Measurement Error and Misclassification in Interval-Censored Life History Data

by

Bethany J. Giddings White

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics - Biostatistics

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In practice, data are frequently incomplete in one way or another. It can be a significant challenge to make valid inferences about the parameters of interest in this situation. In this thesis, three problems involving such data are addressed. The first two problems involve interval-censored life history data with mismeasured covariates. Data of this type are incomplete in two ways. First, the exact event times are unknown due to censoring. Second, the true covariate is missing for most, if not all, individuals. This work focuses primarily on the impact of covariate measurement error in progressive multi-state models with data arising from panel (i.e., interval-censored) observation. These types of problems arise frequently in clinical settings (e.g. when disease progression is of interest and patient information is collected during irregularly spaced clinic visits). Two and three state models are considered in this thesis. This work is motivated by a research program on psoriatic arthritis (PsA) where the effects of error-prone covariates on rates of disease progression are of interest and patient information is collected at clinic visits (Gladman et al. 1995; Bond et al. 2006). Information regarding the error distributions were available based on results from a separate study conducted to evaluate the reliability of clinical measurements that are used in PsA treatment and follow-up (Gladman et al. 2004). The asymptotic bias of covariate effects obtained ignoring error in covariates is investigated and shown to be substantial in some settings. In a series of simulation studies, the performance of corrected likelihood methods and methods based on a simulation-extrapolation (SIMEX) algorithm (Cook & Stefanski 1994) were investigated to address covariate measurement error. The methods implemented were shown to result in much smaller empirical biases and empirical coverage probabilities which were closer to the nominal levels.

The third problem considered involves an extreme case of interval censoring known as current status data. Current status data arise when individuals are observed only at a single point in time and it is then determined whether they have experienced the event of interest. To complicate matters, in the problem considered here, an unknown proportion of the population will never experience the event of interest. Again, this type of data is incomplete in two ways. One assessment is made on each individual to determine whether or not an event has occurred. Therefore, the exact event times are unknown for those

who will eventually experience the event. In addition, whether or not the individuals will ever experience the event is unknown for those who have not experienced the event by the assessment time. This problem was motivated by a series of orthopedic trials looking at the effect of blood thinners in hip and knee replacement surgeries. These blood thinners can cause a negative serological response in some patients. This response was the outcome of interest and the only available information regarding it was the seroconversion time under current status observation. In this thesis, latent class models with parametric, nonparametric and piecewise constant forms of the seroconversion time distribution are described. They account for the fact that only a proportion of the population will experience the event of interest. Estimators based on an EM algorithm were evaluated via simulation and the orthopedic surgery data were analyzed based on this methodology.

Acknowledgements

I would like to take this opportunity to express my upmost gratitude and appreciation to my supervisors, Dr. Richard J. Cook and Dr. Grace Y. Yi for their guidance, valuable insight and support over the course of my graduate studies. I have learned a lot from them. I would also like to thank my departmental committee members, Dr. Mary E. Thompson, Dr. K. Stephen Brown for reviewing my thesis and sharing their constructive suggestions and comments. Thanks also to Dr. Suzanne L. Tyas (University of Waterloo) and Dr. Y. Paul Peng (Queen's University) for agreeing to sit on my thesis committee. Their helpful feedback on my work was very much appreciated. I am also very grateful to Ker-Ai Lee for her help with statistical programming over the past couple of years.

Many thanks are extended to Dr. Dafna Gladman (University of Toronto) and Dr. Theodore Warkentin (McMaster University) for providing the Psoriatic Arthritis and Orthopedic Surgery study data. These studies motivated the development of much of this work.

This research was supported by the Natural Sciences and Engineering Council of Canada (NSERC) and through several research assistantships with Dr. Richard J. Cook. I am grateful for this funding as it gave me freedom to concentrate my efforts on this research.

Last, but certainly not least, I would like to express my love and heartfelt thanks to my wonderful husband, Adam White, loving parents, Percy and Mary Giddings, and the rest of my family and friends for their tremendous support, encouragement and patience throughout this journey. You are all special people and I am very blessed to have you in my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Life history data are frequently collected for use in investigations within disciplines such as medicine, epidemiology, biology, sociology, economics, engineering and actuarial science (Kalbfleisch & Prentice 2002; Lawless 2003). Such data arise when individuals are observed over time and information on the occurrence of one or more events for these individuals is collected. Unfortunately, this type of data are often incomplete in practice. Exact event times are often unknown and the covariate measurements collected may be prone to error. In this thesis, methodology dealing with different types of incomplete life history data will be explored.

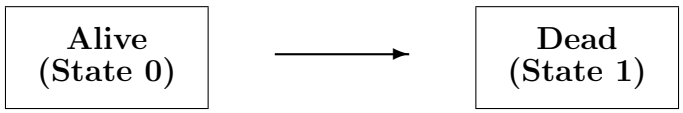
1.1 Interval Censored Life History Data

Life history data can be represented in two closely related ways. One is the *multi-state* framework, in which case a multi-state model is used to feature the data. A multi-state model is a model for a stochastic process in which a response can occupy one of a set of possible discrete states at any time. The second way that life history data can be represented is through the *event occurrence* framework. Counting processes can be used to formulate models under this framework. In contrast to the multi-state framework, in this case it is the number of occurrences of a particular event in a given time interval that is recorded. Some problems lend themselves naturally to the multi-state framework and some to the event occurrence framework, although many problems are amenable to both

(Kalbfleisch & Lawless 1999). Consider recurrent events for example. Recurrent events arise when transient events can occur repeatedly to an individual over time. Examples include seizures suffered by persons with epilepsy, damaged joints in a patient suffering from arthritis and failures of a piece of equipment or software. An even more complex setting involves consideration of multiple events that may occur simultaneously, either once or repeatedly, to individuals over time (Kalbfleisch & Lawless 1999). The number of states and the transitions which are possible are dictated by the problem being considered. Under the multi-state framework, states can be formed by defining categories based on the total number of events experienced. In an application that will be considered in this research, the recurrent event is joint damage in a study of arthritis. One way to define states is by the number of damaged joints so that the states essentially represent the severity of arthritis (Gladman et al. 1995). The data would then consist of a count of the number of joints observed to be damaged over a specific time interval. Alternatively, the states could be defined by different combinations of damaged joints, so that the severity of arthritis would be classified by the relative importance of groups of damaged joints.

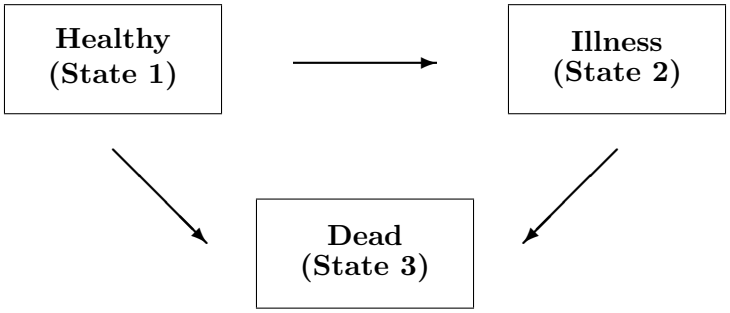
In this thesis the focus will be on the multi-state framework. With an event defined as a transition between two states, multi-state models provide convenient representations for most life history problems. The state structure defines the states and illustrates the possible transitions (Hougaard 1999). Some examples of these structures are given below in Figures 1.1 to 1.4.

Figure 1.1: *Two-state lifetime model involving only one possible transition (i.e. death).*



Ideally, the transition times as well as the states will be recorded for all individuals. However, this is often not the case. For instance, it is rarely the case that all individuals are observed until they enter an absorbing state, so the transition times are *right-censored*.

