*. The at-risk class has higher medical costs than the not-at-risk class throughout the follow-up period, except for the last year. The intercept of the not-at-risk class increases overall, but that of the at-risk class tends to slightly decrease. The time-varying sex effects are similar in both classes. An interesting discrepancy between the two classes is that the time-varying age at entry effect decreases overall in the not-at-risk class but increases in the at-risk class.

Tables 5.2 to 5.5 present the extended GEE estimates and their standard errors (*sw.se*) of $(\alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$ in the longitudinal LCMs for *M0000.1*, *M1001.1*, *M1100.1*, and *M1101*, respectively. When the coefficients are time-varying, their averages and the standard errors of the averages are reported. The yearly binary data provide limited information, so they appear only in Tables 5.2 and 5.3 under the simplest all-constant coefficients, *M0000.1* and *M1001.1*, where only the intercept and age effects are time-varying, respectively. When the sex effect is constant in the regression models of the yearly counts for both latent classes, it is not significant for the at-risk class; see the middle panel of Tables 5.2 and 5.3. On the other hand, the average time-varying effect of sex is significant in Tables 5.4 and 5.5. Figure 5.3 shows that the effect of sex changes over time in a curvilinear fashion. The constant coefficient of sex averages out the changes over time, which results in sex being a significant risk factor for later effects of survivors (Table 5.3). This is the only disagreement with the other results in the tables. Table 5.4 compares the standard error estimates for the extended GEE estimators of α, *sw.se*, and *sw.se.naive* introduced in Section 5.2, for the yearly counts and costs under *M1100.1*. The estimated standard error *sw.se* can be either greater or smaller than *sw.se.naive* which is derived under the assumption that the GEE

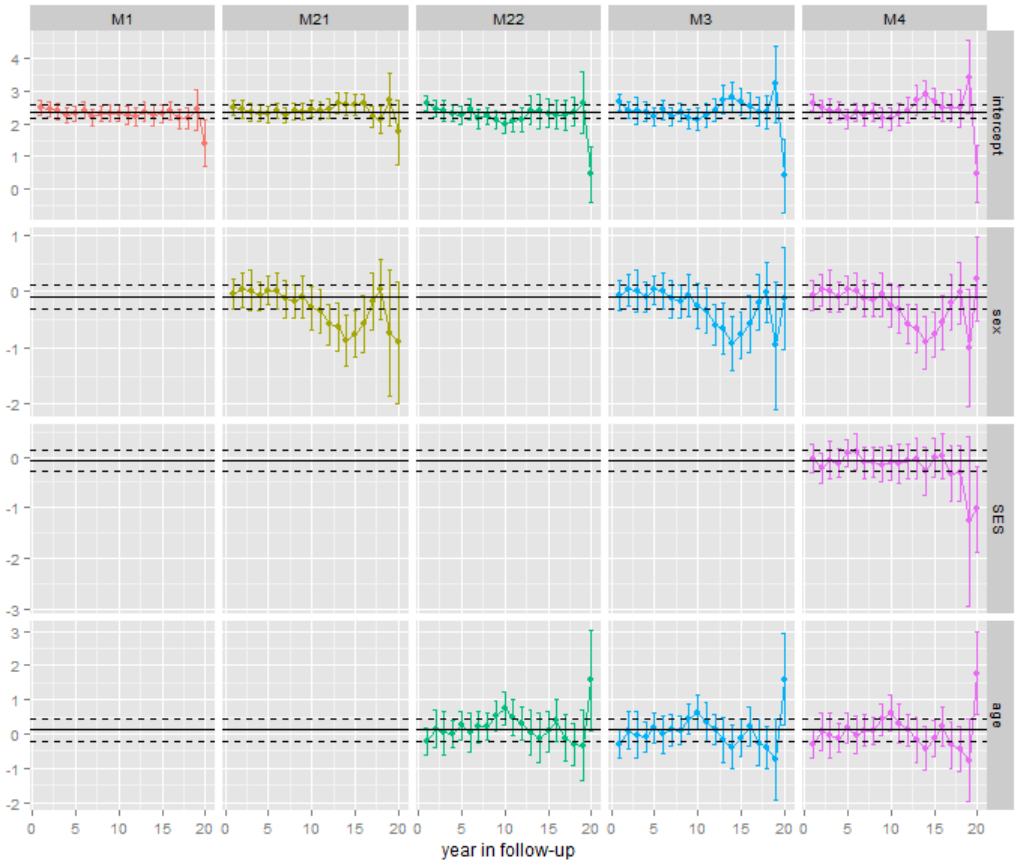

Figure 5.3: Time-dependent coefficients of yearly counts for at-risk class under CS correlation structure.

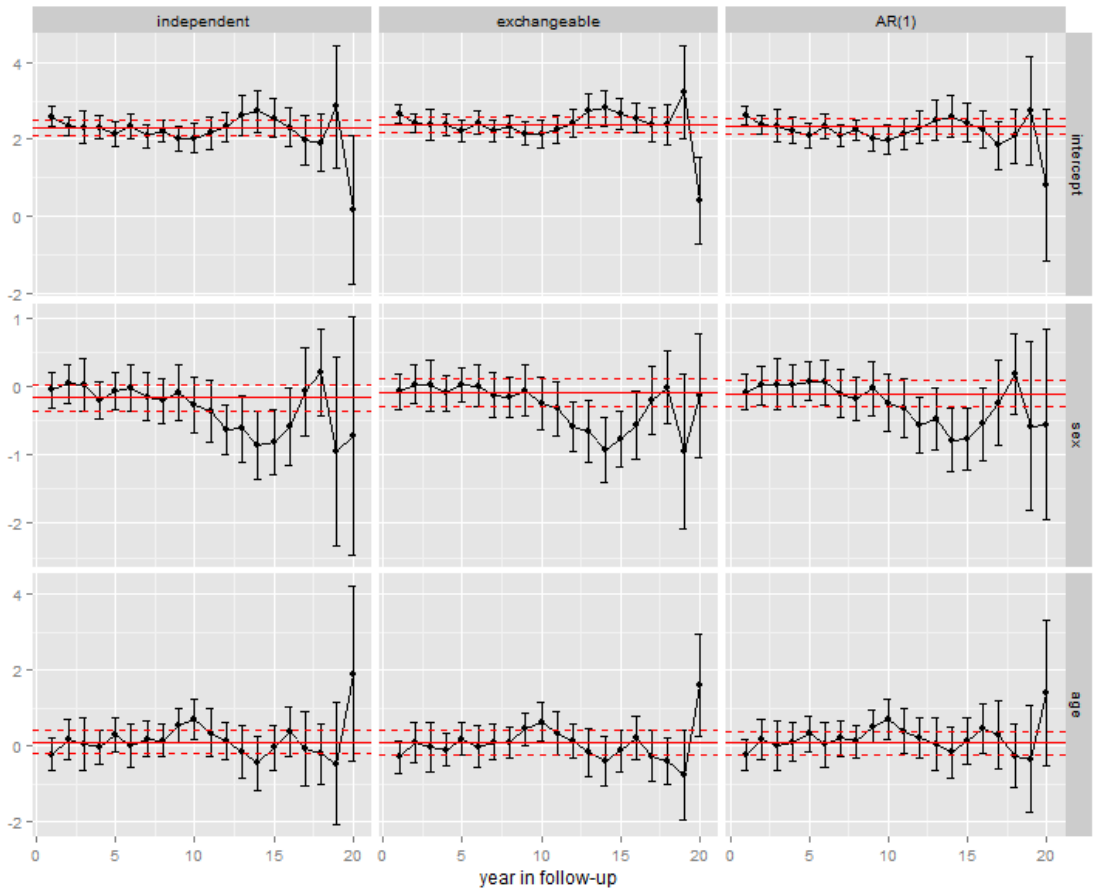

Figure 5.4: Time-dependent coefficients of yearly counts for at-risk classs under *M1101*.

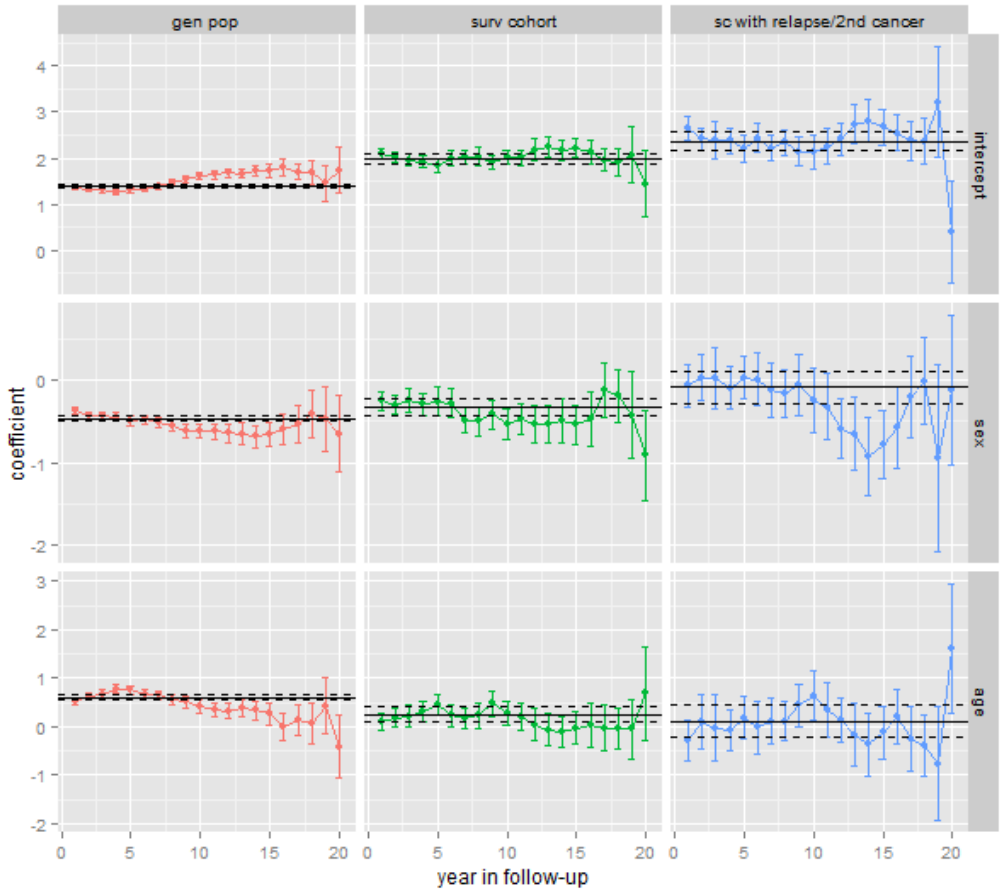

Figure 5.5: Time-dependent coefficients of yearly counts under *M1101.1*: not-at-risk class vs. full SC vs. at-risk class.

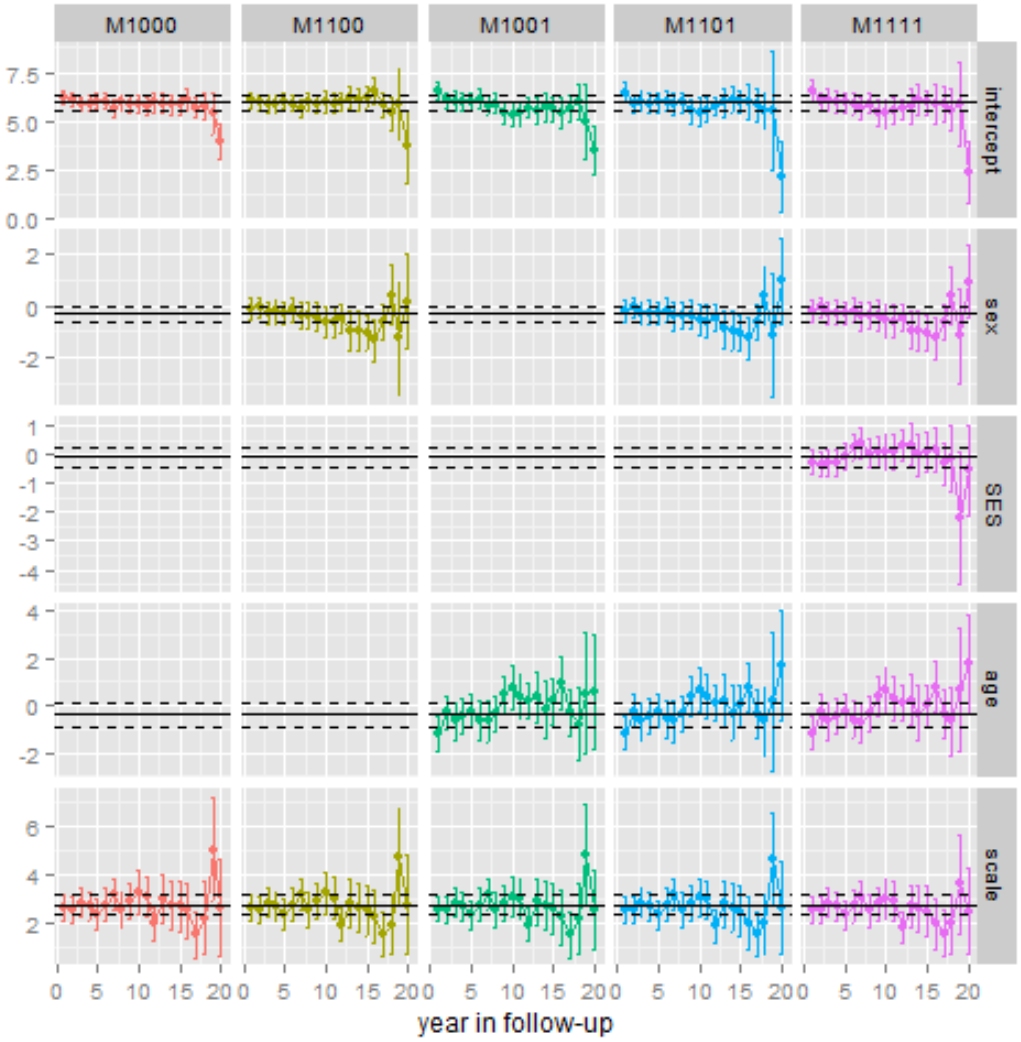

Figure 5.6: Time-dependent coefficients of yearly costs for at-risk class under CS correlation structure.