Figure 1.2: *Three-state illness-death model.*



Right censoring may have an impact on inference if the censoring mechanism is dependent (or conditionally dependent) on the event occurrence mechanism (i.e. the observed sample including incomplete observations is not representative of the population in absence of censoring) (Andersen & Keiding 2002). Other forms of incomplete data can arise when individuals are excluded from the study based on their stage in the process. For instance, this may occur if only individuals who experience a precipitating event are included in the study (*truncation*) (Matthews & Cook 2005). Again, if this is not going to be accounted

Figure 1.3: *Competing risks model (i.e. multiple modes of failure) involving K possible transitions.*

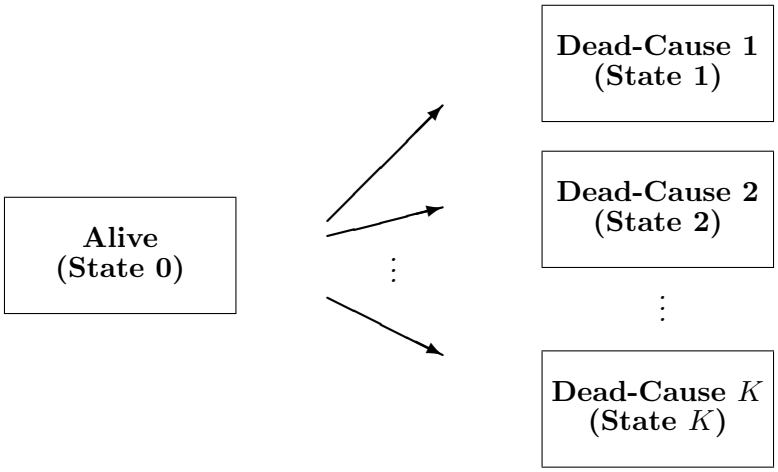
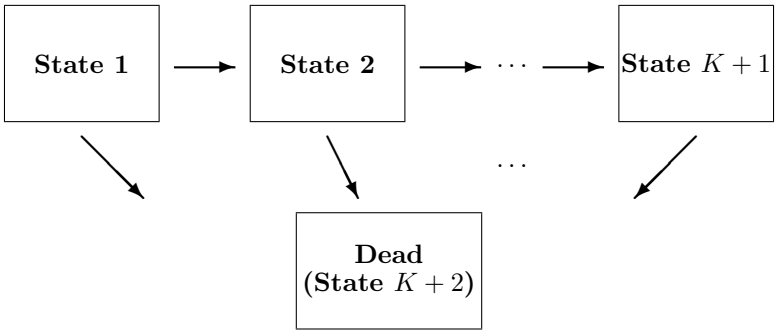


Figure 1.4: *Progressive model involving* $2K+1$ *possible transitions (including the possibility of transition to an absorbing state).*



for explicitly in the analysis, it is important that these individuals do not experience systematically higher or lower risks of experiencing the event(s) over the unobserved durations than the population of interest.

Sometimes individuals are observed at prespecified assessment times and their states are determined only at these times. Information about transitions between successive observation times is unavailable. This type of data are sometimes referred to as *panel data* in the context of multi-state models (Kalbfleisch & Lawless 1989) or *interval-censored lifetime data* in survival analysis. This type of data arises naturally in settings such as clinical trials where patients are examined by physicians periodically and their states are assessed at those visits. As in the case of censoring and truncation, inference in this case may be affected if the life history process and the follow-up process are not independent. If this is the case, the follow-up process may contain information on the life history process so both processes must be modeled simultaneously to ensure the validity of inference. If they are independent it is a much simpler problem since only the life history process must be modeled. Therefore, in the panel data case, it is usually assumed that the follow-up times are specified in advance or that the follow-up process is independent of the life history process (Grüger et al. 1991). However, these are often unrealistic assumptions in clinical settings. Grüger et al. (1991) present additional noninformative assessment schemes under which standard statistical inferences are still valid:

- assessment at regularly spaced intervals,
- any assessment scheme (regular or irregular) that has been fixed in advance,
- random assessment times that are independent of the life histories of the subjects under study, and
- in the case of clinical studies, a *doctor's care assessment scheme* where the doctor monitoring a patient is permitted to set the next assessment time depending on the state the patient occupies at the current assessment.

Patient self-selection of assessment times may be informative so if this is the case, the follow-up process must be taken into account in the likelihood to conduct valid statistical inference (Grüger et al. 1991). Even though we do not obtain complete information regarding the movements through states for a given individual, the data we do obtain can still provide valuable information regarding the parameters of interest. Modeling and inference in the presence of censoring will be discussed in the following sections.

1.1.1 Two-state Models

The simplest state structure involves two states. For example, a mortality model involves only two states, *Alive* and *Dead*. This structure is illustrated in Figure 1.1. The *Dead* state is called an absorbing state, as once it is entered an individual cannot move back to the *Alive* state. This could represent, for instance, death of an individual or failure of a piece of equipment. All individuals are expected to eventually make the transition between states. However, in practice there are situations where this is not necessarily the case. These will be discussed further for cure rate data in Chapter 4. Under the simpler model, however, there is only one possible transition to consider and all subjects will eventually make the transition. It is characterized by a *hazard function*, $\lambda(t)$, which is a function of parameter(s) and may also be a function of time and covariates. There has been much work done in developing methodology to deal with this type of data (Lawless 2003). Analysis in this situation is referred to as *Lifetime Data Analysis* or *Survival Analysis*. A slightly more complex state structure permits movement back and forth between the two states.

Lifetime data can be characterized by certain distributions. First, let T be a non-negative random variable representing time to failure or death. Depending on how the data are collected and summarized, it may be continuous (the exact time is collected) or discrete (the lifetimes are grouped in some way). Considering continuous T, let the probability density function (*p.d.f.*), of T be denoted by $f_T(t)$. Then the cumulative distribution function (*c.d.f.*) is given by

$$F_T(t) = P(T \leq t) = \int_0^t f_T(x)dx. \tag{1.1}$$

From this, we can define the survivor function, which is the probability that an individual survives to time t.

$$\mathcal{F}_T(t) = P(T > t) = 1 - F_T(t) = \int_t^\infty f_T(x)dx. \tag{1.2}$$

$\mathcal{F}_T(t)$ is a non-increasing continuous function with $\mathcal{F}_T(0) = 1$ and $\lim_{t\to\infty} \mathcal{F}_T(t) = 0$. An extremely important function in the characterization of life history data is the hazard function (or transition intensity function, in the case of multi-state models). It is essentially the instantaneous probability of death, failure or transition between states at time t, given the individual survives or remains in the current state up to time t. It can be written as

$$\lambda(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f_T(t)}{\mathcal{F}_T(t)}. \tag{1.3}$$

Equivalently, if we let

$$N(t) = \begin{cases} 1, & \text{if event occurs at time t} \\ 0, & \text{otherwise} \end{cases},$$

then the hazard function is

$$\lambda(t) = \lim_{\Delta t\to 0} \frac{P(\Delta N(t) = 1 | T \geq t)}{\Delta t}, \tag{1.4}$$

where $\Delta N(t) = N(t + \Delta t) - N(t)$. This expression is similar to those for the transition intensities under a multi-state framework. Another function which is of interest when dealing with life history data is the cumulative hazard function:

$$\Lambda(t) = \int_0^t \lambda(s)ds. \tag{1.5}$$

Any of the functions, $f_T(t)$, $F_T(t)$, $\mathcal{F}_T(t)$, $\lambda(t)$, or $\Lambda(t)$ are sufficient to specify the distribution of T (Lawless 2003). Often, however, the hazard function is used as the basis for analysis.

There are several possible approaches one can take when modeling a lifetime distribution. These include parametric, semi-parametric and nonparametric models. Parametric modeling involves specification of the lifetime distribution up to a vector of unknown parameters $\boldsymbol{\theta}$. The Exponential, Weibull, Log-logistic, Log-normal and Gamma distributions are but a few of the possible candidates for the distribution of T. A generalization to these parametric models involves the more flexible weakly parametric models. Rather than the hazard functions that exist under fully parametric models, a piecewise constant hazard function may be assumed. Parametric models are attractive because estimation and inference are relatively straightforward. However, a specific parametric form has to be deemed appropriate for the data and this is often not a trivial task. An alternative is to carry out nonparametric estimation. These models do not force a functional form on the data. A widely used nonparametric estimate in survival analysis is the Kaplan-Meier or product limit estimate of the survivor function (Kaplan & Meier 1958). This is similar to the standard empirical estimate of the survivor function with some modifications to account for the fact that when dealing with censored data, the number of failure times greater than or equal to a certain time, t, are not usually known exactly (Lawless 2003). Confidence limits on these estimates can also be obtained. Often these nonparametric estimates are used to assess the appropriateness of parametric models when performing diagnostics (Lawless 2003; Matthews & Cook 2005).

Interest frequently lies in investigating the effects of covariates, $\boldsymbol{z}$, on the time to failure or death. To do so we could adopt a *proportional hazards* regression model if it is thought that the covariates have a multiplicative effect on the hazard function:

$$\lambda(t) = \lambda_0(t)\phi(\boldsymbol{z}), \tag{1.6}$$

where $\phi(\cdot)$ is some specified function. This can be a parametric model if the baseline transition intensity, $\lambda_0(t)$, is assumed to have a parametric form. Alternatively, a semi-parametric approach could be taken if this baseline transition intensity is left arbitrary.