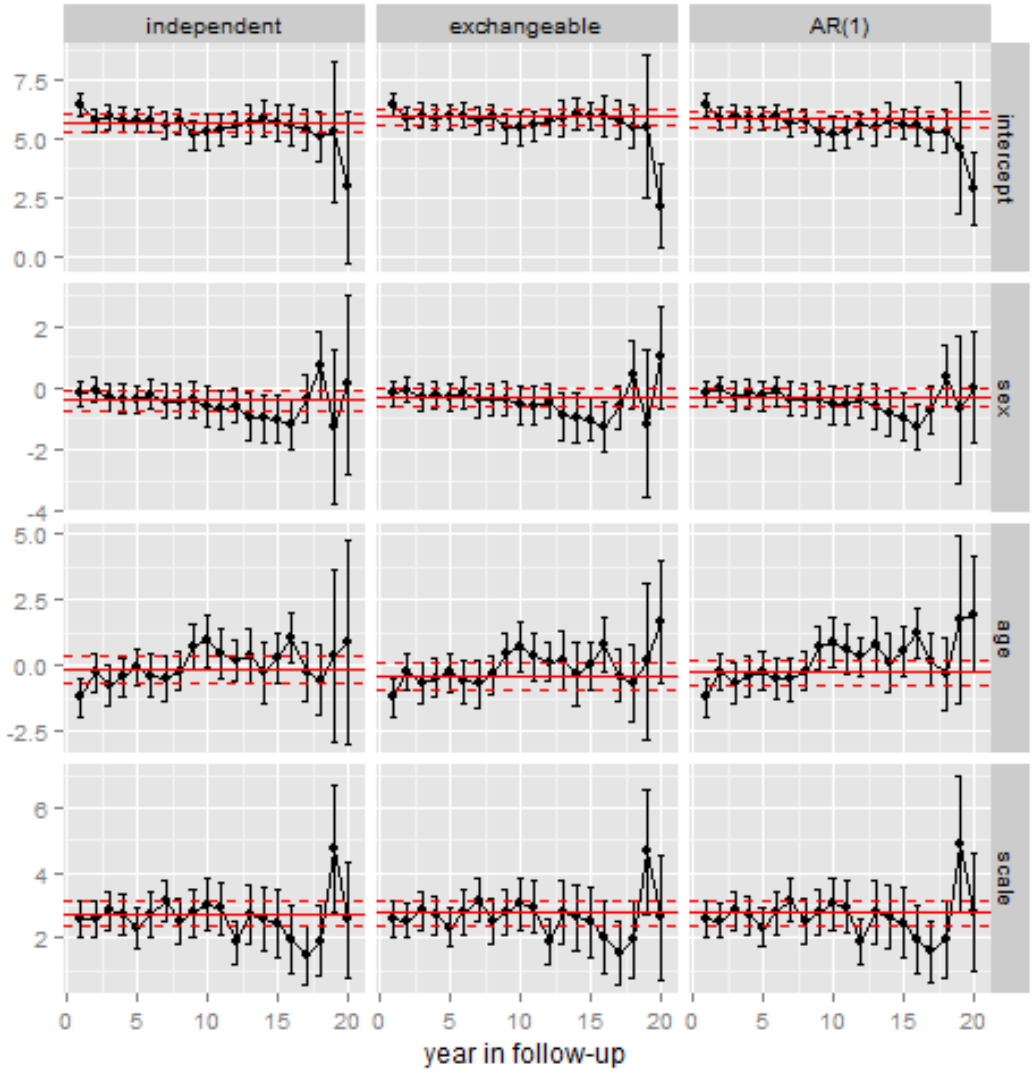

Figure 5.7: Time-dependent coefficients of yearly costs for at-risk class under *M1101*.

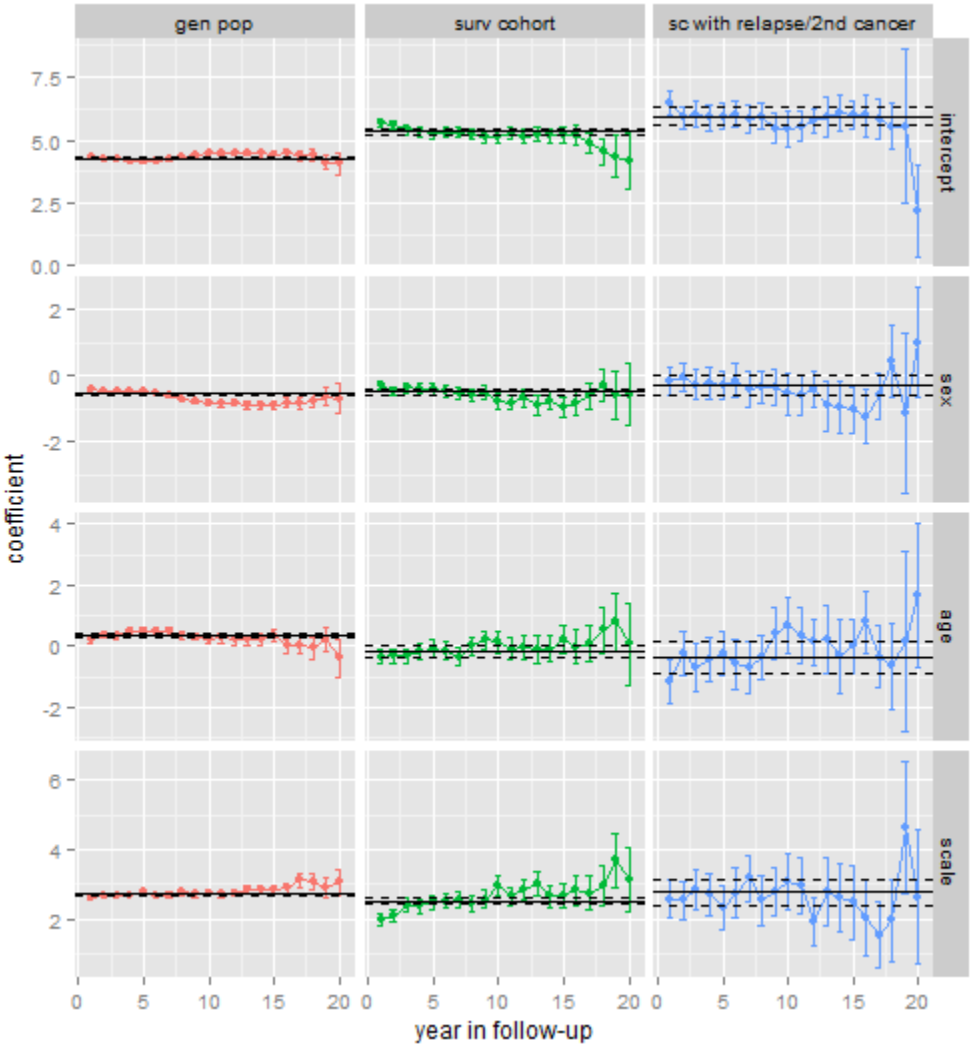

Figure 5.8: Time-dependent coefficients of yearly costs under *M1101.1*: not-at-risk class vs. full SC vs. at-risk class.

estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are true values. Interestingly, if *sw.se.naive* is used for the costs, both sex and SES are significant factors for later effects in the risk model, which does not agree with the other results.

Table 5.5 considers the preferred model *M1101*. We contrast the extended GEE estimates and their standard errors under *M1101.1* and *M1101.2*, under compound symmetric and AR(1) correlation structures, for both yearly counts and costs. Both correlation structures lead to the same significant factors for later effects. Figure 5.8 compares the time-varying coefficients in the at-risk (blue) and not-at-risk (red) classes for yearly costs. We can see that the *sex* effects over time are similar in both latent classes. Females need more medical care especially during about 8 to 17 years after follow-up, which can be their pregnancy period for most women. Therefore, in the corresponding risk model for η, *sex* is not a significant factor for later effects. Both the count and cost analyses found a significantly higher risk of later effects associated with *treatment with radiation therapy rather than other treatments*. The analysis of counts also identified *diagnosis in 1980s rather than 1990s* and *treatment with chemo but no radiation rather than other treatments* as significant risk factors of later effects. This indicates that different choices of the metrics for classification can lead to the identification of different risk factors. In our context, the results show that survivors diagnosed in the 1990s have fewer physician visits than those diagnosed in the 1980s, but the yearly medical costs may not be lower, although we adjusted for inflation. This may be because new treatments are more expensive. The survivors who initially received chemotherapy but not radiation may see physicians more often but do not necessarily cost more than the other survivors.

5.3.2 Results Under Latent Class Models: Subject-Specific Modelling

In the regression models for longitudinal response $\mathbf{Y}$ in each of the latent classes, even for the same covariates, the regression coefficients can differ from subject to subject because of heterogeneity. Such models are called subject-specific models or mixed effects models, which are another extension of GLMs for longitudinal data (Diggle *et al.*, 2002). Contrasted to marginal models, mixed-effects models specify only one regression model for subject-specific mean and the within-subject correlation is induced from random effects. Marginal models for continuous measurements are equivalent in coefficients to linear mixed effects models and have a determined relationship in the variance-covariance functions (Fitzmaurice *et al.*, 2012). Therefore, we can estimate the random effects for each latent class after estimating the marginal models by the GEE estimators from Section 5.3.1.

In contrast with marginal models, (5.2) and (5.3) for the at-risk class, in mixed effects models, the covariate effects and within-subject association are modelled through a single equation by introducing random effects b_i for each subject i in the $\eta = 1$ class,

$$l\{E(Y_{ij}|\eta_i = 1, \mathbf{Z}_{ij}, b_i)\} = \beta_j^{'}\mathbf{Z}_{ij} + b_i^{'}\mathbf{X}_{ij}, \tag{5.12}$$

Table 5.2: Analysis of Yearly Binaries vs. Counts vs. Costs by LCMs: Under *M0000.1* Compound Symmetric[a]

	M0000.1					
	binaries		counts		costs	
Factor	*estimate*	*sw.se*	*estimate*	*sw.se*	*estimate*	*sw.se*
α estimates in the Risk Model						
intercept	-0.123	(0.851)	-0.303	(0.464)	0.134	(0.356)
male (vs. female)	-0.226	(0.772)	-0.443	(0.326)	-0.260	(0.285)
SES high (vs. low)	0.914	(1.027)	0.337	(0.322)	0.292	(0.290)
age at diagnosis	0.844	(1.294)	-0.450	(0.530)	-0.498	(0.455)
diag in 90s (vs. 80s)	0.764	(0.485)	**-0.513**	(0.259)	0.199	(0.173)
treatment (vs. other)						
chemo no rad	**0.980**	(0.538)	**0.572**	(0.279)	0.304	(0.201)
rad no chemo	3.061	(3.275)	**1.010**	(0.415)	**1.520**	(0.470)
both	**1.698**	(0.897)	**1.282**	(0.382)	**1.014**	(0.269)
β estimates in the Regression Model for the At-risk Class						
	GEE estimates based on $\delta = 1$ subgroup					
intercept	**3.133**	(0.435)	**2.367**	(0.105)	**5.944**	(0.188)
male (vs. female)	**-0.665**	(0.324)	-0.095	(0.103)	**-0.305**	(0.162)
SES high (vs. low)	-0.328	(0.295)	-0.065	(0.106)	-0.102	(0.166)
standardized age	-0.657	(0.494)	0.111	(0.169)	-0.370	(0.263)
dispersion/scale parameter	**1.12**	(0.587)	**11.1**	(1.4)	**2.770**	(0.195)
correlation parameter	0.36	(0.235)	**0.318**	(0.038)	**0.393**	(0.049)
θ estimates in the Regression Model for the Not-at-risk Class						
	GEE estimates based on general population					
intercept	**1.466**	(0.029)	**1.389**	(0.018)	**4.277**	(0.021)
male (vs. female)	**-0.559**	(0.025)	**-0.469**	(0.018)	**-0.558**	(0.019)
SES high (vs. low)	-0.043	(0.024)	**-0.067**	(0.018)	**-0.051**	(0.020)
standardized age	-0.022	(0.040)	**0.604**	(0.030)	**0.348**	(0.032)
dispersion/scale parameter	**0.958**	(0.009)	**10.3**	(0.582)	**2.723**	(0.014)
correlation parameter	**0.214**	(0.005)	**0.384**	(0.013)	**0.332**	(0.005)

[a]Significant effect with p-value ≤ 0.05 in **Boldface**.

Table 5.3: Analysis of Yearly Binaries vs. Counts vs. Costs by LCMs: Under *M1001.1* Compound Symmetric[a]

	M1001.1					
	binaries		counts		costs	
Factor	*estimate*	*sw.se*	*estimate*	*sw.se*	*estimate*	*sw.se*
α estimates in the Risk Model						
intercept	0.529	(0.503)	0.004	(0.402)	0.191	(0.318)
male (vs. female)	-0.218	(0.414)	**-0.503**	(0.296)	-0.313	(0.250)
SES high (vs. low)	0.624	(0.529)	0.304	(0.300)	0.270	(0.254)
age at diagnosis	-0.374	(0.697)	-0.501	(0.518)	-0.256	(0.395)
diag in 90s (vs. 80s)	**0.724**	(0.425)	**-0.798**	(0.241)	0.028	(0.179)
treatment (vs. other)						
chemo no rad	0.427	(0.392)	**0.556**	(0.233)	0.278	(0.182)
rad no chemo	0.715	(0.692)	**1.310**	(0.397)	**1.707**	(0.528)
both	**0.853**	(0.505)	**1.348**	(0.354)	**0.927**	(0.242)
β estimates in the Regression Model for the At-risk Class						
	GEE estimates based on $\delta = 1$ subgroup					
intercept	**3.190**[b]	(0.542)[c]	**2.220**[b]	(0.127)[c]	**5.630**[b]	(0.208)[c]
male (vs. female)	**-0.686**	(0.334)	-0.096	(0.102)	**-0.306**	(0.155)
SES high (vs. low)	-0.342	(0.312)	-0.067	(0.107)	-0.102	(0.159)
standardized age	-0.749[b]	(0.602)[c]	0.199[b]	(0.183)[c]	-0.081[b]	(0.288)[c]
dispersion/scale parameter	1.17	(1.11)	**10.8**	(1.35)	**2.690**[b]	(0.225)[c]
correlation parameter	0.387	(0.385)	**0.329**	(0.040)	**0.398**	(0.048)
θ estimates in the Regression Model for the Not-at-risk Class						
	GEE estimates based on general population					
intercept	**1.350**[b]	(0.035)[c]	**1.502**[b]	(0.033)[c]	**4.237**[b]	(0.029)[c]
male (vs. female)	**-0.567**	(0.025)	**-0.494**	(0.019)	**-0.549**	(0.019)
SES high (vs. low)	-0.041	(0.025)	**-0.067**	(0.019)	**-0.050**	(0.020)
standardized age	0.009[b]	(0.054)[c]	**0.435**[b]	(0.059)[c]	**0.253**[b]	(0.047)[c]
dispersion/scale parameter	**0.955**	(0.009)	**9.97**	(0.529)	**2.818**[b]	(0.025)[c]
correlation parameter	**0.217**	(0.005)	**0.38**	(0.013)	**0.334**	(0.005)

[a]Significant Effect with P-value ≤ 0.05 in **Boldface**.

[b]Average values over 20 estimates

[c]se of the 20 averaged estimates

Table 5.4: Analysis of Yearly Counts and Yearly Costs by LCMs: *sw.se* vs. *sw.se.naive* Under *M1100.1* Compound Symmetric[a]

	M1100.1					
	counts			costs		
Factor	*estimate*	*sw.se*	*sw.se.naive*	*estimate*	*sw.se*	*sw.se.naive*
α estimates in the Risk Model						
intercept	-0.231	(0.423)	(0.347)	0.209	(0.329)	(0.235)
male (vs. female)	-0.312	(0.337)	(0.236)	*-0.279*	(0.258)	(0.163)
SES high (vs. low)	0.341	(0.329)	(0.254)	*0.299*	(0.257)	(0.164)
age at diagnosis	0.367	(0.544)	(0.381)	-0.425	(0.400)	(0.267)
diag in 90s (vs. 80s)	**-1.105**	(0.260)	(0.255)	0.009	(0.174)	(0.160)
treatment (vs. other)						
chemo no rad	**0.534**	(0.259)	(0.263)	0.293	(0.187)	(0.186)
rad no chemo	**1.476**	(0.474)	(0.479)	**1.812**	(0.518)	(0.478)
both	**1.498**	(0.393)	(0.365)	**0.991**	(0.244)	(0.228)
β estimates in the Regression Model for the At-risk Class						
	GEE estimates based on $\delta = 1$ subgroup					
intercept	**2.399**[b]	(0.116)[c]		**5.839**[b]	(0.214)[c]	
male (vs. female)	**-0.312**[b]	(0.123)[c]		**-0.482**[b]	(0.203)[c]	
SES high (vs. low)	-0.062	(0.109)		-0.096	(0.158)	
standardized age	0.014	(0.174)		-0.380	(0.250)	
dispersion/scale parameter	**10.697**	(1.307)		**2.690**[b]	(0.223)[c]	
correlation parameter	**0.324**	(0.041)		**0.399**	(0.048)	
θ estimates in the Regression Model for the Not-at-risk Class						
	GEE estimates based on general population					
intercept	**1.440**[b]	(0.026)[c]		**4.269**[b]	(0.027)[c]	
male (vs. female)	**-0.506**[b]	(0.039)[c]		**-0.687**[b]	(0.029)[c]	
SES high (vs. low)	**-0.063**	(0.019)		**-0.049**	(0.020)	
standardized age	**0.569**	(0.031)		**0.347**	(0.032)	
dispersion/scale parameter	**10.034**	(0.534)		**2.804**[b]	(0.026)[c]	
correlation parameter	**0.379**	(0.013)		**0.333**	(0.005)	

[a]Significant effect with p-value ≤ 0.05 in **Boldface**.

[b]Average values over 20 estimates

[c]se of the 20 averaged estimates

Table 5.5: Analysis of Yearly Counts and Yearly Costs by LCMs: Under *M1101* Compound Symmetric vs. AR(1)[a]

	M1101.1				*M1101.2*			
	counts		costs		counts		costs	
Factor	*estimate*	*sw.se*	*estimate*	*sw.se*	*estimate*	*sw.se*	*estimate*	*sw.se*
α estimates in the Risk Model								
intercept	0.179	(0.435)	0.196	(0.314)	-0.120	(0.488)	0.389	(0.350)
male (vs. female)	-0.329	(0.341)	-0.286	(0.247)	-0.217	(0.331)	-0.155	(0.262)
SES high (vs. low)	0.365	(0.342)	0.280	(0.248)	0.123	(0.322)	0.224	(0.260)
age at diagnosis	0.097	(0.590)	-0.302	(0.389)	-0.014	(0.539)	-0.232	(0.422)
diag in 90s (vs. 80s)	**-1.347**	(0.283)	0.017	(0.178)	**-1.161**	(0.277)	-0.237	(0.199)
treatment (vs. other)								
chemo no rad	**0.474**	(0.246)	0.269	(0.181)	**0.769**	(0.288)	0.234	(0.197)
rad no chemo	**1.524**	(0.525)	**1.729**	(0.509)	**1.228**	(0.423)	**1.566**	(0.575)
both	**1.463**	(0.413)	**0.946**	(0.241)	**1.514**	(0.406)	**0.961**	(0.282)
β estimates in the Regression Model for the At-risk Class								
	GEE estimates based on $\delta = 1$ subgroup							
intercept	**2.360**[b]	(0.128)[c]	**5.664**[b]	(0.232)[c]	**2.203**[b]	(0.148)[c]	**5.479**[b]	(0.217)[c]
male (vs. female)	**-0.293**[b]	(0.124)[c]	**-0.421**[b]	(0.201)[c]	**-0.257**[b]	(0.128)[c]	**-0.420**[b]	(0.197)[c]
SES high (vs. low)	-0.078	(0.111)	-0.094	(0.159)	-0.050	(0.107)	-0.087	(0.156)
standardized age	0.070[b]	(0.186)[c]	-0.071[b]	(0.287)[c]	0.208[b]	(0.186)[c]	0.244[b]	(0.269)[c]
dispersion parameter	**10.59**	(1.302)	**2.641**[b]	(0.224)[c]	**10.92**	(1.330)	**2.659**[b]	(0.220)[c]
correlation parameter	**0.331**	(0.042)	**0.401**	(0.048)	**0.739**	(0.034)	**0.798**	(0.029)
θ estimates in the Regression Model for the Not-at-risk Class								
	GEE estimates based on general population							
intercept	**1.537**[b]	(0.036)[c]	**4.324**[b]	(0.032)[c]	**1.468**[b]	(0.034)[c]	**4.299**[b]	(0.031)[c]
male (vs. female)	**-0.546**[b]	(0.040)[c]	**-0.697**[b]	(0.030)[c]	**-0.553**[b]	(0.040)[c]	**-0.724**[b]	(0.029)[c]
SES high (vs. low)	**-0.062**	(0.019)	**-0.049**	(0.020)	**-0.063**	(0.019)	**-0.053**	(0.020)
standardized age	**0.399**[b]	(0.060)[c]	**0.235**[b]	(0.047)[c]	**0.439**[b]	(0.058)[c]	**0.314**[b]	(0.046)[c]
dispersion parameter	**10.029**	(0.537)	**2.801**[b]	(0.025)[c]	**10.338**	(0.574)	**2.802**[b]	(0.025)[c]
correlation parameter	**0.381**	(0.013)	**0.333**	(0.005)	**0.776**	(0.009)	**0.725**	(0.003)

[a] Significant effect with p-value ≤ 0.05 in **Boldface**.

[b] Average values over 20 estimates

[c] se of the 20 averaged estimates

where $\mathbf{X}_{ij}$ is a subset of $\mathbf{Z}_{ij}$. The mixed effects models explicitly distinguish between-subject and within-subject sources of variability (Fitzmaurice *et al.*, 2012). When the link function $l(\cdot)$ is the identity function, (5.12) is a linear mixed effects model. The part of population-averaged mean of Y_{ij} in (5.12), $\beta_j^{'}\mathbf{Z}_{ij}$, is the same as in (5.2), and the $\boldsymbol{\beta}$'s have the same interpretation. The second part in the right hand side of (5.12) is the additional subject-specific mean for i in the at-risk class. Assume we have the random effects b_i with mean 0's and variance-covariance G_1, and a vector of residuals $\mathbf{e}_i$ of (5.12) with mean 0's and diagonal variance $R_{1i} = \text{Var}(\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i)$, and that b_i and $\mathbf{e}_i$ are independent of each other. When the linear mixed effects model specified for the $\eta = 1$ class, the variance of $\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i$ can be derived as $\Sigma_{1i} = \mathbf{X}_i G_1 \mathbf{X}_i^{'} + R_{1i}$, a compound symmetric form; see the derivation of Fitzmaurice *et al.* (2012). In this form, G_1 introduces the within-subject correlation and R_{1i} presents the between-subject variation. Thus, the introduction of random effects, b_i, for the $\eta = 1$ class induces correlation among the components of $\mathbf{Y}_i$, for which the variance-covariance function has a determined relationship with a compound symmetric specification in the marginal model (5.3). Therefore, for continuous responses, the marginal model specification (5.2) and (5.3) with a compound symmetric correlation and the the linear mixed effects model specification (5.12) are equivalent. After the model parameters are estimated, it is possible to obtain predictions of the subject-specific effects, b_i, or of the subject-specific response trajectories, $\beta_j^{'}\mathbf{Z}_{ij} + b_i^{'}\mathbf{X}_{ij}$, from the linear mixed effects model. The best linear unbiased predictor (BLUP) of b_i is $E(b_i|\eta_i = 1, \mathbf{Y}_i, \mathbf{Z}_i) = G_1\mathbf{X}_i^{'}\Sigma_{1i}^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\beta}}^{'}\mathbf{Z}_i)$. Replacing the estimated parameters in the variance-covariance function, the resulting predictor, the empirical BLUP, is

$$\hat{b}_i = \hat{G}_1\mathbf{X}_i^{'}\hat{\Sigma}_{1i}^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\beta}}^{'}\mathbf{Z}_i). \tag{5.13}$$

These relationships between marginal models and mixed effects models are based on the linear functional form, which is not the case for GLMMs. For nonlinear link functions, the fixed effects in GLMMs are not comparable to the regression parameters in marginal models. One exceptional case where they are almost comparable is for repeated counts data when a log-linear model is adopted and the model has a single random intercept (Fitzmaurice *et al.*, 2012). The compound symmetric specification for within-subject association is equivalent to the random intercept model (Lee and Nelder, 2004), $\log\{E(Y_{ij}|\eta_i = 1, \mathbf{Z}_{ij}, b_i)\} = \beta_0 + b_i + \beta_1^{'}\mathbf{Z}_{ij}$, where b_i is a singular random effect and $\beta = (\beta_0, \beta_1^{'})^{'}$.

Mixed effects models for the $\eta = 0$ class can be introduced as

$$l\{E(Y_{ij}|\eta_i = 0, \mathbf{Z}_{ij}, c_i)\} = \theta_j^{'}\mathbf{Z}_{ij} + c_i^{'}\mathbf{X}_{ij}. \tag{5.14}$$

The relationships between the marginal models (5.4) and (5.5) and the mixed effects models (5.14) are the same as for the $\eta = 1$ class. In particular, when the link function $l(\cdot)$ is the identity function and (5.14) follows the same assumptions and notation as in the $\eta = 1$

class, the variance of $\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i$ can be derived as $\Sigma_{0i} = \mathbf{X}_i G_0 \mathbf{X}_i^{'} + R_{0i}$ and the empirical BLUP of the random effects c_i in the $\eta = 0$ class is

$$\hat{c}_i = \hat{G_0}\mathbf{X}_i^{'}\hat{\Sigma}_{0i}^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\theta}}^{'}\mathbf{Z}_i). \tag{5.15}$$

We specify the population-averaged part of linear mixed effects models (5.12) and (5.14) for each latent class and the risk model for η in the LCMs for yearly costs the same as in Section 5.3.1. As an illustration, the linear mixed effects models for yearly costs are specified with only a random intercept:

$$E(Y_{ij}|\eta_i = 1, \mathbf{Z}_i, b_i) = \beta_{0j} + b_i + \beta_{1j} sex_i + \beta_{2j} SES_i + \beta_{3j} age_i \tag{5.16}$$

$$E(Y_{ij}|\eta_i = 0, \mathbf{Z}_i, c_i) = \theta_{0j} + c_i + \theta_{1j} sex_i + \theta_{2j} SES_i + \theta_{3j} age_i \tag{5.17}$$

As discussed earlier, these LCMs with random effects are equivalent to those ones with CS correlation marginal specifications for yearly costs. The parameters in variance functions have a simple determined relationship. For the $\eta = 1$ class, $G_1 = \phi_1\sigma_1$ and $R_i = \phi_1(1-\sigma_1)I_{J_i}$, where I_{J_i} is a $J_i \times J_i$ identity matrix; same for the $\eta = 0$ class. The two kinds of models have these relationships theoretically, while numerically, linear mixed effects models are estimated by MLE and marginal models are estimated by GEE methods. For example, Tables 5.2 and 5.5 included the LCMs for yearly costs under *M0000.1* and *M1101.1* estimated by the extended GEE method. We estimated the regression models (5.16) and (5.17) by MLE in R, and the results are similar to the corresponding equivalent models in Tables 5.2 and 5.5. The differences only appear at about the third digit after decimal point. Therefore, the LCMs with random effects under *M0000.1* and *M1101.1* have the same estimates of regression parameters in Tables 5.2 and 5.5 and the estimated variance parameters can be calculated from the scale and correlation parameters in those two tables. Under *M0000.1*, the within-subject and between-subject variations are 1.09 and 1.68 respectively for the at-risk class, and 0.90 and 1.82 respectively for the not-at-risk class. Under *M1101.1*, the within-subject and between-subject variations are 1.06 and 1.58 respectively for the at-risk class, and 0.93 and 1.89 respectively for the not-at-risk class. We conduct risk classification under the subject-specific models based on these two models in the next section.

5.4 Analysis IV.B of CAYACS Physician Claims: Risk Classification/Prediction in Cohort by Yearly Costs

One of CAYACS's important goals is to conduct risk classification/prediction in the survivor cohort. This section classifies the survivor cohort into at-risk and not-at-risk classes based on analysis of yearly costs. We consider two types of classification procedures, one using

estimated subject-specific means (Section 5.4.1) and the other via three estimators for the risk probability (Section 5.4.2). An approximate ROC method is proposed in Section 5.4.3 to evaluate the performance of the second type of risk classification procedures. In addition, Section 5.4.4 presents estimates of the risk probability for the individuals with their dynamically updated information. It exemplifies how to use the classification procedures of Section 5.4.2 adaptively for prediction. The proposed risk prediction procedures are applied and evaluated in Section 5.4.5 using the additional RSC information collected during the the study follow-up.

5.4.1 Risk Classification by Subject-Specific Mean

This subsection demonstrate risk classification by subject-specific means estimated from the LCM under *M0000*, where all coefficients are time-independent for the yearly costs. We make use of the merit of longitudinal analysis to conduct risk classification based on the subject-specific effects b_i and c_i, i.e., to predict η_i given $\mathbf{Z}_i$, b_i and c_i. The empirical BLUP formula (5.13) and (5.15) show that $\hat{b}_i$ and $\hat{c}_i$ for each subject are linear combinations of multiple responses from subject i, given $\eta = 1$ or $\eta = 0$ respectively. They summarize subject-specific information of responses $\mathbf{Y}_i$ after adjusted by the covariates. We develop a risk classification strategy based on subject-specific effects of the LCM from the idea of Fisher's (linear or quadratic) discriminant analysis for multivariate variables; for example, see Jobson (2012). We can consider the risk classification problem is to seek partitioning the longitudinal responses $\mathbf{Y}$ into at-risk and not-at-risk regions, also adjusted by covariates $\mathbf{Z}$. Fisher's idea was to transform multivariate variables to a univariate variable in a linear function form. For longitudinal responses $\mathbf{Y}$ whose dimension is different from subject to subject, we can summarize them to subject-specific mean given covariates $\mathbf{Z}$.

To evaluate distributions of means with random effects for the two classes, the continuous covariate *age at entry*, which is equivalent to diagnosis age for survivors, is truncated into 4 categories, 0 to 5, 5 to 10, 10 to 15 and 15 to 20, and the mean of each category is used to calculate the means of responses. Therefore, we have 16 covariate combination cohorts (2 for sex, 2 for SES and 4 for age). The parameters in (5.17) are estimated from the general population, and the distribution of estimated subject-specific mean of the not-at-risk class $\hat{E}(Y_{ij}|\eta_i = 0, \mathbf{Z}_i, c_i)$ is evaluated for each of the 16 cohorts. The distribution of estimated subject-specific mean of the at-risk class $\hat{E}(Y_{ij}|\eta_i = 1, \mathbf{Z}_i, b_i)$ in (5.16) is also evaluated for each of the 16 cohorts according to estimates of the parameters from the RSC subgroup of survivor cohort. The distributions of $\hat{E}(Y_{ij}|\eta_i = 0, \mathbf{Z}_i, c_i)$ (green) vs. $\hat{E}(Y_{ij}|\eta_i = 1, \mathbf{Z}_i, b_i)$ (red) are contrasted for each of the 16 cohorts in each plot of Figure 5.9, respectively.