The *relative risk function*, $\phi(\boldsymbol{z})$, can take on various parametric forms such as:

- *log-linear form*: $\phi(\boldsymbol{z};\boldsymbol{\beta}) = \mathrm{e}^{\boldsymbol{\beta}'\boldsymbol{z}}$,
- *linear form*: $\phi(\boldsymbol{z};\boldsymbol{\beta}) = 1 + \boldsymbol{\beta}'\boldsymbol{z}$, and
- *logistic form*: $\phi(\boldsymbol{z};\boldsymbol{\beta}) = \log(1 + e^{\boldsymbol{\beta}'\boldsymbol{z}})$.

The *Cox Model*, which is simply a proportional hazards model with a log-linear relative risk, is widely used in practice.

Inference may be conducted based on the adopted model via maximum likelihood estimation. The first step here is to determine the likelihood function based on the probability of observing the data that were actually collected. This will be a function of the unknown parameters that we can maximize to determine which values of the parameters are most likely to give rise to the observed data. Suppose $\boldsymbol{\theta}$ is a p-dimensional vector of unknown parameters upon which the distribution of random variable Y depends. For p.d.f. $f(y;\boldsymbol{\theta})$, the likelihood function based on a random sample $y_1, y_2, ..., y_n$ is $\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta})$. The maximum likelihood estimate, $\hat{\boldsymbol{\theta}}$, is usually found by maximizing the log-likelihood function with respect to $\boldsymbol{\theta}$. The log-likelihood function would be $l(\boldsymbol{\theta}) = \log(\mathcal{L}(\boldsymbol{\theta})) = \sum_{i=1}^{n} \log(f(y_i;\boldsymbol{\theta}))$ and the maximum likelihood estimate would be obtained by setting the score functions to $\mathbf{0}$; $U_j(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta})/\partial\theta_j = 0$ for $j = 1, 2, ..., p$. Denote the maximum likelihood estimator based on a sample of size n as $\hat{\boldsymbol{\theta}}_n$. Then, under certain mild regularity conditions, $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathcal{I}^{-1}(\boldsymbol{\theta})$. The matrix $\mathcal{I}(\boldsymbol{\theta})$ is called the Fisher or expected information matrix and its (i,j) element is defined as: $\mathcal{I}_{ij}(\boldsymbol{\theta}) = E\left(-\partial^2 l(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j\right)$, $i, j = 1, 2, ..., p$. The estimator $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}$ and the observed information matrix $n^{-1}I(\hat{\boldsymbol{\theta}})$ is a consistent estimator of $n^{-1}\mathcal{I}(\boldsymbol{\theta})$, where the (i,j) element of $I(\hat{\boldsymbol{\theta}})$ is $\left(-\partial^2 l(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j\right)\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. These and several other asymptotic results involving maximum likelihood estimators lead to useful inferences (Lawless 2003).

Likelihood functions are presented below for several types of incomplete data that arise in lifetime data analysis. For the purposes of this development we will assume T

is continuous. Likelihood function formulation follows in a similar way for discrete time, T. Assume a parametric form is appropriate so the distribution of T is specified up to an unknown vector of parameters, $\boldsymbol{\theta}$. Therefore, we have p.d.f. $f_T(t;\boldsymbol{\theta})$, c.d.f. $F_T(t;\boldsymbol{\theta})$, survivor function $\mathcal{F}_T(t;\boldsymbol{\theta})$ and hazard function $\lambda(t;\boldsymbol{\theta})$. Consider first the case where all n subjects are observed until their failure time or time to death. Therefore, we observe $t_1, t_2, t_3, ..., t_n$ and the likelihood function is:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_T(t_i;\boldsymbol{\theta}). \tag{1.7}$$

Unfortunately complete survival data are usually not obtained for each subject in practice. Most datasets include complete data on some subjects and incomplete data on others.

The most common type of incomplete data arises due to right censoring. This occurs when the study ceases or an individual is lost to follow-up prior to experiencing the event of interest. Type I Censoring describes the situation where each subject has a fixed potential censoring time $r_i > 0$ such that T_i is observed if $T_i \leq r_i$ (Lawless 2003). Therefore, for individuals with a right-censored event time (i.e. r_i is less than the failure time for subject i), all we know is that their event time is larger than their censoring time. For each subject i, $i = 1, 2, ..., n$, the data collected in the presence of right censoring would be (u_i, δ_i), where $\delta_i = I(T_i \leq r_i)$ and $u_i = \min(T_i, r_i)$. Then, the likelihood function is:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[f_T(u_i;\boldsymbol{\theta})\right]^{\delta_i} \left[\mathcal{F}_T(u_i;\boldsymbol{\theta})\right]^{1-\delta_i} \\
&= \prod_{i=1}^{n} f_T(u_i;\boldsymbol{\theta}) \left[\lambda_T(u_i;\boldsymbol{\theta})\right]^{\delta_i - 1}.
\end{aligned}$$

An extension of this involves a random censoring time, R, rather than a fixed potential censoring time for each subject. Sometimes the censoring process is linked to the time to event process and therefore must be taken into account when estimating the parameters of interest. Another variation of right censoring is Type II Censoring. It involves the situation where only the s shortest lifetimes are observed where s is chosen in advance. The objective in these types of schemes is the efficient use of study resources. In this case $t_{(1)} \leq t_{(2)} \leq t_{(3)} \cdots \leq t_{(s)}$ are observed. The study is then stopped and censored event

times are recorded for the remaining subjects. The likelihood function is then based on the joint distribution of order statistics (Lawless 2003).

Left-censored data arise when the event is known to have occurred prior to a certain time, l, but the exact time is unknown. For instance, consider a study investigating the age at development of a particular health condition. People may enter the study having already been diagnosed, however, there may be no record of the exact time of onset (Lee & Wang 2003). Now the data for subject i with left censoring time l_i would be (u_i, η_i), where $\eta_i = I(l_i \leq T_i)$ and $u_i = \max(T_i, l_i)$. For subjects with left-censored event times, the contribution to the likelihood would be $F_T(u_i; \boldsymbol{\theta})$. Therefore the likelihood in the presence of Type I right and left censoring is:

$$\mathcal{L}(\boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} \left[f_T(u_i;\boldsymbol{\theta})\right]^{\delta_i \eta_i} \left[\mathcal{F}_T(u_i;\boldsymbol{\theta})\right]^{1-\delta_i} \left[F_T(u_i;\boldsymbol{\theta})\right]^{1-\eta_i}.$$

Interval censoring is quite common in survival analysis. This would occur if subjects were being examined at intermittent times and the event of interest occurred between assessment times. Sun (2006) provides an excellent survey of methodology for interval-censored life history data. The exact event time is unknown; it is known only to lie between the two examination times, c_i and d_i, say. That is, $c_i \leq T_i < d_i$ for subject i. The contribution to the likelihood function by this individual will be $F_T(d_i;\boldsymbol{\theta}) - F_T(c_i;\boldsymbol{\theta})$ or equivalently, $\mathcal{F}_T(c_i;\boldsymbol{\theta}) - \mathcal{F}_T(d_i;\boldsymbol{\theta})$. Let $\Delta_i = I(T_i \leq c_i)$ and $\Gamma_i = I(c_i \leq T_i < d_i)$. The other types of data can be considered as special cases of interval-censored data. Specifically, $c_i = d_i$ when an exact event time is observed, $d_i = \infty$ for right-censored data and $c_i = 0$ for left-censored data. Given the notation introduced above, we can build a likelihood function for complete observations and the types of censored data discussed above. With the data for subject i given as $(c_i, d_i, \Delta_i, \Gamma_i)$ for $i = 1, 2, 3, ..., n$, the likelihood function will be:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f_T(c_i;\boldsymbol{\theta})\right]^{\Delta_i\Gamma_i} \left[F_T(d_i;\boldsymbol{\theta}) - F_T(c_i;\boldsymbol{\theta})\right]^{(1-\Delta_i)\Gamma_i} \left[F_T(c_i;\boldsymbol{\theta})\right]^{\Delta_i(1-\Gamma_i)} \left[\mathcal{F}_T(d_i;\boldsymbol{\theta})\right]^{(1-\Delta_i)(1-\Gamma_i)}. \quad (1.8)$$

The first contribution is from observed event times, the second from interval-censored event times, the third from left-censored times (with censoring times, c_i) and the fourth from right-censored times (with censoring times d_i). Note that these likelihood functions are

based on the assumption that the censoring times are fixed for each subject. If this is not reasonable, the censoring process must be modeled and incorporated into the analysis.

Truncation is another form of incomplete data that arise in life history data. In the presence of truncation, likelihoods are expressed based on conditional distributions. The data do not appear to be different for censored and truncated data. The main difference between censoring and truncation is that truncation actually has an impact on the units selected for the sample, whereas censoring results in incomplete data on the life history process for a unit in the study (Commenges 2002). Truncated data will not be considered in this research.