Each of these plots depicts the subject-specific means of the at-risk and not-at-risk classes given specific covariates. The membership of an individual in the survivor cohort is unobserved, while his/her subject-specific mean can be evaluated by the LCM and compare to see which class he/she is closer to. Formally speaking, it is from the idea of linear

discriminant analysis (LDA) which starts from the Bayesian perspective: we want to find the distribution of η given a subject-specific mean, while we have distributions of the subject-specific means given classes. A prior distribution of classes in the population need to be assumed, which is π_1, π_0 and $\pi_1 + \pi_0 = 1$ in this risk classification problem. Typically, when we choose $\pi_1 = \pi_0 = 0.5$ regardless of $\mathbf{Z}$, the same density value of the two distributions points the decision boundary for classification on x-axis (blue lines) in Figure 5.9. The range of these decision boundary values is from 4.62 to 5.23 among the 16 cohorts.

For any individual i in the survivor cohort, his/her subject-specific mean by the LCM is calculate by $\hat{E}(\mathbf{Y}_i|\mathbf{Z}_i, b_i, c_i) = p(\mathbf{Z}_i;\hat{\alpha})[\mu_1(\mathbf{Z}_i;\hat{\beta}) + \hat{b}_i] + \{1 - p(\mathbf{Z}_i;\hat{\alpha})\}[\mu_0(\mathbf{Z}_i;\hat{\theta}) + \hat{c}_i]$, where $\hat{b}_i$ and $\hat{c}_i$ are evaluated from the empirical BLUP (5.13) and (5.15), respectively. This value is then compared to the decision boundary given corresponding covariates in Figure 5.9. When it is larger than the boundary value, this subject belongs to the at-risk class; the not-at-risk class, otherwise. The risk classification result for the survivor cohort is listed in Table 5.6 and compared to the RSC status. The false negative rate is 26% (63/237) when 56% of the survivors are classified in the *at-risk* class.

This risk classification strategy is intuitive for *M0000*, where we assume that repeated responses from the same subject have the same mean regardless of time. This may not be true all the time. The limitation of this procedure is the difficulty to allow time-varying effects on longitudinal responses.

Table 5.6: Comparison of Risk Classification by Subject-Specific Means and RSC Status

Criterion	$\hat{E}(\mathbf{Y}_i\|\mathbf{Z}_i, b_i, c_i)$ vs. Figure 5.9	
δ_i	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$ (56%)
0	643	729
1	**63**	174

5.4.2 Risk Classification by Risk Probability

Another way to conduct risk classification is by estimated risk probabilities of η_i for each subject with available information and choosing a particular cut-off value. The estimated risk probabilities of η we consider are as follows.
(i) $\hat{P}(\eta_i = 1|\mathbf{Z}_i) = p(\mathbf{Z}_i;\hat{\alpha}_L)$, which is a risk probability estimation based only on subject i's covariates;
(ii) $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)$ through plugging in $(\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\theta}}_L)$ in

$$P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i]P(\eta_i = 1|\mathbf{Z}_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i]P(\eta_i = 1|\mathbf{Z}_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i]P(\eta_i = 0|\mathbf{Z}_i)}. \tag{5.18}$$

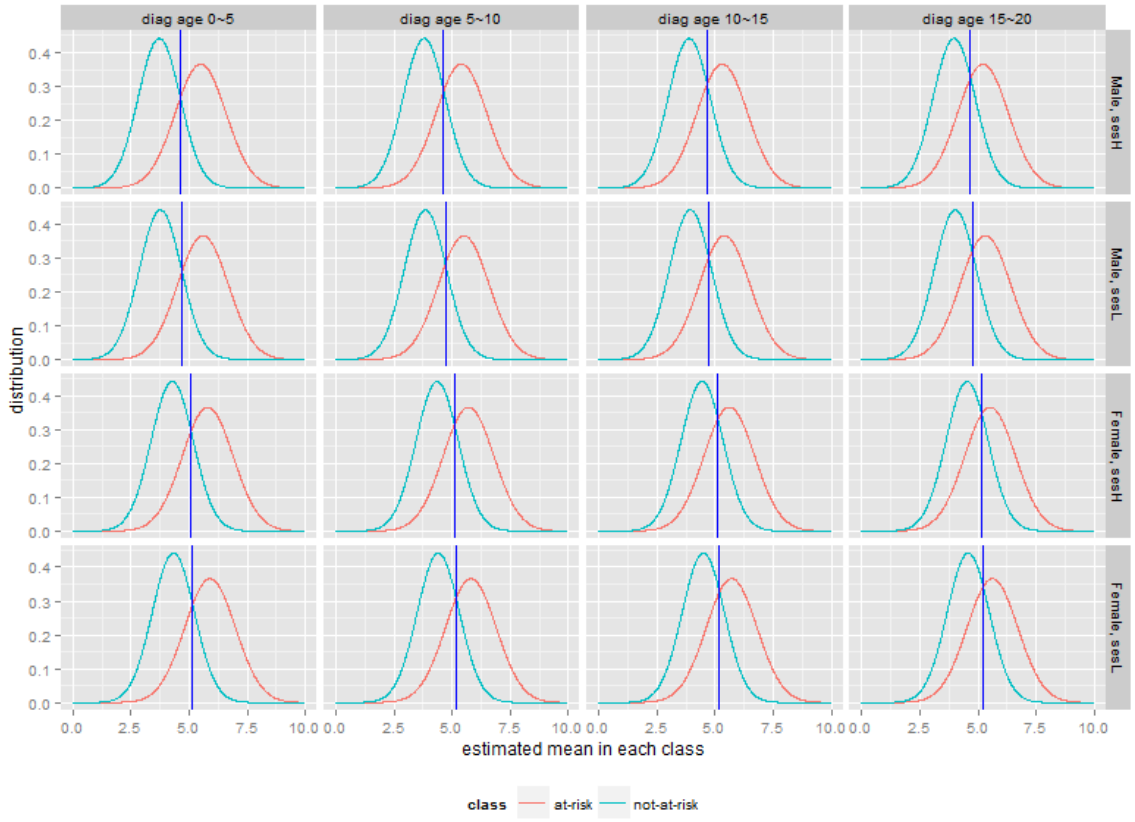

Figure 5.9: Distributions of estimated means of each class with random intercepts.

Even subjects with the same characteristics, i.e., the same $\mathbf{Z}_i$, may have different risk probabilities after we include their physician-claim data $\mathbf{Y}_i$.
(iii) $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i)$ with $(\alpha, \boldsymbol{\beta}, \boldsymbol{\theta}) = (\hat{\alpha}_L, \hat{\boldsymbol{\beta}}_L, \hat{\boldsymbol{\theta}}_L)$ and $\hat{b}_i$, $\hat{c}_i$ estimated by empirical BLUP in

$$P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Z}_i, b_i, c_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Z}_i, b_i, c_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i, c_i]P(\eta_i = 0|\mathbf{Z}_i, b_i, c_i)}. \tag{5.19}$$

Moreover, after estimating the random effects, $\hat{b}_i$ and $\hat{c}_i$, of each subject, we can add more class information to predict the probability of $\eta = 1$ given $\mathbf{Y}$, $\mathbf{Z}$, b, and c. But $P(\eta_i = 1|\mathbf{Z}_i, b_i, c_i)$ in (5.19) is not available in our LCMs. One way to get it is to jointly model the longitudinal responses $\mathbf{Y}$ and the latent indicator η through sharing the same random effects. Another way is to approximate $P(\eta_i = 1|\mathbf{Z}_i, b_i, c_i)$ by the second risk probability $P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)$ to get an approximation of $P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i)$. The approximate to $P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\tilde{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i; \alpha, \boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i, c_i]P(\eta_i = 0|\mathbf{Y}_i, \mathbf{Z}_i)}. \tag{5.20}$$

We evaluate these three estimators for risk probability under *M1101.1* for illustration. When $\mathbf{Y}_i$ is a response vector, $[\mathbf{Y}_i|\eta_i, \mathbf{Z}_i]$ is the joint pdf of $\mathbf{Y}_i$ given η_i. We assume that it follows a multivariate normal distribution with mean $\boldsymbol{\mu}_{\eta i}$ and variance $\Sigma_{\eta i}$, while given b_i and c_i, the joint pdf of $[\mathbf{Y}_i|\eta_i, \mathbf{Z}_i, b_i, c_i]$ is independent. These three estimated risk probabilities of $\eta = 1$ are calculated for each subject in the full cohort, and Figure 5.10 shows the histograms of the estimates. We see that the histogram of risk probability (i) is concentrated in between 0.45 to 0.9. By using more information in $P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)$, the risk probability (ii) distributes from 0 to 1. And the distribution of risk probability (iii) is more towards extreme values, so (iii) is less sensitive to the choice of cut-off value. To study the variability of the three risk probabilities, we estimate 9 replicates of them by parametric bootstraps, which are correspondingly placed on the right hand side of each histogram in Figure 5.10. We can see that the risk probability (i) varies a lot, while the other two are relatively stable.

The estimated proportion of subjects in the at-risk class is about 62.1% based on $\sum_{i=1}^{n} \hat{P}(\eta_i = 1|\mathbf{Z}_i)/n$. The estimated proportion based on $\sum_{i=1}^{n} \hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)/n$ is about 62.5%; while it is about 63.4% based on $\sum_{i=1}^{n} \hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i)/n$. Based on $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)$, we calculate the estimated proportion at-risk in different treatment cohorts: about 59.0% of the survivors with chemotherapy but no radiation are in the at-risk class; about 82% of those with radiation but no chemotherapy are in the at-risk class; about 73.9% of those with both chemotherapy and radiation are in the at-risk class; and about 50.2% of those with other treatments are in the at-risk class.

To conduct classification by the three estimated risk probabilities, we need to choose a cut-off value. When the estimated risk probability is larger than the cut-off value, $\hat{\eta}_i = 1$;

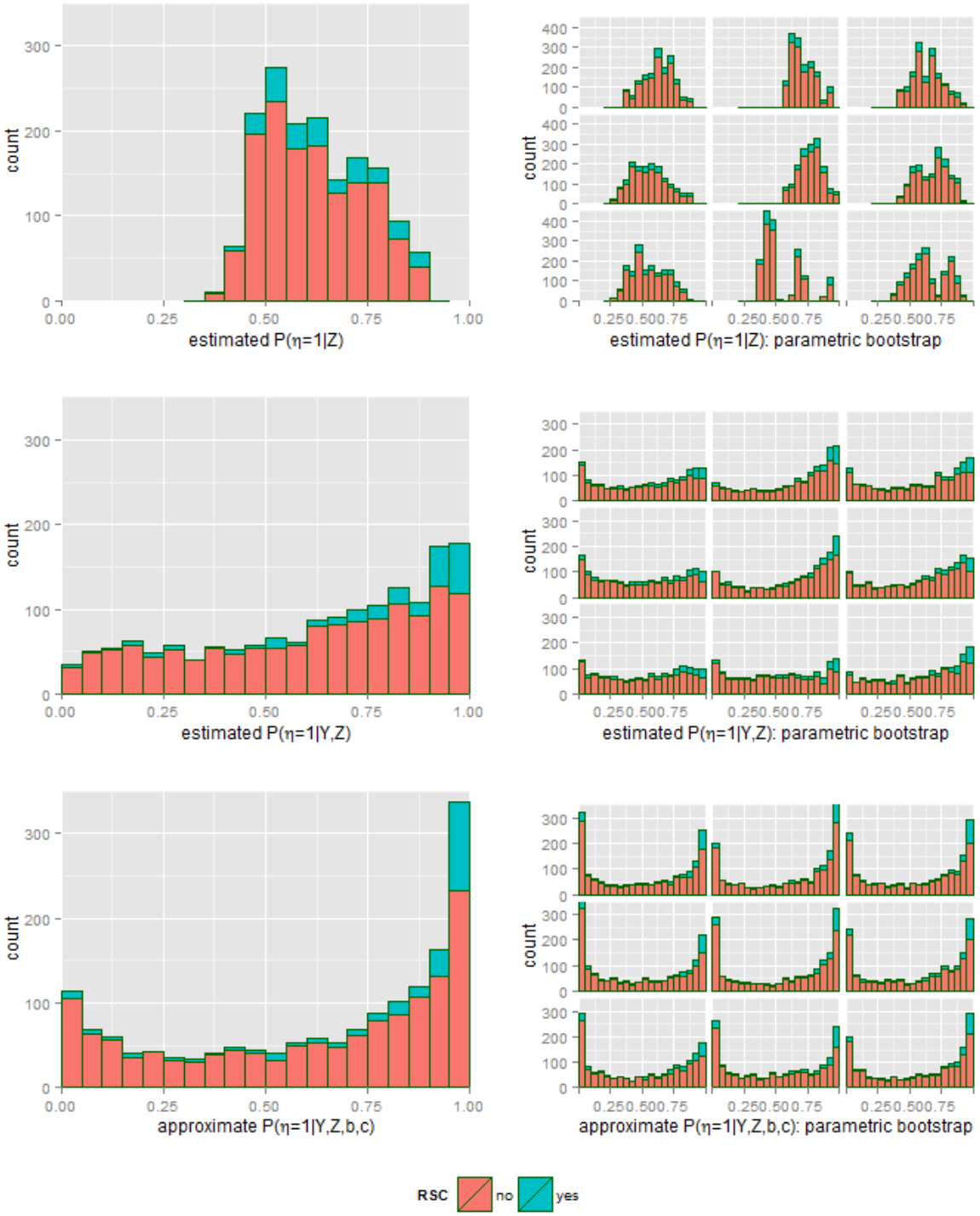

Figure 5.10: Histograms of estimated risk probabilities of η for the full survivor cohort and parametric bootstraps.

otherwise, $\hat{\eta}_i = 0$. Table 5.7 presents the risk classification results by these three estimated risk probabilities at different cut-off values, and compared with RSC status.

Table 5.7: Comparison of Risk Classification and RSC Status at Different Cut-off Values

Criterion	$\hat{P}(\eta_i = 1\|\mathbf{Z}_i)$		$\hat{P}(\eta_i = 1\|\mathbf{Y}_i, \mathbf{Z}_i)$		app $\hat{P}(\eta_i = 1\|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i)$	
δ_i	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$
cut-off = 0.1		(100%)		(95%)		(89%)
0	0	1372	79	1293	170	1202
1	**0**	237	**7**	230	**12**	225
cut-off = 0.2		(100%)		(88%)		(82%)
0	0	1372	188	1184	262	1110
1	**0**	237	**14**	223	**21**	216
cut-off = 0.3		(100%)		(81%)		(78%)
0	0	1372	284	1088	336	1036
1	**0**	237	**24**	213	**26**	211
cut-off = 0.4		(99%)		(75%)		(73%)
0	8	1364	379	993	406	966
1	**2**	235	**25**	212	**31**	206
cut-off = 0.5		(82%)		(68%)		(67%)
0	262	1110	480	892	491	881
1	**32**	205	**33**	204	**37**	200
cut-off = 0.6		(52%)		(60%)		(61%)
0	674	698	591	781	573	799
1	**102**	135	**49**	188	**49**	188
cut-off = 0.7		(30%)		(49%)		(54%)
0	983	389	754	618	674	698
1	**151**	86	**64**	173	**59**	178
cut-off = 0.8		(9%)		(36%)		(45%)
0	1260	112	928	444	815	557
1	**198**	39	**95**	142	**74**	163
cut-off = 0.9		(0%)		(22%)		(31%)
0	1372	0	1126	246	1008	364
1	**237**	0	**130**	107	**101**	136

5.4.3 Evaluation of Risk Classification by Risk Probability

Now, we want to evaluate and compare the performance of the three risk probability estimators in risk classification. Formally, in a classification problem, a ROC curve, which is sensitivity against 1-specificity, performs this task if we know the true value of η for at-risk membership. However, η is latent in our application. Given the available RSC status, par-

tial information about the latent indicator η, we propose the following approximate ROC method.

To estimate the two conditional probabilities, sensitivity and specificity, one needs the numbers of true negative (TN), false positive (FP), false negative (FN) and true positive (TP). Instead, we have only information summarized in 2 by 2 contingency tables such as Table 5.7, which are based on to RSC status (δ) instead of η. Under our additional model assumptions, the $\delta = 1$ group belongs to at-risk class ($\eta = 1$), while the $\delta = 0$ group is a combination of $\eta = 1$ and $\eta = 0$. We assume the same sensitivity for with ($\delta = 1$) or without ($\delta = 0$) RSC groups. We can use the observed sensitivity to approximate sensitivity of the classifier. Here, the observed sensitivity is the ratio of the number of $\hat{\eta} = 1$ given $\delta = 1$ divided by the number of $\delta = 1$.

Figure 5.11 plots the approximate false negative rate (1-sensitivity) and the classified at-risk rate (rate of $\hat{\eta} = 1$) of the three estimated risk probabilities at different cut-off values. The false positive rate is positive proportional to the classified at-risk rate, i.e., the higher the classified at-risk rate results in higher false positive rate. The false negative rate for $P(\eta = 1|\mathbf{Z})$ is $102/237 = 43\%$, and drops to $49/237 = 21\%$ for $P(\eta = 1|\mathbf{Y}, \mathbf{Z})$ when cut-off is 0.6.

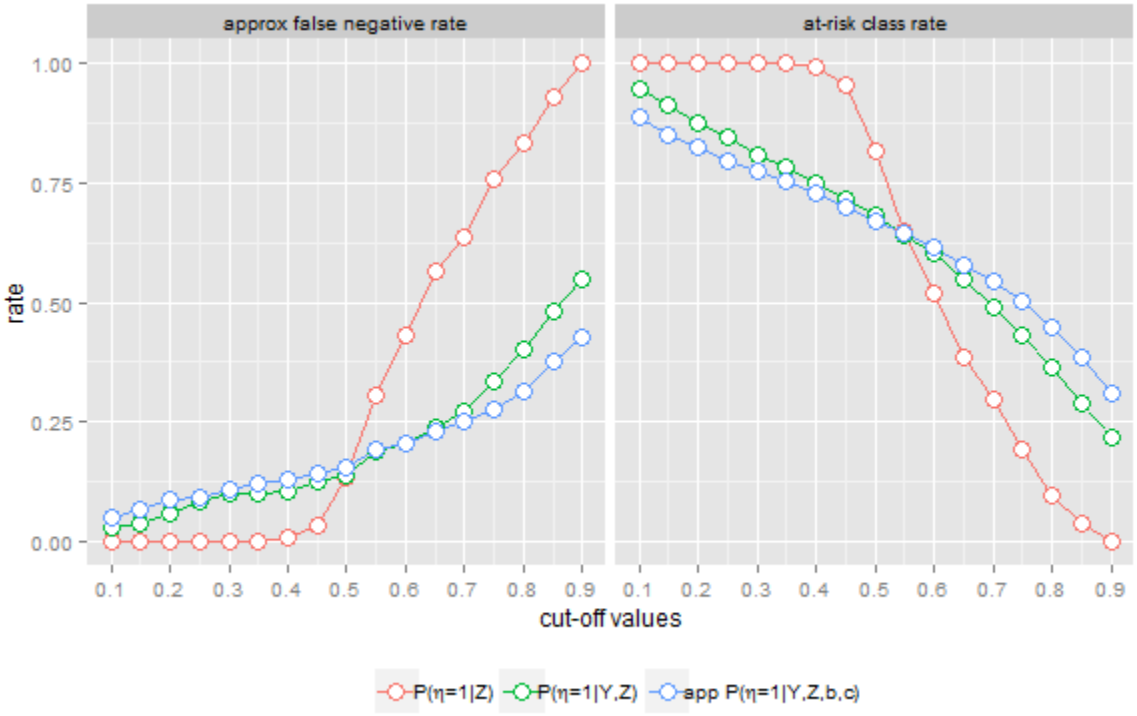

Figure 5.11: Approximate false negative rate and classified at-risk class rate at different cut-off values.

To get specificity, besides the same sensitivity assumption, we also need assume the prevalence of at-risk, which is the rate of $\eta = 1$ in survivors. At each prevalence level, we can decompose the numbers of $\hat{\eta} = 0$ and $\hat{\eta} = 1$ in the $\delta = 0$ group into real $\eta = 1$ and $\eta = 0$ classes to calculate TN and FP, then $1 - \text{specificity} = \frac{\text{FP}}{\text{TN+FP}}$. Figure 5.12 shows approximate ROC curves at different prevalence levels and contrasts with the three estimated risk probabilities.

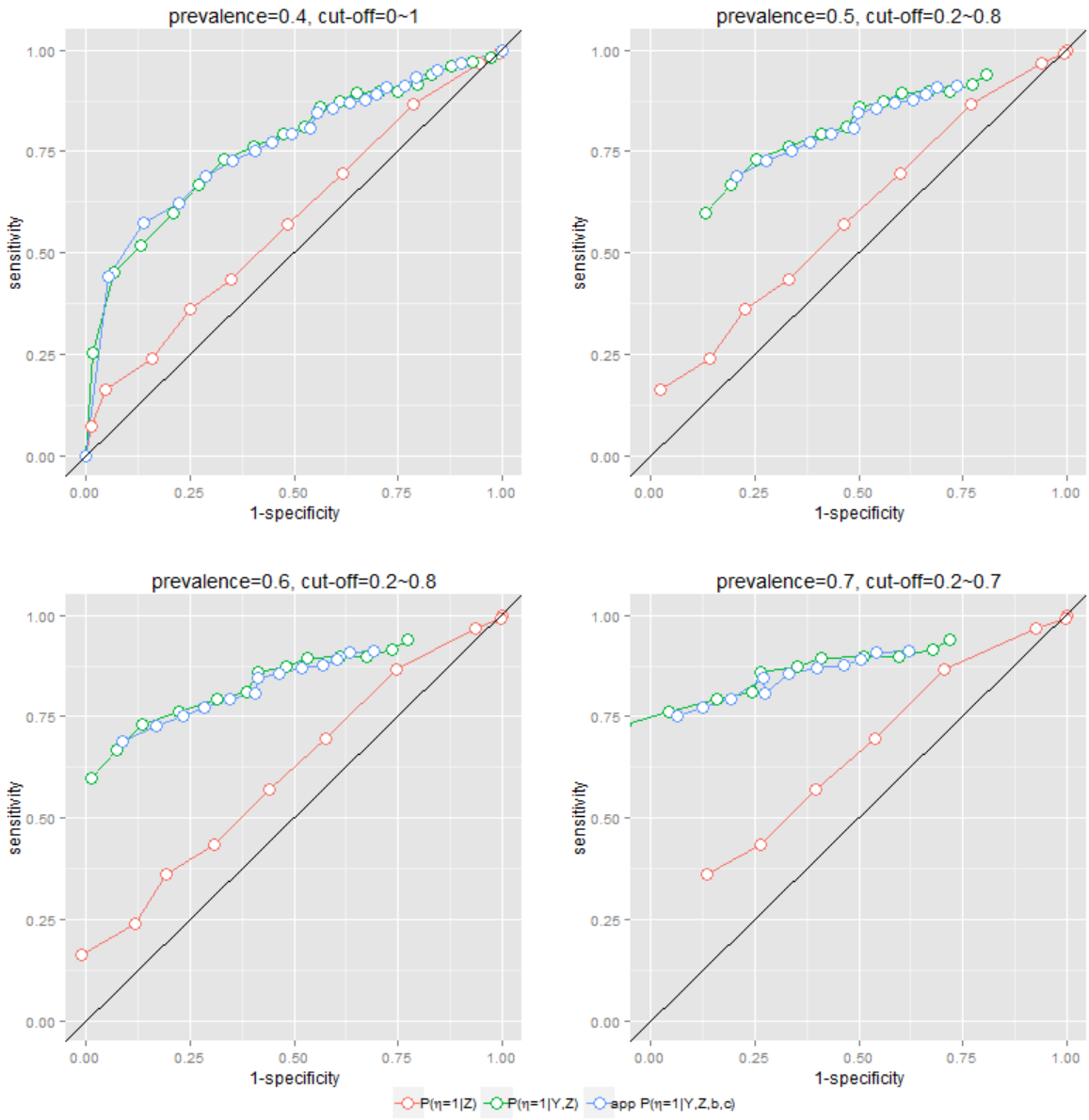

Figure 5.12: Approximate ROC at different prevalence.

From these four plots in Figure 5.12, we can see that both the risk probabilities (ii) and (iii) dramatically improve risk probability (i), $\hat{P}(\eta_i = 1|\mathbf{Z}_i)$. The performance of risk probabilities (ii) and (iii) are quite close when cut-off values are between 0.5 and 0.75, and risk probability (ii) is slightly better than (iii) in this range. In other ranges, risk probability (iii) is much better than (ii). In general, risk probability (iii) is more stable and less sensitive to cut-off value choices.

5.4.4 Dynamic Estimates for Risk Probability

We consider the risk probability $P(\eta = 1|\mathbf{Z})$ which only depends on individual's characteristics, as the risk probability at time 0. For each survivor in the cohort, we can estimate their risk probabilities dynamically at different follow-up years. For example, we can estimate the dynamic risk probabilities at time 0 by $\hat{P}(\eta = 1|\mathbf{Z})$, at time 5 years by $\hat{P}(\eta = 1|\mathbf{Z}, \mathbf{Y}_5)$ up to 5 years observations after follow-up, and at every 5 more years until use all 20 years observations. Figure 5.13 are histograms of dynamic estimated risk probabilities at time 0, up to 5 years, up to 10 years, up to 15 years, and up to 20 years, respectively, for the full survivor cohort.

We illustrate this procedure by a sub-cohort of 40 subjects, who are diagnosed with cancer in 1981 and followed until 2006. They are the people who have all 20 years follow-up. Within this sub-cohort, 9 of them have RSC and 31 of them don't have. We evaluate the following five dynamic risk probabilities: $\hat{P}(\eta = 1|\mathbf{Z}) \mapsto \hat{P}(\eta = 1|\mathbf{Z}, \mathbf{Y}_5) \mapsto \hat{P}(\eta = 1|\mathbf{Z}, \mathbf{Y}_{10}) \mapsto \hat{P}(\eta = 1|\mathbf{Z}, \mathbf{Y}_{15}) \mapsto \hat{P}(\eta = 1|\mathbf{Z}, \mathbf{Y}_{20})$ for each of them, and plot these dynamic risk probability curves for each individual in the upper panel of Figure 5.14 and contrast them separately according to their RSC status. We also classify them into at-risk (in red) and not-at-risk (in grey) classes. As we assumed, the RSC group belongs to at-risk class, while the no RSC group is a combination and need to be classified. By using cut-off value 0.5 at the 20's year, the no RSC group is categorized to two classes.

We take the means at each time point within each of the classes and plot at the bottom panel of Figure 5.14 in different RSC groups, respectively. In the two classes of the no RSC group, the at-risk class (red line) is quite similar to the red line in the RSC group. At time 0, the at-risk and not-at-risk classes in the no RSC group are quite close. As more and more data are collected, the two latent classes are separated obviously.

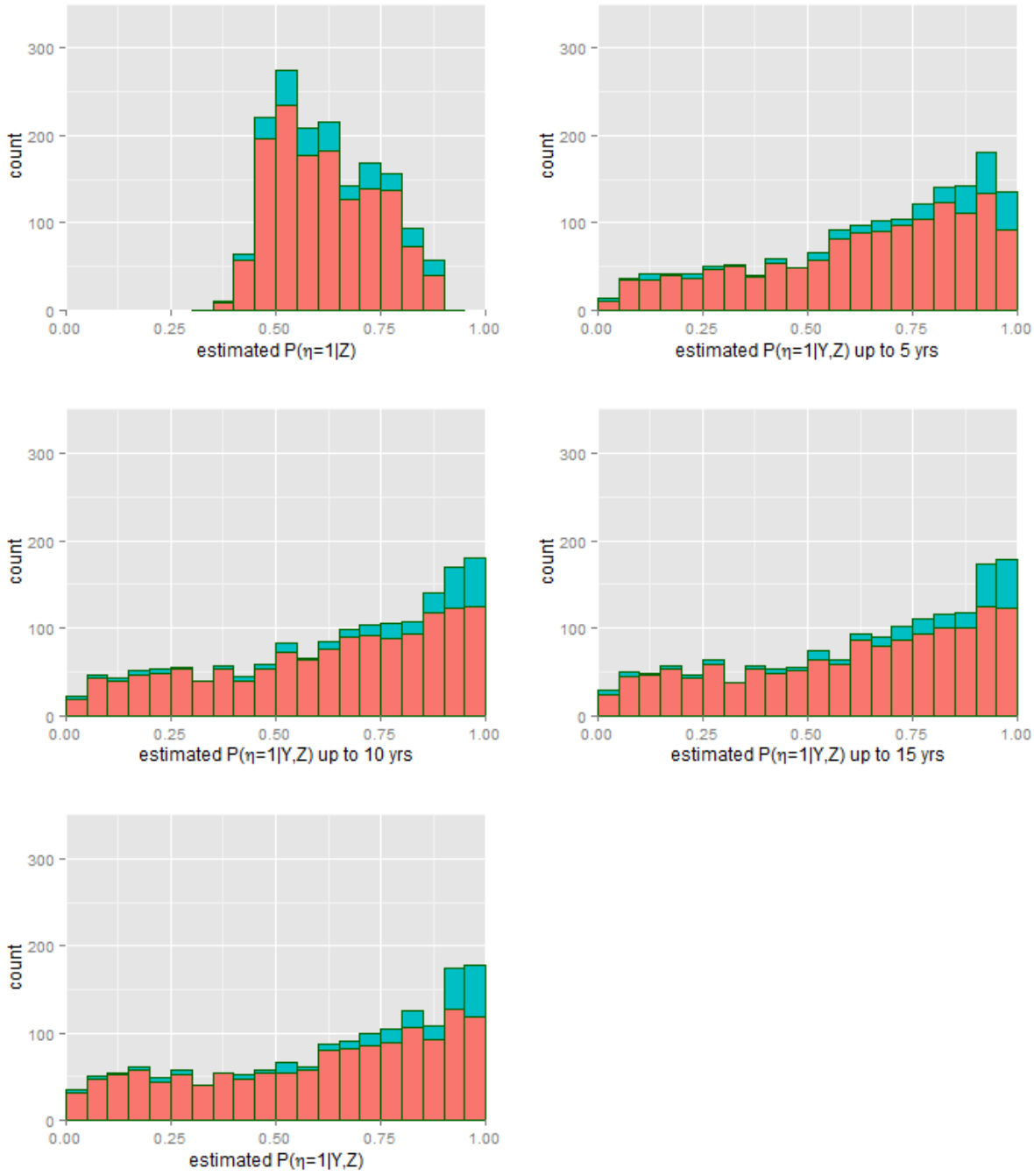

Figure 5.13: Histograms of dynamic estimates for risk probability for full survivor cohort.