If a semi-parametric approach is taken and the proportional hazards model, (1.6), deemed appropriate, inference on the parameters of interest can be conducted using the partial likelihood function rather than the full likelihood function. The partial likelihood function is obtained by factoring the full likelihood function into conditional probabilities and discarding the terms which involve nuisance parameters (Lawless 2003). This is much simpler since the baseline hazard function and any parameters upon which it depends are not included in the partial likelihood. However, there is usually a loss of information when this method is used and this loss is difficult to assess (Lawless 2003). Additional information on the derivation of this partial likelihood can be found in Cox (1975) and Matthews & Cook (2005) and information on its asymptotic properties can be found in Andersen et al. (1993).

1.1.2 Multi-state Models

State structures and analyses become more complex as the number of states and possible transitions increase. Examples of these include illness-death models (Figure 1.2), competing risks models (Figure 1.3) and progressive models (Figure 1.4). As in the case of the two-state structure, it is often the intensities associated with the transitions which are of interest in the analysis. Roughly speaking these represent the instantaneous probability of transition at time t. The intensities are frequently modeled as a function of covariates that are believed to be relevant to the process. These covariates may be fixed or time-varying;

however, if a time-varying covariate process may be influenced by the life history process (these are known as internal covariates), the covariate process must be modeled in addition to the life history process and interpretation of model parameters will not be as straightforward as in the case of external covariates (Kalbfleisch & Prentice 2002). Transition intensities can also be modeled as a function of time if it is believed that they vary with time (Kalbfleisch & Lawless 1999).

Under a general multi-state framework interest often lies in transitions between states or in the durations of the sojourns in states or times between successive state transitions. When only panel data are available, the exact transition times are unknown and therefore an analysis of the durations of the sojourns is not convenient. In this case, it is the transitions between states that are of interest. There are many different ways to model data under the multi-state framework. Hougaard (1999) gives a concise review of such models. We will consider first a single sample problem. The most commonly adopted model is based on Markov processes. A process $\{Y(t),\ t \geq 0\}$ with state space $1, ..., K$ can be modeled as a continuous-time Markov process if for all $0 \leq s \leq t$ and $j, k = 1, ..., K$,

$$P\left\{Y(t) = k | Y(s) = j, Y(u) = y(u), 0 \leq u < s\right\} = P\left\{Y(t) = k | Y(s) = j\right\}. \tag{1.9}$$

To obtain an expression for the transition intensities for a general multi-state model similar to the two-state version in (1.4), let

$$N_{jk}(t) = \begin{cases} 1, & \text{if transition } j \to k \text{ occurs at time t} \\ 0, & \text{otherwise} \end{cases},$$

and $\boldsymbol{N(t)} = \{N_{jk}(t) : j, k = 1, 2, ..., K\}$. Then the transition intensities can be written as

$$\lambda_{jk}(t|H(t)) = \lim_{\Delta t \to 0} \frac{P\left(\Delta N_{jk}(t) = 1 | H(t)\right)}{\Delta t}, \tag{1.10}$$

where $H(t) = \{\boldsymbol{N(s)}, 0 \leq s < t\}$ is the state path or history up to time t (Kalbfleisch & Lawless 1999). It is appropriate to omit the state history of an individual in (1.10) and write $\lambda_{jk}(t|H(t)) = \lambda_{jk}(t)$ if the Markovian Property assumption given in (1.9) is reasonable. This property holds if the conditional distribution of the future states given the current and past states depends only on the current state and is independent of the past

state path (Ross 1993). Under this model, analysis is greatly simplified as the transition intensities do not depend on the entire state path. A further simplification is achieved when the intensities are time homogeneous. In this case, $\lambda_{jk}(t) = \lambda_{jk}$ is independent of t for all $j, k = 1, ..., K$. It is often the case that the assumption of time-homogeneity is not appropriate. A useful compromise which still exploits some of the favorable properties of the time-homogeneous models is the use of piecewise constant transition intensities. In the presence of covariates, piecewise constant baseline transition intensities with 4-10 pieces have been found to be generally robust even when the true underlying intensities are smooth functions (Lawless & Zhan 1998). Semi-Markov models are appropriate when the transition probabilities depend on the time since the last transition as well as the current state (Kalbfleisch & Lawless 1999). Again, since the transition times are unknown in the panel data case, semi-Markov models are not readily adopted. However, some other general non-Markovian models which do not depend on transition times, can be applied to panel data. In these models, the transition intensities can be permitted to depend on the past state path in any way. This can be accommodated for progressive state structures. However, for non-progressive state structures, these general models are not feasible since it is often extremely difficult, if not impossible, to write general formulas for the transition probabilities (Hougaard 1999).

A general multi-state model with state space $1, 2, ..., K$ can be described via the following transition intensity matrix, $Q(t)$:

$$Q(t) = \begin{pmatrix} -\sum_{j=2}^{K} \lambda_{1j}(t) & \lambda_{12}(t) & \cdots & \lambda_{1,K-1}(t) & \lambda_{1K}(t) \\ \lambda_{21}(t) & -\sum_{j=1,j\neq 2}^{K} \lambda_{2j}(t) & \cdots & \lambda_{2,K-1}(t) & \lambda_{2K}(t) \\ \lambda_{31}(t) & \lambda_{32}(t) & \cdots & \lambda_{3,K-1}(t) & \lambda_{3K}(t) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{K-1,1}(t) & \lambda_{K-1,2}(t) & \cdots & -\sum_{j=1,j\neq K-1}^{K} \lambda_{K-1,j}(t) & \lambda_{K-1,K}(t) \\ \lambda_{K1}(t) & \lambda_{K2}(t) & \cdots & \lambda_{K,K-1}(t) & -\sum_{j=1}^{K-1} \lambda_{Kj}(t) \end{pmatrix} \tag{1.11}$$

Note from the above matrix, the diagonal elements are given by $\lambda_{kk}(t) = -\sum_{j=1,j\neq k}^{K} \lambda_{kj}(t)$ for $k = 1, 2, ..., K$. It is convention to define these transition intensities in this manner to

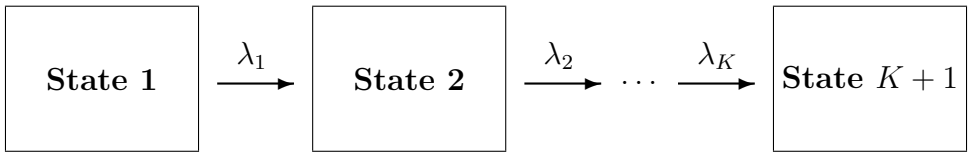
satisfy the constraints $\sum_{j=1}^{K} \lambda_{kj}(t) = 0$, $k = 1, 2, ...K$ (Kalbfleisch & Lawless 1999). Fixed covariates can be easily incorporated into the formulations by expressing the transition intensities as a function of time and the covariates, $\lambda_{jk}(t) = g(t, \boldsymbol{z})$, for some non-negative function g. A multiplicative model is frequently used in practice. For a given individual (with subject subscripts suppressed) we often adopt models of the form

$$\lambda_{jk}(t) = \lambda_{0jk}(t) \exp(\boldsymbol{\beta}'_{jk} \boldsymbol{z}_{jk}), \tag{1.12}$$

where $\lambda_{0ij}(t)$ are the baseline transition intensities which may or may not depend on t and $\boldsymbol{\beta}_{jk}$ is a vector of regression coefficients associated with fixed covariates of interest, $\boldsymbol{z}_{jk}$, $j, k = 1, 2, ..., K$. Here, the baseline transition intensities and the regression coefficients are permitted to vary across the possible transitions. This is analogous to the proportional hazards model commonly applied in lifetime data analysis (see (1.6)).

A progressive state structure, such as that presented in Figure 1.4, is much simpler than the general K-state model specified by the transition intensity matrix in (1.11). Consider $K+1$ distinct states that individuals may occupy at any given time. These could represent disease stages, for instance. Suppose the transition intensities between the states are of interest and the last state in the progression (State $K + 1$) is an absorbing state and can only be reached through transition from State K. The state structure associated with this problem (Figure 1.5) is a slightly simpler version of that given in Figure 1.4. In addition,

Figure 1.5: *Progressive model involving K transition intensities.*



suppose that there are n individuals who are monitored periodically over the course of the study so that for subject i there are m_i sets of observations at times $\{u_{ij}; j = 1, ..., m_i\}$. Each set of observations will include the state occupied by individual i and may also include measurements on covariates. For the purposes of this work we will consider only fixed

covariates (i.e. baseline values of covariates). Therefore, for subject i, the data consist of $(u_{ij}, y_i(u_{ij}), \boldsymbol{z}_i)$ for $j = 1, 2, ..., m_i$. We will assume that all subjects enter the study at time 0 ($u_{i0} = 0$ for $i = 1, 2, ...n$) in State 1 ($y_i(u_{i0}) = 1$ for $i = 1, 2, ...n$).

To model this process, assume a multiplicative model similar to (1.12) is appropriate. In particular, let

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp(\boldsymbol{\beta_k}' \boldsymbol{z_i}), \tag{1.13}$$

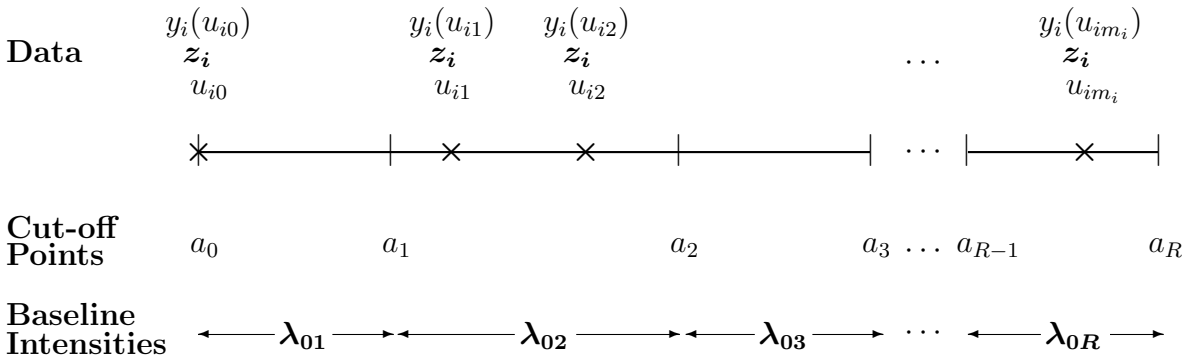
for $k = 1, ..., K$, represent the intensity associated with the $k \to k+1$ transition for subject i. Suppose $\lambda_{ik}(t)$ depends on time via a piecewise constant baseline intensity with R parts:

$$\lambda_{0k}(t) = \begin{cases} \lambda_{0k1}, & a_0 \leq t < a_1 \\ \lambda_{0k2}, & a_1 \leq t < a_2 \\ \lambda_{0k3}, & a_2 \leq t < a_3 \\ \quad \vdots & \quad \vdots \\ \lambda_{0kR}, & a_{R-1} \leq t < a_R \end{cases} . \tag{1.14}$$

Let $\boldsymbol{\lambda}_{0r}$ represent a vector of the baseline intensities for all transitions for $t \in [a_{r-1}, a_r)$ so that $\boldsymbol{\lambda_{0,r}} = (\lambda_{01r}, \lambda_{02r}, \cdots, \lambda_{0Kr})'$ for $r = 1, ..., R$. An extension of this model may allow for different numbers of piecewise constant baseline intensities for each transition. That is, rather than having the same number of intensities for each transition, R, we could have R_k, a number which depends on the transition, k. Considering the simpler model given in (1.14), an illustration of what may be observed for a given subject and the time-varying baseline transition intensities are displayed in Figure 1.6. Clearly the set-up can be quite complicated so care must be taken when constructing expressions to be used in estimation.

For general multi-state models as outlined above, the data obtained are interval-censored. To obtain the likelihood function, we require an expression for transition probabilities rather than the intensities. Unfortunately, in the case of general state structures, there is often no closed form for the transition probabilities which means a likelihood function cannot be formulated. Kalbfleisch & Lawless (1985) describe a method to obtain maximum likelihood estimates with panel data for general state structures under the assumption of time homogeneous intensities. Consider panel data under a continuous-time Markov

Figure 1.6: *An illustration of the observation process and the underlying baseline intensities in effect over time for an arbitrary subject, i.*



model with transition intensity matrix given by $Q(t)$ in (1.11). This method is primarily applicable for time-homogenous models where $\lambda_{jk}(t) = \lambda_{jk}$ although extensions to incorporate some simple forms of non-homogeneity are possible (Kalbfleisch & Lawless 1985; Gentleman et al. 1994). Suppose that a K-state time-homogeneous multi-state model is appropriate and let the transition intensities be characterized up to a vector of p functionally independent parameters, $\boldsymbol{\theta}$, so the transition intensity matrix is $Q(\boldsymbol{\theta}) = [\lambda_{jk}(\boldsymbol{\theta})]_{(K\times K)}$. Let $P(\cdot;\boldsymbol{\theta}) = [p_{jk}(\cdot;\boldsymbol{\theta})]_{(K\times K)}$ represent the transition probability matrix. Since we are dealing with a time-homogeneous problem, we have $P(s, s+t) = P(0,t) = P(t)$. Then, solving the forward Kolmogorov differential equation with $s = 0$, $dP(t;\boldsymbol{\theta})/dt = P(t;\boldsymbol{\theta})Q(\boldsymbol{\theta})$, with boundary condition $P(0;\boldsymbol{\theta}) = I$ gives the unique solution

$$P(t;\boldsymbol{\theta}) = \exp\left(Q(\boldsymbol{\theta})t\right). \tag{1.15}$$

It is of interest to estimate $\boldsymbol{\theta}$, so suppose a random sample of n individuals is observed at times $u_0, u_1, ..., u_m$. If we denote the number of individuals that transition from state j to k between u_{l-1} and u_l by n_{jkl} and condition on the initial distribution of individuals among the states, the likelihood function can be written as

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{l=1}^{m}\left\{\prod_{j,k=1}^{K}\left[p_{jk}(u_l - u_{l-1};\boldsymbol{\theta})\right]^{n_{jkl}}\right\}. \tag{1.16}$$

The function that we need to maximize with respect to $\boldsymbol{\theta}$ is the log-likelihood function:

$$l(\boldsymbol{\theta}) = \sum_{l=1}^{m} \sum_{j,k=1}^{K} n_{jkl} \log[p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})]. \tag{1.17}$$

If we were to proceed with the Newton-Raphson algorithm, first and second derivatives of the log-likelihood function would be required. These are as follows:

$$S_r(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_r} = \sum_{l=1}^{m} \sum_{j,k=1}^{K} \left[n_{jkl} \frac{\partial p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})/\partial \theta_r}{p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})} \right], \ r = 1, 2, ..., p, \tag{1.18}$$

and

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} = \sum_{l=1}^{m} \sum_{j,k=1}^{K} n_{jkl} \left\{ \frac{\partial^2 p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})/\partial \theta_r \partial \theta_s}{p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})} - \frac{[\partial p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})/\partial \theta_r]\,[\partial p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})/\partial \theta_s]}{p_{jk}^2(u_l - u_{l-1}; \boldsymbol{\theta})} \right\}. \tag{1.19}$$

These derivatives can be extremely difficult to obtain analytically since the transition probabilities are often complex functions of the intensities, if they can be written in closed form at all. However, given the form of the transition probability matrix in (1.15), we can take advantage of a canonical decomposition to help calculate these derivatives. If for a given value of $\boldsymbol{\theta}$, the transition intensity matrix, $Q(\boldsymbol{\theta})$, has distinct eigenvalues $d_1, d_2, ..., d_K$ and eigenvectors, $\boldsymbol{D_1}, \boldsymbol{D_2}, ..., \boldsymbol{D_K}$, which are summarized in matrix A such that $A = (\boldsymbol{D_1}, \boldsymbol{D_2}, \boldsymbol{D_3}, \cdots \boldsymbol{D_K})$, then we can use matrix decomposition to obtain

$$P(t; \boldsymbol{\theta}) = ADA^{-1}, \tag{1.20}$$

where $D = \text{diag}(e^{d_1 t}, e^{d_2 t}, ..., e^{d_K t})$. The first derivatives can then be calculated as:

$$\frac{\partial P(t; \boldsymbol{\theta})}{\partial \theta_r} = AV_r A^{-1}, \tag{1.21}$$

where $r = 1, 2, ..., p$ and

$$V_r = \begin{cases} \frac{g_{jk}^{(r)}\left(e^{d_j t} - e^{d_k t}\right)}{d_j - d_k} & j \neq k \\ g_{jj}^{(r)} t e^{d_j t} & j = k \end{cases}, \tag{1.22}$$

where $j, k = 1, 2, ..., K$ and $g_{jk}^{(r)}$ is the (j, k) entry in $G^{(r)} = A^{-1}\left[\partial Q(\boldsymbol{\theta})/\partial \theta_r\right] A$. This derivation is given in Jennrich & Bright (1976) and Kalbfleisch & Lawless (1985).