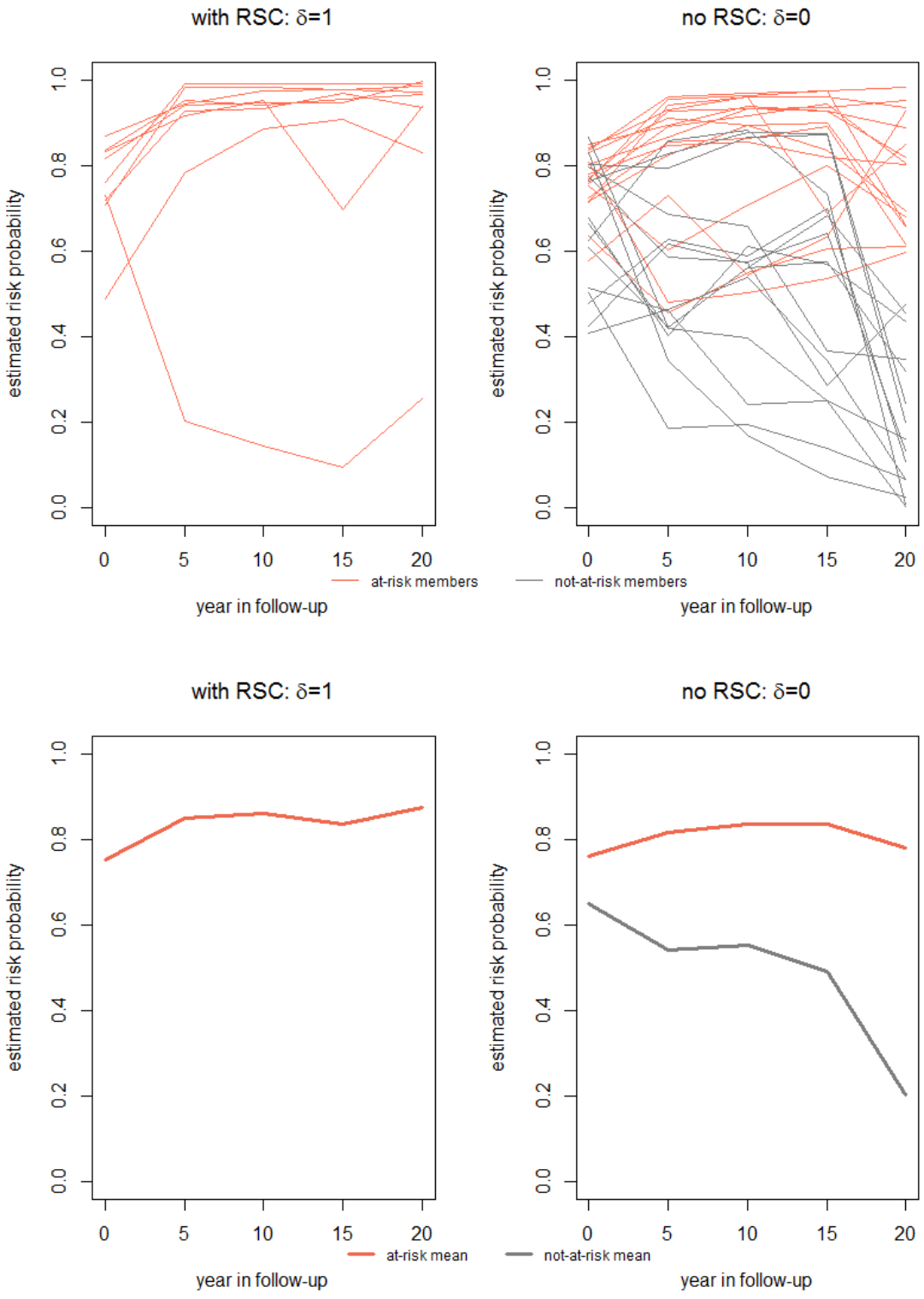

Figure 5.14: Dynamic estimates for risk probability of survivors diagnosed in 1981 and followed until 2006 (40 in total) and means of each risk class.

5.4.5 Risk Prediction for RSC during Follow-up

As in Section 3.7.2, to validate the model and estimating procedure, we predict the RSC during follow-up group using the LCM, when it is not used in model fitting. This application uses $\delta = \delta_0$ to fit the LCM. The $\delta = 1$ subset has $s = 165$ survivors with RSC before follow-up, and $v = 75$ survivors had RSC during the follow-up. As in Figures 5.1 and 5.2, Figures 5.15 and 5.16 compare yearly counts and costs in the $\delta = 1$ subset with the full cohort and the general population, respectively. The sample size of the $\delta = 1$ subset became too small to fit the model in the final few years. We illustrate this application using the yearly costs for the first 16 years of the follow-up under *M1101.1*. We estimated $\hat{P}(\eta_i = 1|\mathbf{Z}_i)$ and $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)$ after fitting the model and classified the cohort into at-risk and not-at-risk groups using these estimated risk probabilities as in Section 5.4.2. Table 5.8 compares the risk classification and prediction with the RSC status.

Table 5.8: Comparison of Risk Prediction and RSC Status

Criterion	$\hat{P}(\eta_i = 1\|\mathbf{Z}_i) > 0.6$		$\hat{P}(\eta_i = 1\|\mathbf{Y}_i, \mathbf{Z}_i) > 0.6$		either > 0.6	
Classification	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$ (54%)	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$ (60%)	$\hat{\eta}_i = 0$	$\hat{\eta}_i = 1$ (74%)
RSC status						
No	674	722	590	779	367	1002
Before follow-up	**78**	87	**37**	128	**24**	141
During follow-up	**22**	53	**20**	55	**7**	68

The false negative rates for predicting RSC during follow-up by $P(\eta = 1|\mathbf{Z})$ and $P(\eta = 1|\mathbf{Y}, \mathbf{Z})$ are $22/75 = 29\%$ and $20/75 = 27\%$, respectively. Predicting $\hat{\eta}_i = 1$ provided any of the estimated probabilities are larger than 0.6, only $7/75 = 9\%$ are falsely predicted to be in the $\eta = 0$ class.

5.5 Summary and Discussion

We have extended the extended GEE estimator developed in Chapter 3 for cross-sectional counts to longitudinal LCMs. We have shown that the procedure can readily be used for any type of longitudinal data provided it is suitable for GEE approaches. We have applied the methodology to the CAYACS data in yearly binaries, counts, and costs. We have also conducted risk classification and prediction based on the longitudinal LCMs. In addition to marginal models and the extended GEE procedure for the longitudinal LCMs, we considered mixed effects models for each latent class and developed risk classification based on the prediction of random effects for the LCMs.

In further investigation we could consider continuous time scales rather than yearly data, viewing the longitudinal data as stochastic processes.

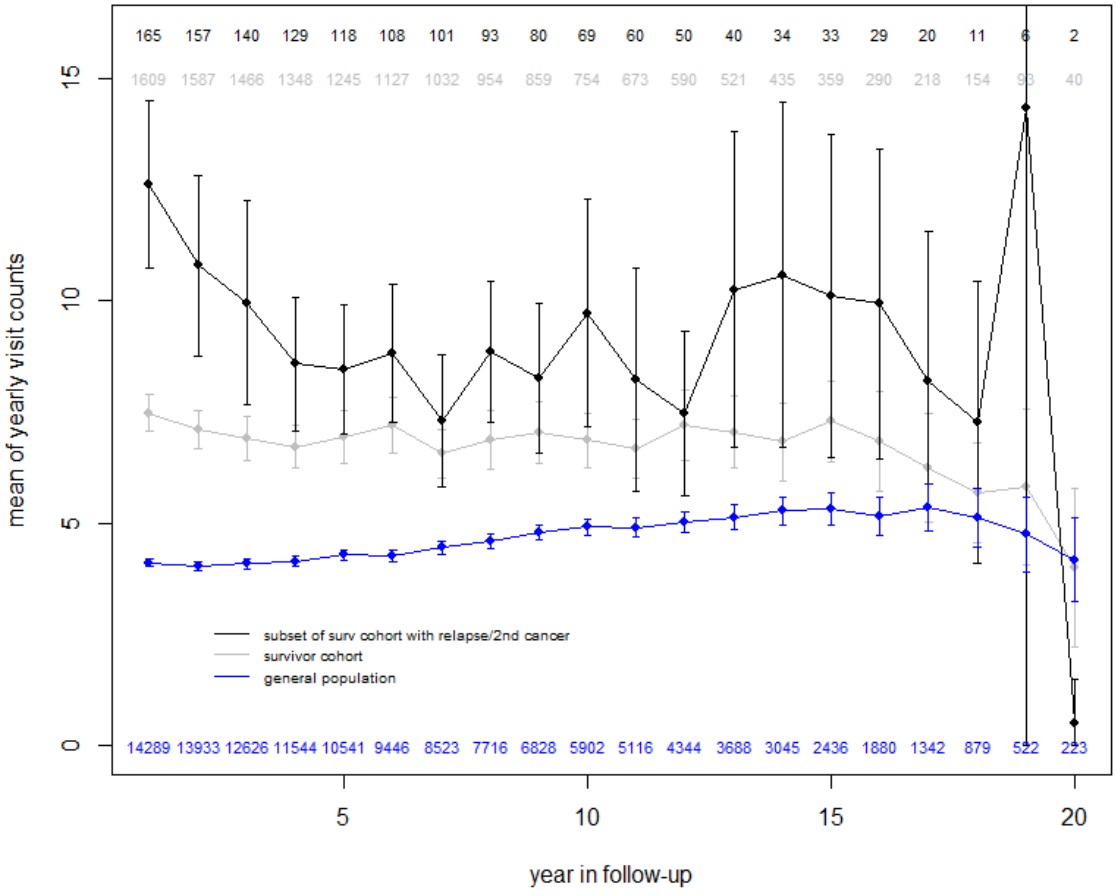

Figure 5.15: Mean and CI of yearly counts during follow-up: Subset of survivor cohort with RSC (before follow-up) vs. full survivor cohort vs. general population.

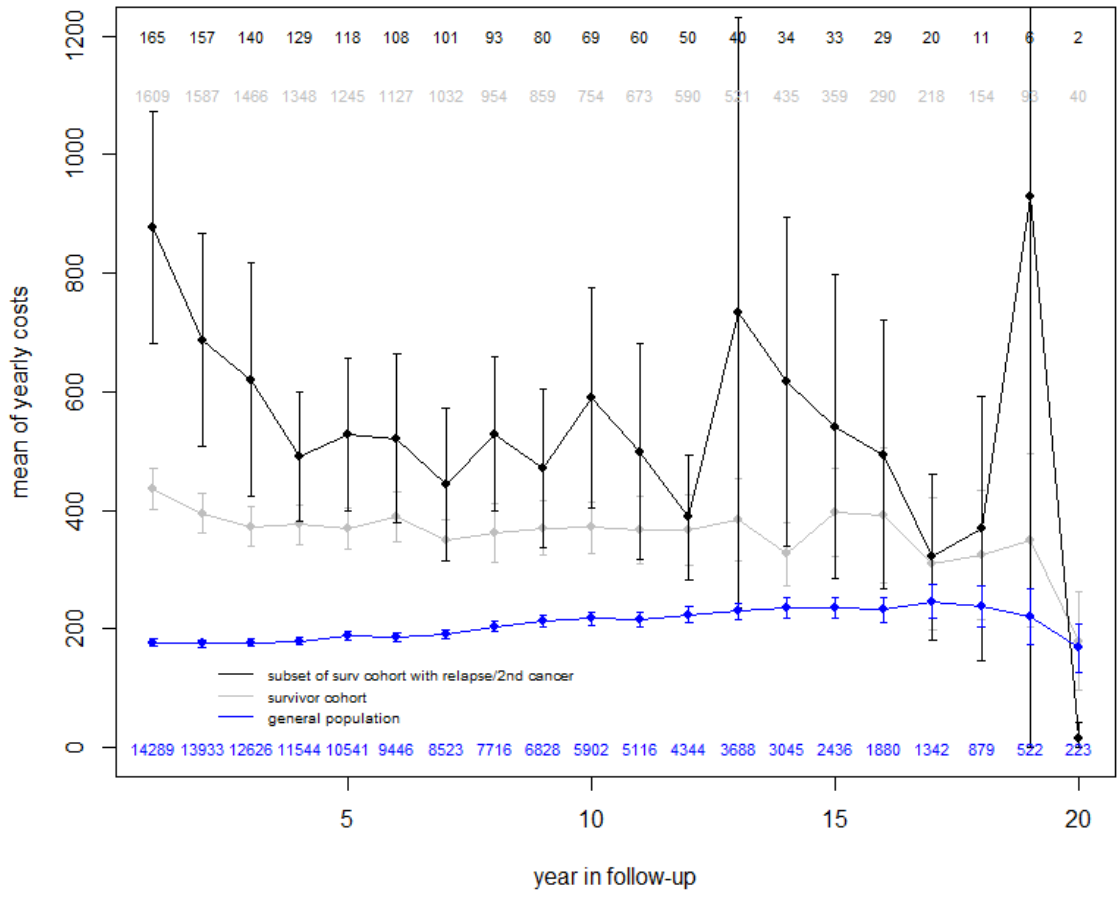

Figure 5.16: Mean and CI of yearly costs during follow-up: Subset of survivor cohort with RSC (before follow-up) vs. full survivor cohort vs. general population.

Chapter 6

Final Discussion

6.1 Summary

This thesis was motivated and illustrated by the CAYACS physician claim project, but the proposed modelling and inference procedures can be applied rather broadly to data with multiple unobserved classes. We formulated LCMs and developed associated inference procedures and applications. We also established the asymptotic properties of the estimators and conducted simulation studies to study their finite-sample performance. We analyzed the CAYACS data using all the methodologies as an illustration.

The CAYACS project primarily studies the risk of later effects arising from the original diagnosis. We formulated the risk assessment problem into LCMs. When survivors experience later effects, they need more physician care and thus have higher medical costs than those in the general population of the same gender and age. We defined the two subgroups in the survivor cohort as at-risk and not-at-risk classes. The at-risk membership indicator η is latent.