The commonly used Newton-Raphson algorithm requires the second derivatives of the log-likelihood. Here, however, a quasi-Newton procedure is outlined, where the second derivative given in (1.19) is replaced with its expectation, leading to an algorithm which only requires first derivatives. Let $N_j(u_{l-1}) = \sum_{k=1}^{K} n_{jkl}$ be the number of individuals in state j at time u_{l-1}. Since $\sum_{k=1}^{K} \partial^2 p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})/\partial\theta_r \partial\theta_s = 0$, then by first taking the expectation conditional on $N_j(u_{l-1})$, the (r, s) component of the information matrix is:

$$E\left\{-\frac{\partial^2 l}{\partial\theta_r \partial\theta_s}\right\} = \sum_{l=1}^{m}\sum_{j,k=1}^{K}\left[\frac{E\left(N_j(u_{l-1})\right)}{p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})}\frac{\partial p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})}{\partial\theta_r}\frac{\partial p_{jk}(u_l - u_{l-1}; \boldsymbol{\theta})}{\partial\theta_s}\right]. \quad (1.23)$$

This quantity can be approximated by $M_{rs}(\boldsymbol{\theta})$, which is simply (1.23) with $E\left(N_j(u_{l-1})\right)$ replaced by $N_j(u_{l-1})$. These estimates are summarized in matrix $M(\boldsymbol{\theta}) = [M_{rs}(\boldsymbol{\theta})]_{(p\times p)}$. Then, the quasi-Newton procedure proceeds in the following way:

- Begin with initial values $\boldsymbol{\theta}_0$,
- Obtain an updated estimate by $\boldsymbol{\theta}^{(r)} = \boldsymbol{\theta}^{(r-1)} + M(\boldsymbol{\theta}^{(r-1)})^{-1} S(\boldsymbol{\theta}^{(r-1)})$,
- Repeat until convergence is reached.

Computation of these derivatives is facilitated by (1.15), (1.20) and (1.21). A good initial estimate, $\boldsymbol{\theta}_0$ results in convergence to the maximum likelihood estimate, $\hat{\boldsymbol{\theta}}$ such that $M(\hat{\boldsymbol{\theta}})^{-1}$ is an estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ and if $\boldsymbol{\theta}$ is an interior point of the parameter space, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ will have a multivariate Normal limiting distribution as $n \to \infty$ (Kalbfleisch & Lawless 1985).

This method can also accommodate different observation times for each individual. However, the amount of computing time increases linearly with the number of distinct time intervals in the sample (Kalbfleisch & Lawless 1985). The above discussion was based on all subjects entering at the beginning of the study and remaining under observation until the end. However, this method is appropriate when people enter and leave at different times, as long as their event time distribution does not differ from the other subjects'. This method also works for some simple non-homogeneous cases as outlined in Kalbfleisch & Lawless (1985). It is possible to incorporate covariates in the model. However, if interest

lies in continuous covariates, discrete covariates with many levels or simultaneous consideration of a large number of covariates, this method will require a great deal of computation and therefore will be very difficult to implement (Kalbfleisch & Lawless 1989). In this case, covariate values may have to be grouped to apply this method (Kalbfleisch & Lawless 1985).

When considering progressive models, such as that introduced in Figure 1.5, one can then take advantage of the simplified state-structure to construct a likelihood function. Fortunately, a closed form for the transition probabilities is available under a progressive, time-homogeneous Markov model (Satten 1999). Under a $K+1$ state model, with the intensity of moving from state k to state $k+1$ denoted as δ_k and considering the case where there are no covariates, the probability of being in stage k_2 conditional on being in stage k_1 at time zero can be written as

$$P_{k_1,k_2}(t) = \begin{cases} \sum_{k=k_1}^{k_2} C_{k_1,k,k_2} e^{-\delta_k t}, & k_1 \leq k_2 \\ 0, & k_1 > k_2, \end{cases} \tag{1.24}$$

with the coefficients C_{k_1,k,k_2} given by

$$C_{k_1,k,k_2} = \frac{\prod_{l=k_1}^{k_2-1} \delta_l}{\prod_{\substack{l=k_1 \\ l\neq k}}^{k_2} (\delta_l - \delta_k)}, \quad k_1 \leq k \leq k_2,$$

where $C_{k,k,k} = 1$. Using the notation introduced previously for the observation times and the states occupied at these times, under time-homogeneity the likelihood function would be

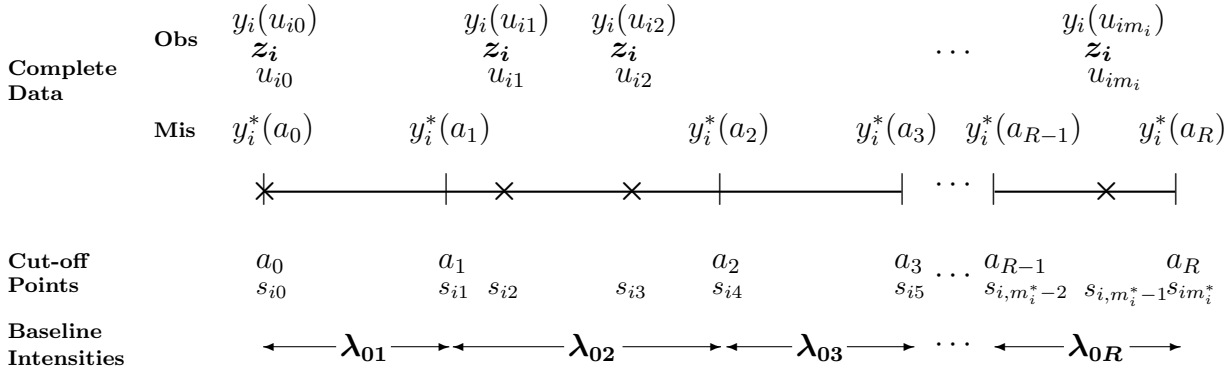
$$\mathcal{L}(\boldsymbol{\delta}\,) = \prod_{j=1}^{m_i-1} P_{y_i(u_{ij}),y_i(u_{i,j+1})}(\Delta u_{ij}),$$

where $\Delta u_{ij} = u_{i,j+1} - u_{ij}$.

This likelihood function cannot be directly applied in the case of the model given in (1.13) with baseline transition intensities specified in (1.14) since the transition intensities are not time homogeneous; they are assumed to be piecewise constant over time.

However, due to the fact that they are constant over certain time intervals (i.e. the baseline intensities, $\boldsymbol{\lambda_{0r}}$, are constant in $[a_{r-1}, a_r)$), we can use (1.24) to construct the likelihood function for this problem. As is evident from Figure 1.6, individuals can undergo a wide range of observation patterns. The observation times generally do not fall on the cut-off points (a_r, $r = 1, ..., R$), so we cannot set up the likelihood assuming constant intensities between visits. However, it is possible to build a complete data likelihood assuming the states occupied by the individuals at the cut-off points, a_r, $r = 1, 2, ..., R$, are observed in addition to those at the assessment times. To do this, we must introduce additional notation. Let $y_i^*(a_r)$ be the (unobserved) state occupied by subject i at time a_r, where $r = 1, 2, ..., R$. Therefore the complete data for subject i would consist of $\{y_i(u_{ij}), y_i^*(a_r); j = 0, 1, 2, ..., m_i, r = 1, 2, ..., R\}$. In addition, let $S_i = \left\{s_{i,0}, s_{i,1}, ...s_{i,m_i^*}\right\}$ represent the set of ordered u_{ij}'s and a_r's for subject i so assuming no observation time u_{ij} is chosen as a cut-off point a_r, $m_i^* = m_i + R$. Figure 1.6 has been modified to incorporate this new notation in Figure 1.7. In addition to the above notation, let $\boldsymbol{\theta}$ represent the set

Figure 1.7: *An illustration of the observation process (including both observed (obs) and unobserved or missing (mis) data) and the underlying baseline intensities in effect over time for an arbitrary subject, i.*



of all unknown parameters in the model (these could include covariate effects as well as the unknown baseline intensities) and define

$$\mathcal{Y}_i(s_{ij^*}) \doteq \begin{cases} y_i(u_{ij}), & \text{if } s_{ij^*} = u_{ij} \\ y_i^*(a_r), & \text{if } s_{ij^*} = a_r \end{cases}, \tag{1.25}$$

for $i = 1, 2, ..., n, j = 0, 1, ..., m_i, j^* = 0, 1, 2, ..., m_i^*$ and $r = 1, 2, ..., R$. The complete data likelihood can then be expressed in the following way:

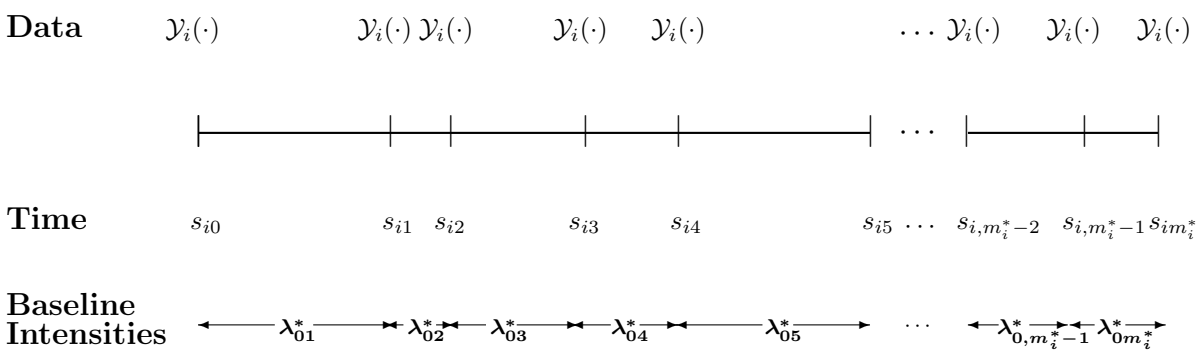
$$\begin{aligned}
\mathcal{L}_{complete}(\boldsymbol{\theta}\ ) &= P\left\{y_i(u_{ij}), y_i^*(a_r) | y_i(0) = 1; i = 1, 2, ..., n, j = 0, 1, ..., m_i, r = 1, 2, ..., R\right\} \\
&= P\left\{\mathcal{Y}_i(s_{ij^*}) | \mathcal{Y}_i(0) = 1; i = 1, 2, ..., n, j^* = 0, 1, 2, ..., m_i^*\right\} \\
&= \prod_{i=1}^{n} \prod_{j^*=1}^{m_i^*-1} P_{\mathcal{Y}_i(s_{ij^*}), \mathcal{Y}_i(s_{i,j^*+1})}(s_{i,j^*+1} - s_{ij^*}).
\end{aligned}$$

On each interval, $[s_{ij^*}, s_{i,j^*+1})$ for $i = 1, 2, ..., n$ and $j^* = 0, 1, 2, ..., m_i^*$, the set of baseline transition intensities is constant at $\boldsymbol{\lambda^*_{0j^*}} = \left(\lambda^*_{01j^*}, \lambda^*_{02j^*}, \cdots, \lambda^*_{0Kj^*}\right)'$, with values equal to one of the R pieces, $\boldsymbol{\lambda_{0r}}$, $r = 1, 2, ..., R$. Specifically,

$$\boldsymbol{\lambda^*_{0j^*}} = \boldsymbol{\lambda_{0r}} \text{ , when } [s_{ij^*}, s_{i,j^*+1}) \subseteq [a_{r-1}, a_r).$$

When the covariates included in the model are fixed or it is assumed that the covariate remains constant between visits, we have essentially divided up the problem into R smaller ones, each based on a time-homogeneous Markov model. This revised observation timeline is displayed in Figure 1.8.

Figure 1.8: *An illustration of the complete observation process and the underlying baseline intensities in effect over time for an arbitrary subject, i.*



The complete data likelihood, conditional on all subjects being in state 1 at time 0, can be expressed in a closed form using (1.24) and the model for the intensities given in (1.13) and (1.14):

$$\mathcal{L}_{complete}(\boldsymbol{\theta}\ ) = \prod_{i=1}^{n} \prod_{j^*=0}^{m_i^*-1} \left\{ \sum_{l=\mathcal{Y}_i(s_{ij^*})}^{\mathcal{Y}_i(s_{i,j^*+1})} C_{\mathcal{Y}_i(s_{ij^*}),l,\mathcal{Y}_i(s_{i,j^*+1})} \exp\left\{ \left[\lambda^*_{0,l,j^*} \exp(\boldsymbol{\beta_l}' z_i) \right] (s_{i,j^*+1} - s_{ij^*}) \right\} \right\}, \tag{1.26}$$

with the coefficients, $C_{\mathcal{Y}_i(s_{ij^*}),l,\mathcal{Y}_i(s_{i,j^*+1})}$ equal to

$$C_{\mathcal{Y}_i(s_{ij^*}),l,\mathcal{Y}_i(s_{i,j^*+1})} = \frac{\prod_{h=\mathcal{Y}_i(s_{ij^*})}^{\mathcal{Y}_i(s_{i,j^*+1})-1} \lambda^*_{0,h,j^*} \exp(\boldsymbol{\beta_h}' z_i)}{\prod_{\substack{h=\mathcal{Y}_i(s_{ij^*}) \\ h\neq l}}^{\mathcal{Y}_i(s_{i,j^*+1})} \left[\lambda^*_{0,h,j^*} \exp(\boldsymbol{\beta_h}' z_i) - \lambda^*_{0,l,j^*} \exp(\boldsymbol{\beta_l}' z_i) \right]},$$

for $\mathcal{Y}_i(s_{ij^*}) \leq l \leq \mathcal{Y}_i(s_{i,j^*+1})$. Since this likelihood involves missing data (i.e. the states occupied at the cut-off points), a natural way to proceed with the maximization is via the EM Algorithm.

One can obtain maximum likelihood estimates by way of the EM algorithm by using an iterative two step approach (Dempster et al. 1977). After selecting reasonable initial values for the parameters of interest, $\hat{\boldsymbol{\theta}}^{(\mathbf{0})}$, and letting *obs* indicate the observed data and *mis* indicate the missing data, the algorithm proceeds in the following manner:

1. *Expectation Step (E-Step)*
Define $Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\boldsymbol{r-1})}\ \right) \doteq E_{mis|obs}\left\{ \log\left[\mathcal{L}_{complete}(\hat{\boldsymbol{\theta}}\ ) \right]; \hat{\boldsymbol{\theta}}^{(\boldsymbol{r-1})}\ \right\}$.

2. *Maximization Step (M-Step)*
Obtain $\hat{\boldsymbol{\theta}}^{(\boldsymbol{r})}$ through maximization of $Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\boldsymbol{r-1})}\right)$ with respect to $\boldsymbol{\theta}$ for $r = 1, 2, ...$

Steps 1 and 2 are repeated until convergence is reached.

In order to use this method to obtain maximum likelihood estimates from (1.26), we can express the complete data likelihood in an equivalent, yet more convenient manner based

on additional notation. Let $\boldsymbol{V_i} = (V_{i,1}, V_{i,2}, ..., V_{i,m_i^*})'$ be a m_i^*- dimensional random vector representing the state path for subject i. At the same time, let $\boldsymbol{v_i} = (v_{i,1}, v_{i,2}, ..., v_{i,m_i^*})'$ be an observed state path. Since not all m_i^* states are observed, let $\mathcal{P}_i$ be the set of all possible paths for subject i. All values in positions of $\boldsymbol{v_i} \in \mathcal{P}_i$ corresponding to the observed states will be equal to the actual observed states while other positions can be any state greater than or equal to the last observed state and less than or equal to the next observed state in this progressive model. Finally, we can express the complete data likelihood, conditional on the initial state occupied, as follows:

$$\mathcal{L}_{complete}(\boldsymbol{\theta}\ ) = \prod_{i=1}^{n} \left\{ \prod_{\boldsymbol{v_i}\ \in \mathcal{P}_i} \left[\prod_{j^*=1}^{m_i^*} P\left(\mathcal{Y}_i(s_{ij^*}) = v_{ij^*} | \mathcal{Y}_i(s_{i,j^*-1}) = v_{i,j^*-1}\right) \right]^{I(\boldsymbol{V_i}\ = \boldsymbol{v_i}\ )} \right\}, \tag{1.27}$$

where $I(\cdot)$ is an indicator function. It then follows that the log-likelihood is:

$$l_{complete}(\boldsymbol{\theta}\ ) = \sum_{i=1}^{n} \left\{ \sum_{\boldsymbol{v_i}\ \in \mathcal{P}_i} \left[I(\boldsymbol{V_i}\ = \boldsymbol{v_i}\ ) \sum_{j^*=1}^{m_i^*} \log\left(P\left(\mathcal{Y}_i(s_{ij^*}) = v_{ij^*} | \mathcal{Y}_i(s_{i,j^*-1}) = v_{i,j^*-1}\right)\right) \right] \right\}. \tag{1.28}$$

Now, the only random quantity in this expression consists of the n indicator variables given by $I(\boldsymbol{V_i}\ = \boldsymbol{v_i}\ )$, $i = 1, 2, ..., n$. It follows then, that the E-Step first involves finding the expectation of these indicators with respect to their distribution, given what was observed, and based on the estimate of $\boldsymbol{\theta}$ from the previous iteration. Since (1.28) is linear in the indicators, $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta^{(r-1)}}\right)$ is obtained by replacing $I(\boldsymbol{V}_i = \boldsymbol{v}_i)$, $i = 1, 2, ..., n$ with their corresponding expectations in (1.28).