We first specified the LCM as a mixture Poisson distribution for cross-sectional counts. We presented the MLE and proposed a pseudo-MLE procedure that used the supplementary data to estimate the parameters in the $\eta = 0$ class (Chapter 2). Simulation studies showed that the pseudo-MLE is more efficient than the MLE and robust to $\eta = 0$ class misspecification. However, both methods lack robustness to distribution misspecification of the $\eta = 1$ class, especially when estimating the parameters in the risk model. In the analysis of CAYACS visits, we suggested that the starting time of the follow-up for the general population should be chosen to match that of a random survivor with the same sex and birth year. This makes the general population a better comparison group in terms of age at entry and length of observation period. This made a noticeable difference in the quasi-Poisson regression, which showed that the visit counts were affected nonlinearly by the aging of the subject. It is therefore necessary to study the CAYACS data longitudinally. The similarity in the real-data analysis by the MLE and pseudo-MLE in which the general

population ($\mathcal{Q}$) was used, validated the pseudo-MLE procedure. The approach to choosing the starting time for the comparison group may prove useful for practitioners, especially in longitudinal studies.

Chapter 3 developed robust estimating procedures for the LCM. We introduced a binary variable δ as partial information about the latent risk indicator η, where $\delta = 1$ was a subgroup of $\eta = 1$. We proposed three pseudo-MLEs and compared them to the MLE under a mixture Poisson distribution for the counts. These likelihood-based methods validated the use of δ as a partially observed η, but they suffered when the distribution of the counts was misspecified. To obtain more robust statistical methods, we proposed two sets of extended GEEs for the parameters in the LCM. We developed two types of extended GEE estimators for each set of extended GEEs by using the supplementary dataset for the estimation of one class alone or together with the partial information about the other class. The estimators from the extended GEEs where we specified the mean and variance functions but not the underlying distribution for each of the latent classes were robust to distribution misspecification and maintained satisfactory efficiency. The computational advantage of the extended GEE methods was clear, since the estimating equations are much simpler than likelihood functions. We adapted the extended GEEs (Chapter 5) to longitudinal counts and medical costs.

In Chapters 4 and 5 we summarized the CAYACS data in yearly counts and medical costs. Longitudinal analysis of the LCMs can explore questions for which cross-sectional analysis is not suitable. In Chapter 4, we analyzed the full survivor cohort and compared it with the general population separately and together, using conventional longitudinal approaches. We demonstrated the merits of longitudinal analysis. For example, the time-varying effects of age at entry showed that the cohort and the population have different visit trends as they age.

Chapter 5 adapted the robust extended GEE method developed for LCMs (Chapter 3) to the longitudinal data analyzed in Chapter 4 by adjusting for within-subject correlation for both yearly counts and medical costs. The analysis of the longitudinal data under LCMs achieved our objective of risk assessment; we explored the visit trends over time for each latent class. We investigated different ways to classify the cohort due to the risk of later effects by specifying the regression model in each class as either marginal or subject-specific models.

6.2 Future Investigation

In this thesis we have built pseudo-MLEs and extended GEE procedures for LCMs and comprehensively analyzed the CAYACS data using conventional approaches and our LCMs. There are several possibilities for future investigation.

When we extended the GEE-based method from cross-sectional to longitudinal data, we introduced within-subject correlation, with the subjects being independent of one another. It would be interesting to extend the proposed modelling and inferential procedures to investigate the potential correlation of the subjects, similarly to, for example, the approach of Lee *et al.* (2006).

We used physician-claim records to determine the medical care related to the later effects experienced by survivors. The records were collected from the BC MSP and did not include oncologist visits. The BC Cancer Agency does not yet have electronic oncology records available, and it has a different payment system. In addition, oncologists do not provide regular care for survivors after the age of 18. A combination of all the relevant medical information including BC MSP physician visits, oncologist visits, and periods of hospitalization would lead to a more reliable risk assessment.

In our research, the latent risk indicator η is constant over time. A time-dependent indicator $\eta(t)$ could accommodate evolving cohorts. Another way to extend the LCMs is to consider more than two classes. The determination of number of classes can be viewed as a model selection problem, or can conduct cluster analysis to pre-determine the number of classes.

Another issue is noninformative censoring. For example, the death rates during the follow-up period are about 4%, 17%, and 0.4% for the cohort, the RSC subgroup, and the population, respectively. This may not be an issue for cross-sectional analysis, since the cross-sectional data summarized the information for the entire follow-up period. Moreover, the death rates are quite low in the full cohort and the population. However, for the RSC subgroup we saw sudden drops in the visit counts and medical costs in the final few years. This may have occurred because survivors with serious later effects had died. Future research may need to adopt informative censoring, especially for the RSC subgroup.

One advantage of longitudinal analysis over cross-sectional analysis is the ability to deal with time-varying covariates. The SES of young survivors may change more over time than that of their peers, which could be an important factor. We used the SES at the study entry as a constant covariate, but it could be updated using census data and recorded address changes. Another recommendation is that one should choose the time scale for the longitudinal analysis according to the study objectives. As mentioned before, a calendar time scale could be used for administrative purposes. For an individual time scale, the starting point could be the individual's birthday, allowing the age and cohort effects to be totally separated. The response $\mathbb{Y} = Y(t)$ could instead be viewed as a mixture of stochastic processes using real time scales. It would be challenging to develop LCMs and the corresponding inference procedures for stochastic processes.

The physician visit counts and costs are highly correlated. We can construct the counting process of visits ($N(t)$) and costs corresponding to each visit (c_j) together as $C(t) = \sum_{j=1}^{N(t)} c_j$. This compound Poisson process may serve as a better model of the CAY-

ACS data and answer more scientific questions. For example, costs corresponding to each visit may measure the severity of the visit.

6.2.1 Generalized Methods of Moments

We did some work on the GMM approach and it showed promise. In Chapter 3, we proposed two sets of extended GEEs for the same set of parameters in the LCM. Therefore, the number of estimating equations is greater than the number of parameters. This scenario falls into the GMM framework, so we could combine the estimating equations using a GMM estimator. In simulation setting 3, the GMM estimators performed well even when the assumption that the at-risk class has the same distribution of visit counts as the $\delta = 1$ subgroup was not valid and the extended GEE estimators were slightly biased. The GMM estimator for LCMs deserves further theoretical and numerical study.

GMM was developed in 1982 as a generalization of the method of moments in econometrics (Hansen, 1982). It has since been widely applied to analyze economic and financial data, and numerous inference techniques based on GMM estimators have been developed (Hall, 2005). For recent developments, see Yin (2009) and Yin *et al.* (2011). As is well known, the optimality of the MLE stems from its distributional assumption for the data. The statistical properties of the MLE are quite sensitive to this assumption, and the MLE is often computationally burdensome. In contrast, the GMM framework provides a computationally convenient method that avoids the need to specify the likelihood function (Hall, 2005).

GMM has had a significant impact in econometrics. Hall (2005) declared "a set of population moment conditions which are deduced from the assumptions of the econometric model" to be the cornerstone of GMM estimation. We constructed a GMM framework based on statistical estimating functions and developed a GMM estimator robust to distribution misspecification, as well as the model assumptions in Section 3.2, since some estimating equations without those model assumptions play an important role in GMM estimators by weighting. As stated in Section 3.4.1, the type J GEE estimator did not use the model assumptions of Section 3.2; it ignored the information about δ. As a result, it suffered nonidentifiability problems in the LCM, i.e., it failed to distinguish classes. However, it may be a good candidate if we wish to relax Assumption (ii). In contrast, the type P extended GEE estimators applied Assumption (iii) explicitly. This simplified the implementation, avoided nonidentifiability problems, and was totally robust to distribution misspecification. However, they were not robust enough for the model assumptions. Every GEE estimator is double-edged. If we can use the GMM framework, combining the estimating equations properly to balance the different assumptions, we will be able to develop GMM estimators for the LCM that avoid nonidentifiability problems and are also robust to distribution misspecification and the model assumptions. Simulation studies have demonstrated the

advantages of GMM, at least for the estimates of α in the risk model. Further theoretical derivation of GMM for the LCM is needed.

Another advantage of the GMM framework is that the overidentifying restrictions can be used in diagnostic tests of misspecification of the population moment conditions. We could apply these theories to develop tests for the assumption that $\delta = 1$ is a subgroup of the $\eta = 1$ class.

Bibliography

Boxall, P. C. and Adamowicz, W. L. (2002). Understanding heterogeneous preferences in random utility models: A latent class approach. *Environmental and Resource Economics*, **23**(4), 421–446.

De Angelis, L. (2013). Latent class models for financial data analysis: Some statistical developments. *Statistical Methods & Applications*, **22**(2), 227–242.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, **39**(1), 1–38.

Desantis, S. M., Andrés Houseman, E., Coull, B. A., Nutt, C. L., and Betensky, R. A. (2012). Supervised Bayesian latent class models for high-dimensional data. *Statistics in Medicine*, **31**(13), 1342–1360.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Ekholm, A., Jokinen, J., McDonald, J. W., and Smith, P. W. (2012). A latent class model for bivariate binary responses from twins. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**(3), 493–514.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied Longitudinal Analysis*, volume 998. John Wiley & Sons.

Formann, A. K. and Kohlmann, T. (1998). Structural latent class models. *Sociological Methods & Research*, **26**(4), 530–565.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics*, **56**(4), 1055–1067.

Ghosh, J., Herring, A. H., and Siega-Riz, A. M. (2011). Bayesian variable selection for latent class models. *Biometrics*, **67**(3), 917–925.

Godambe, V. P. (1991). *Estimating Functions*. Clarendon Press Oxford.

Gong, G. and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, **9**(4), 861–869.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**(2), 215–231.

Grisolía, J. M. and Willis, K. G. (2012). A latent class model of theatre demand. *Journal of Cultural Economics*, **36**(2), 113–139.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press, Oxford.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, **56**(4), 1030–1039.

Hall, D. B. and Shen, J. (2010). Robust estimation for zero-inflated Poisson regression. *Scandinavian Journal of Statistics*, **37**(2), 237–252.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**(4), 1029–1054.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233.

Jobson, J. (2012). *Applied multivariate data analysis: volume II: Categorical and Multivariate Methods*. Springer Science & Business Media.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton, Mifflin.

Lee, A. H., Wang, K., Scott, J. A., Yau, K. K., and McLachlan, G. J. (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*, **15**(1), 47–61.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, **19**(2), 219–238.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. H., and Clark, L. C. (2000). A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. *Statistics in Medicine*, **19**(10), 1303–1318.

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, **86**(413), 96–107.

Magidson, J. and Vermunt, J. (2002). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing Research*, **20**(1), 36–43.

McBride, M., Rogers, P., Sheps, S., Glickman, V., Broemeling, A., Goddard, K., Hu, J., Lorenzi, M., Peacock, S., Pritchard, S., *et al.* (2010). Childhood, adolescent, and young adult cancer survivors research program of British Columbia: Objectives, study design, and cohort characteristics. *Pediatric Blood & Cancer*, **55**(2), 324–330.

McBride, M., Lorenzi, M., Page, J., Broemeling, A., Spinelli, J., Goddard, K., Pritchard, S., Rogers, P., and Sheps, S. (2011). Patterns of physician follow-up among young cancer survivors. *Canadian Family Physician*, **57**(12), e482–e490.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, **11**(1), 59–67.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC press.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**(437), 162–170.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. Wiley.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, **29**(1), 81–117.

of Health, B. C. M. (2013). Medical services plan (MSP) payment information file. Data extract published by Population Data BC.

Pepe, M. S. and Janes, H. (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics*, **8**(2), 474–484.

Reboussin, B. A., Liang, K.-Y., and Reboussin, D. M. (1999). Estimating equations for a latent transition model with multiple discrete indicators. *Biometrics*, **55**(3), 839–845.

Reid, N. (2010). Likelihood inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**(5), 517–525.

Schlattmann, P. (2009). *Medical Applications of Finite Mixture Models*. Springer.

Shih, Y.-C. T. and Tai-Seale, M. (2012). Physicians' perception of demand-induced supply in the information age: A latent class model analysis. *Health Economics*, **21**(3), 252–269.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, **88**(422), 421–427.

van Smeden, M., Naaktgeboren, C. A., Reitsma, J. B., Moons, K. G., and de Groot, J. A. (2014). Latent class models in diagnostic studies when there is no reference standard – A systematic review. *American Journal of Epidemiology*, **179**(4), 423–431.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**(1), 5–42.

Varki, S. and Chintagunta, P. K. (2004). The augmented latent class model: Incorporating additional heterogeneity in the latent class model for panel data. *Journal of Marketing Research*, **41**(2), 226–233.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, **17**(1), 33–51.

Vermunt, J. K. and Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis*, **41**(3), 531–537.

Wang, H., Hu, X. J., McBride, M. L., and Spinelli, J. J. (2014). Analysis of counts with two latent classes, with application to risk assessment based on physician-visit records of cancer survivors. *Biostatistics*, **15**(2), 384–397.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, **39**(2), 95–101.

Ware, J. H., Lipsitz, S., and Speizer, F. E. (1988). Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine*, **7**(1–2), 95–107.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**(3), 439–447.

Yang, H., O'Brien, S., and Dunson, D. B. (2011). Nonparametric Bayes stochastically ordered latent class models. *Journal of the American Statistical Association*, **106**(495), 807–817.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis*, **4**(2), 191–207.

Yin, G., Ma, Y., Liang, F., and Yuan, Y. (2011). Stochastic generalized method of moments. *Journal of Computational and Graphical Statistics*, **20**(3), 714–727.

Yuan, K.-H. and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis*, **65**(2), 245–260.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**(1), 121–130.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**(4), 1049–1060.

British Columbia Ministry of Health [creator] (2013): Medical Services Plan (MSP) Payment Information File. Population Data BC [publisher]. Data Extract. MOH (2012). http://www.popdata.bc.ca/data

British Columbia Ministry of Health [creator] (2013): Medical Services Plan (MSP) Consolidation File. Population Data BC [publisher]. Data Extract. MOH (2012). http://www.popdata.bc.ca/data

BC Vital Statistics Agency [creator] (2012): Vital Statistics Deaths. Population Data BC [publisher]. Data Extract BC Vital Statistics Agency (2012). http://www.popdata.bc.ca/data

Appendix A

Application of EM Algorithm for the Likelihood-based Estimations in Chapter 3

An alternative way to get the MLE of ϕ described in Section 3.2.1 is to apply the EM-algorithm via a full-data likelihood function. Let the "full data" be $\mathcal{F} = \{(Y_i, \eta_i, \delta_i, T_i, \mathbf{Z}_i) : i = 1, \ldots, n\}$ in this application. We can verify the conditions that ensure that the resulting sequence of estimates converges to the MLE $\hat{\phi}$ from $L(\phi; \mathcal{P})$. The likelihood function of ϕ based on $\mathcal{F}$ is

$$\begin{aligned}
L(\phi; \mathcal{F}) &= \prod_{i=1}^{n} [Y_i, \eta_i, \delta_i | T_i, \mathbf{Z}_i; \phi] \\
&= \prod_{i=1}^{n} [Y_i | \eta_i, \delta_i, T_i, \mathbf{Z}_i][\eta_i, \delta_i | \mathbf{Z}_i] \\
&= \prod_{i=1}^{n} [Y_i | \eta_i, T_i, \mathbf{Z}_i][\eta_i, \delta_i | \mathbf{Z}_i] \\
&= \prod_{i=1}^{n} [Y_i | \eta_i = 1, T_i, \mathbf{Z}_i; \beta]^{\eta_i} \prod_{i=1}^{n} [Y_i | \eta_i = 0, T_i, \mathbf{Z}_i; \theta]^{1-\eta_i} \\
&\times \prod_{i=1}^{n} q(\mathbf{Z}_i; \rho)^{\delta_i \eta_i} [p(\mathbf{Z}_i; \alpha) - q(\mathbf{Z}_i; \rho)]^{(1-\delta_i)\eta_i} [1 - p(\mathbf{Z}_i; \alpha)]^{(1-\delta_i)(1-\eta_i)}.
\end{aligned}$$

Then the $\mathcal{F}$ log-likelihood function is

$$l_{\mathcal{F}}(\phi) = l_{\mathcal{F}1}(\beta) + l_{\mathcal{F}2}(\theta) + l_{\mathcal{F}3}(\rho, \alpha),$$

where

$$l_{\mathcal{F}1}(\beta) = \sum_{i=1}^{n} \eta_i \log[Y_i | \eta_i = 1, T_i, \mathbf{Z}_i; \beta], \tag{A.1}$$

$$l_{\mathcal{F}2}(\theta) = \sum_{i=1}^{n} (1-\eta_i) \log[Y_i|\eta_i = 0, T_i, \mathbf{Z}_i; \theta], \tag{A.2}$$

$$l_{\mathcal{F}3}(\rho, \alpha) = \sum_{i=1}^{n} \delta_i \log q(\mathbf{Z}_i; \rho) + (\eta_i - \delta_i) \log[p(\mathbf{Z}_i; \alpha) - q(\mathbf{Z}_i; \rho)] + (1-\eta_i) \log[1 - p(\mathbf{Z}_i; \alpha)]. \tag{A.3}$$

The EM algorithm for the MLE of ϕ iterates between an E-step and an M-step that maximizes the $\mathcal{F}$ likelihood function. The computational advantage is obvious, since the full-data log-likelihood is the summation of three terms, each of which depends on only β, θ, and (ρ, α), separately. Starting with the initial values $(\rho^{(0)}, \alpha^{(0)}, \beta^{(0)}, \theta^{(0)})$, at the lth iteration of the algorithm with $l \geq 1$ and the $(l-1)^{th}$ estimates $(\rho^{(l-1)}, \alpha^{(l-1)}, \beta^{(l-1)}, \theta^{(l-1)})$, the algorithm updates the estimates as follows:

E-Step For $i = 1, \ldots, n$, $\eta_i^{(l)} = 1$ if $\delta_i = 1$, otherwise calculate $\eta_i^{(l)} = E\{\eta_i | Y_i, \delta_i = 0, T_i, \mathbf{Z}_i; \phi^{(l-1)}\}$ via

$$E(\eta|Y, \delta = 0, T, \mathbf{Z}; \phi) = \frac{[Y|\eta = 1, T, \mathbf{Z}; \beta] \frac{p(\mathbf{Z};\alpha) - q(\mathbf{Z};\rho)}{1 - q(\mathbf{Z};\rho)}}{[Y|\eta = 1, T, \mathbf{Z}; \beta] \frac{p(\mathbf{Z};\alpha) - q(\mathbf{Z};\rho)}{1 - q(\mathbf{Z};\rho)} + [Y|\eta = 0, T, \mathbf{Z}; \theta] \frac{1 - p(\mathbf{Z};\alpha)}{1 - q(\mathbf{Z};\rho)}}. \tag{A.4}$$

M-Step Obtain $\beta^{(l)}$, $\theta^{(l)}$, $(\rho^{(l)}, \alpha^{(l)})$ by separately maximizing $l_{\mathcal{F}1}(\beta; \boldsymbol{\eta}^{(l)})$, $l_{\mathcal{F}2}(\theta; \boldsymbol{\eta}^{(l)})$, and $l_{\mathcal{F}3}(\rho, \alpha; \boldsymbol{\eta}^{(l)})$ with respect to β, θ, and (ρ, α), respectively.

The EM algorithm can be applied to the three pseudo-MLEs by using the above procedure with ρ and/or θ fixed to $\hat{\rho}_A$ and $\tilde{\theta}$.

Appendix B

Asymptotic Derivations of the Pseudo-MLEs in Chapter 3

Type A Pseudo-MLE This estimating procedure is equivalent to solving $\frac{\partial l_2(\rho)}{\partial \rho} = 0$ and $\frac{\partial l_1(\phi)}{\partial(\alpha,\beta,\theta)} = 0$, simultaneously, to get zeros at $\hat{\phi}_A = (\hat{\rho}_A, \hat{\alpha}_A, \hat{\beta}_A, \hat{\theta}_A)$. The first-order Taylor expansion of the functions $\frac{\partial l_2(\rho)}{\partial \rho}|_{\rho=\hat{\rho}_A}$ and $\frac{\partial l_1(\phi)}{\partial(\alpha,\beta,\theta)}|_{\phi=\hat{\phi}_A}$, whose values are zeros, yields the following equations:

$$\begin{pmatrix} \frac{1}{\sqrt{n}}\frac{\partial l_2(\rho)}{\partial \rho} \\ \frac{1}{\sqrt{n}}\frac{\partial l_1(\phi)}{\partial(\alpha,\beta,\theta)} \end{pmatrix} \doteq -\frac{1}{n}\begin{pmatrix} \frac{\partial^2 l_2(\rho)}{\partial \rho^2} & 0 \\ \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta,\theta)\partial\rho} & \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta,\theta)^2} \end{pmatrix} \sqrt{n}\Big((\hat{\rho}_A, \hat{\alpha}_A, \hat{\beta}_A, \hat{\theta}_A)' - (\rho, \alpha, \beta, \theta)'\Big). \tag{B.1}$$

The left-hand side of (B.1), by the central limit theorem (CLT), converges to a multivariate normal distribution with mean zero and variance Σ_A, as $n \to \infty$:

$$\begin{pmatrix} \frac{1}{\sqrt{n}}\frac{\partial l_2(\rho)}{\partial \rho} \\ \frac{1}{\sqrt{n}}\frac{\partial l_1(\phi)}{\partial(\alpha,\beta,\theta)} \end{pmatrix} \xrightarrow[n\to\infty]{d} N\Big(0, \Sigma_A\Big).$$

The coefficient of the right-hand side of (B.1) converges to a constant matrix Π_A almost surely, by the strong law of large numbers (SLLN):

$$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2 l_2(\rho)}{\partial \rho^2} & 0 \\ \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta,\theta)\partial\rho} & \frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta,\theta)^2} \end{pmatrix} \xrightarrow[n\to\infty]{a.s.} \Pi_A.$$

Therefore, by Slutsky's Theorem, the asymptotic distribution for the type A pseudo-MLE $(\hat{\rho}_A, \hat{\alpha}_A, \hat{\beta}_A, \hat{\theta}_A)$ is

$$\sqrt{n}\Big((\hat{\rho}_A, \hat{\alpha}_A, \hat{\beta}_A, \hat{\theta}_A)' - (\rho, \alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Pi_A^{-1}\Sigma_A(\Pi_A^{-1})'\Big).$$

Type B Pseudo-MLE This estimating procedure is equivalent to solving the equations $\frac{\partial l(\phi)}{\partial(\rho,\alpha,\beta)} = 0$ and $m(\tilde{\theta} - \theta) = 0$, simultaneously, to get zeros at $\hat{\phi}_B = (\hat{\rho}_B, \hat{\alpha}_B, \hat{\beta}_B, \tilde{\theta})$. The

first-order Taylor expansion of the function $\frac{\partial l(\phi)}{\partial(\rho,\alpha,\beta)}|_{\phi=\hat{\phi}_B}$, whose value is zero, yields the following equation:

$$\frac{1}{\sqrt{n}}\frac{\partial l(\phi)}{\partial(\rho,\alpha,\beta)} \doteq -\frac{1}{n}\frac{\partial^2 l(\phi)}{\partial(\rho,\alpha,\beta)^2}\sqrt{n}\Big((\hat{\rho}_B,\hat{\alpha}_B,\hat{\beta}_B)'-(\rho,\alpha,\beta)\Big)-\frac{1}{n}\frac{\partial^2 l(\phi)}{\partial(\rho,\alpha,\beta)\partial\theta}\sqrt{n}(\tilde{\theta}-\theta). \tag{B.2}$$

By the properties of the likelihood estimator, the left-hand side of (B.2) converges to a multivariate normal distribution with mean zero and variance Σ_B^*, as $n\to\infty$, $\frac{1}{\sqrt{n}}\frac{\partial l(\phi)}{\partial(\rho,\alpha,\beta)} \xrightarrow[n\to\infty]{d} N\Big(0,\Sigma_B^*\Big)$. Combined with the assumed distribution for $\tilde{\theta}$, $\frac{1}{\sqrt{n}}m(\tilde{\theta}-\theta) \xrightarrow[m\to\infty]{d} N\Big(0,r^{-1}AV_s(\theta)\Big)$, and since the two limiting distributions are from two independent samples, the combination vector has a normal distribution

$$\begin{pmatrix} \frac{1}{\sqrt{n}}\frac{\partial l(\phi)}{\partial(\alpha,\beta,\theta)} \\ \frac{1}{\sqrt{n}}m\big(\hat{\theta}-\theta\big) \end{pmatrix} \xrightarrow[n\to\infty]{d} N\Big(0,\Sigma_B\Big)$$

where

$$\Sigma_B = \begin{pmatrix} \Sigma_B^* & 0 \\ 0 & r^{-1}AV_s(\theta) \end{pmatrix}.$$

The coefficients of the right-hand side of (B.2) combined with $\frac{1}{\sqrt{n}}m(\tilde{\theta}-\theta)$ converge to a constant matrix Π_B almost surely, by SLLN:

$$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2 l(\phi)}{\partial(\rho,\alpha,\beta)^2} & \frac{\partial^2 l(\phi)}{\partial(\rho,\alpha,\beta)\partial\theta} \\ 0 & -m \end{pmatrix} \xrightarrow[n\to\infty]{a.s.} \Pi_B.$$

Therefore, by Slutsky's Theorem, the asymptotic distribution for the type B pseudo-MLE $(\hat{\rho}_B,\hat{\alpha}_B,\hat{\beta}_B,\tilde{\theta})$ is

$$\sqrt{n}\Big((\hat{\rho}_B,\hat{\alpha}_B,\hat{\beta}_B,\tilde{\theta})'-(\rho,\alpha,\beta,\theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0,\Pi_B^{-1}\Sigma_B(\Pi_B^{-1})'\Big).$$

Type AB Pseudo-MLE Finding the type AB pseudo-MLE $\hat{\phi}_{AB}=(\hat{\rho}_A,\hat{\alpha}_{AB},\hat{\beta}_{AB},\tilde{\theta})$ is equivalent to solving the equations $\frac{\partial l_2(\rho)}{\partial\rho}=0$, $\frac{\partial l_1(\phi)}{\partial(\alpha,\beta)}=0$, and $m(\tilde{\theta}-\theta)=0$, simultaneously. The first-order Taylor expansion of the functions $\frac{\partial l_2(\rho)}{\partial\rho}|_{\rho=\hat{\rho}_A}$ and $\frac{\partial l_1(\phi)}{\partial(\alpha,\beta)}|_{\phi=\hat{\phi}_{AB}}$, whose values are zero, yields the following equations:

$$\frac{1}{\sqrt{n}}\frac{\partial l_2(\rho)}{\partial\rho} \doteq -\frac{1}{n}\frac{\partial^2 l_2(\rho)}{\partial\rho^2}\sqrt{n}(\hat{\rho}_A-\rho) \tag{B.3}$$

and

$$\frac{1}{\sqrt{n}}\frac{\partial l_1(\phi)}{\partial(\alpha,\beta)} \doteq -\frac{1}{n}\frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta)\partial\rho}\sqrt{n}(\hat{\rho}_A-\rho)-\frac{1}{n}\frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta)^2}\sqrt{n}\Big((\hat{\alpha}_{AB},\hat{\beta}_{AB})'-(\alpha,\beta)\Big)-\frac{1}{n}\frac{\partial^2 l_1(\phi)}{\partial(\alpha,\beta)\partial\theta}\sqrt{n}(\tilde{\theta}-\theta). \tag{B.4}$$

By the CLT, the left-hand sides of (B.3) and (B.4) converge to a multivariate normal distribution with mean zero and variance Σ^*_{AB}, as $n \to \infty$,

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \frac{\partial l_2(\rho)}{\partial \rho} \\ \frac{1}{\sqrt{n}} \frac{\partial l_1(\phi)}{\partial (\alpha,\beta)} \end{pmatrix} \xrightarrow[n\to\infty]{d} N\Big(0, \Sigma^*_{AB}\Big).$$

Combined with the assumed distribution for $\tilde{\theta}$, $\frac{1}{\sqrt{n}} m(\tilde{\theta} - \theta) \xrightarrow[m\to\infty]{d} N\Big(0, r^{-1} A V_{\tilde{\theta}}(\theta)\Big)$, and since the two limiting distributions are from two independent samples, the combination vector has a normal distribution

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \frac{\partial l_2(\rho)}{\partial \rho} \\ \frac{1}{\sqrt{n}} \frac{\partial l_1(\phi)}{\partial (\alpha,\beta)} \\ \frac{1}{\sqrt{n}} m(\tilde{\theta} - \theta) \end{pmatrix} \xrightarrow[n\to\infty]{d} N\Big(0, \Sigma_{AB}\Big)$$

where

$$\Sigma_{AB} = \begin{pmatrix} \Sigma^*_{AB} & 0 \\ 0 & r^{-1} A V_{\tilde{\theta}}(\theta) \end{pmatrix}.$$

The coefficients of the right-hand sides of (B.3) and (B.4) combined with $\frac{1}{\sqrt{n}} m(\tilde{\theta} - \theta)$ converge to a constant matrix Π_{AB} almost surely, by SLLN:

$$-\frac{1}{n} \begin{pmatrix} \frac{\partial^2 l_2(\rho)}{\partial \rho^2} & 0 & 0 \\ \frac{\partial^2 l_1(\phi)}{\partial (\alpha,\beta) \partial \rho} & \frac{\partial^2 l_1(\phi)}{\partial (\alpha,\beta)^2} & \frac{\partial^2 l_1(\phi)}{\partial (\alpha,\beta) \partial \theta} \\ 0 & 0 & -m \end{pmatrix} \xrightarrow[n\to\infty]{a.s.} \Pi_{AB}.$$

Therefore, by Slutsky's Theorem, the asymptotic distribution for the type AB pseudo-MLE $(\hat{\rho}_A, \hat{\alpha}_{AB}, \hat{\beta}_{AB}, \tilde{\theta})$ is

$$\sqrt{n}\Big((\hat{\rho}_A, \hat{\alpha}_{AB}, \hat{\beta}_{AB}, \tilde{\theta})' - (\rho, \alpha, \beta, \theta)'\Big) \xrightarrow[n\to\infty]{d} N\Big(0, \Pi^{-1}_{AB} \Sigma_{AB} (\Pi^{-1}_{AB})'\Big).$$