$$
\begin{aligned}
&E[I(\boldsymbol{V_i} = \boldsymbol{v_i})|y_i(u_{ij}), \boldsymbol{z_i}; \hat{\theta}^{(r-1)}] \\
&= P\left[\boldsymbol{V}_i = \boldsymbol{v}_i|y_i(u_{ij}), \boldsymbol{z_i}; \hat{\theta}^{(r-1)}\right] \\
&= \frac{P\left[\boldsymbol{V}_i=\boldsymbol{v}_i;\hat{\theta}^{(r-1)}\right]}{P[\boldsymbol{V}_i\in\mathcal{P}_i;\hat{\theta}^{(r-1)}]} \\
&= \frac{\prod_{j^*=1}^{m_i^*} P[\mathcal{Y}_i(s_{ij^*})=v_{ij^*}|\mathcal{Y}_i(s_{i,j^*-1})=v_{i,j^*-1};\hat{\theta}^{(r-1)}]}{\sum_{\boldsymbol{v_i}\in\mathcal{P}_i}\left\{\prod_{j^*=1}^{m_i^*} P[\mathcal{Y}_i(s_{ij^*})=v_{ij^*}|\mathcal{Y}_i(s_{i,j^*-1})=v_{i,j^*-1};\hat{\theta}^{(r-1)}]\right\}}
\end{aligned}
$$

Therefore, for this problem the EM Algorithm is as follows:

1. *Expectation Step (E-Step)*

$$
Q\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}^{(\boldsymbol{r-1})}\right) \doteq
$$

$$
\sum_{i=1}^{n}\left\{\sum_{\boldsymbol{v_i}\in\mathcal{P}_i}\left[\frac{\prod_{j^*=1}^{m_i^*} P[\mathcal{Y}_i(s_{ij^*})=v_{ij^*}|\mathcal{Y}_i(s_{i,j^*-1})=v_{i,j^*-1};\hat{\theta}^{(r-1)}]}{\sum_{\boldsymbol{v_i}\in\mathcal{P}_i}\left\{\prod_{j^*=1}^{m_i^*} P[\mathcal{Y}_i(s_{ij^*})=v_{ij^*}|\mathcal{Y}_i(s_{i,j^*-1})=v_{i,j^*-1};\hat{\theta}^{(r-1)}]\right\}}\sum_{j^*=1}^{m_i^*}\log\left[P(\mathcal{Y}_i(s_{ij^*})=v_{ij^*}|\mathcal{Y}_i(s_{i,j^*-1})=v_{i,j^*-1})\right]\right]\right\}.
$$

2. *Maximization Step (M-Step)*
Obtain $\hat{\boldsymbol{\theta}}^{(\boldsymbol{r})}$ through maximization of $Q\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}^{(\boldsymbol{r-1})}\right)$ with respect to $\boldsymbol{\theta}$ for $r = 1, 2, ...$

Steps 1 and 2 are repeated until convergence is reached (i.e. when the difference between successive estimates drops below a specified tolerance).

1.2 Mismeasured Covariates

Data collected in health research frequently involve measurement error in covariates. Study designs can involve either retrospective data collection or prospective data collection. In the former, it is often difficult, if not impossible to determine past exposure levels to a

potential toxin or to accurately determine covariate values which rely on a subject's recall. In prospective studies, it may be difficult to collect accurate covariate information due to practical considerations and cost (Yi & Cook 2005). Sometimes investigators must settle for an imperfect measurement because it is impossible to measure the true value. In other situations it may be possible to obtain a better measurement of the covariate but it is more costly so a less accurate measurement is collected. When the covariates subject to mismeasurement are discrete, they are referred to as *misclassified*; whereas if they are continuous, we are dealing with *measurement error*. Generally, naive estimation approaches which ignore the presence of either result in biased estimates for the parameters of interest. Therefore, it is important that the presence of mismeasured covariates be recognized and accounted for in estimation. Considerable research has been devoted to addressing this issue and accounting for this error. A detailed description of the methods available are described in Fuller (1987) for linear regression models and in Carroll et al. (2006) for nonlinear models. In the following sections, the general effects of mismeasured covariates will be discussed and available methods to address mismeasurement will be briefly described.

1.2.1 General Effects of Mismeasured Covariates

For the purposes of this discussion, let

- Y be a response variable,
- $\boldsymbol{X}$ be a vector of covariates subject to error (true values unknown),
- $\boldsymbol{W}$ be the mismeasured version of $\boldsymbol{X}$, and
- $\boldsymbol{Z}$ be a vector of covariates measured without error.

Suppose, the distribution of the response Y given $(\boldsymbol{X}, \boldsymbol{Z})$ is specified up to unknown parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_X, \boldsymbol{\beta}_Z)$ by a model given by $m(Y, \boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})$. In addition, suppose that the dependence of Y on $(\boldsymbol{X}, \boldsymbol{Z})$ is characterized by the linear predictor, $\boldsymbol{\beta_X}'\boldsymbol{X} + \boldsymbol{\beta_Z}'\boldsymbol{Z}$. Direct use of $\boldsymbol{W}$ in place of $\boldsymbol{X}$ results in biased estimates for $\boldsymbol{\beta_X}$ and can even affect estimation of $\boldsymbol{\beta_Z}$, the parameters associated with the correctly measured covariates (Yi & Cook

2005). The simple linear regression model has been used quite extensively in literature to demonstrate the effect of a mismeasured covariate on estimation of the parameter of interest. It is well known that under this model and assuming the *Classical Error Model*, which will be introduced shortly, a mismeasured covariate results in an estimate of the slope parameter which is biased toward the null. This phenomenon is referred to as attenuation. In addition, the standard error of this estimator based on a naive analysis is often an underestimate of the true standard error (Fuller 1987) and there will be a loss of power to detect significant covariate effects and relationships among the variables (Carroll et al. 2006).

The situation gets much more complicated for more complex regression models. For instance, even in the case of multiple linear regression, any relationship that exists between a covariate measured with error and others measured with or without error can induce bias in the parameter estimators. In fact, there is a tendency for covariate effects based on mismeasured values to be shifted toward those measured with less error (Reeves & Cox 1998). When there are covariates measured with error as well as those without, the presence of error in some may cause bias in the parameter estimates associated with the error-free covariates. In general, the coefficient estimate for an error-free covariate will be biased unless the covariate is independent of the one measured with error (Carroll 1998; Buonaccorsi et al. 2005). For even more complex models, one may find the true effects masked in the presence of additional covariates, absent effects may appear to be significant and estimates may even appear to exhibit an effect which is opposite to the truth. The latter potential was described in the case of a two group ANCOVA where the treatment groups were defined based on an error-free covariate, Z, and a covariate subject to error, X, was measured on all individuals. This problem was due to the fact that the design was not balanced. That is, the mean of X differed across treatments defined by Z or was dependent on Z (Carroll et al. 1995). In the case of binary regression, the presence of measurement error often results in estimates of relative risk that are biased toward 1 (Raboud 1991; Stefanski & Carroll 1985). However, when the majority of subjects experience extreme risks, either very high or very low, relative risk estimates may be biased away from 1 (Stefanski & Carroll 1985). In general, the effects of mismeasured covariates

depend on the model under consideration and the joint distribution of the error process and the variables (i.e. the response variable and covariate(s) measured with and without error) (Carroll 1998). Clearly, mismeasured covariates can have a large and unpredictable impact on estimation and therefore, they must be accounted for in estimation.

1.2.2 Approaches for Mismeasured Covariates

When dealing with measurement error, one must first consider the error distribution or the relationship of $\boldsymbol{W}$ to the unobserved $\boldsymbol{X}$. There are three different approaches that can be taken with respect to the measurement error distribution: parametric, semi-parametric and nonparametric. A nonparametric approach was taken by Pepe & Fleming (1991) when they empirically estimated the likelihood in the presence of mismeasured covariates. Huang & Wang (2000) also took a nonparametric approach to deal with mismeasured covariates in the Cox model with replicate data available. Tsiatis & Davidian (2001) and Kulich & Lin (2000) considered semi-parametric approaches of dealing with mismeasured covariates in survival analysis.

For a parametric approach dealing with continuous covariates, two types of additive error models have been developed that have quite different interpretations. Considering first the simplest of the two, and letting $\boldsymbol{X} = X$, $\boldsymbol{W} = W$ and $\boldsymbol{Z} = Z$ represent scalars rather than vectors for this formulation, the *classical error model* can be expressed as:

$$W = X + U, \tag{1.29}$$

where U is independent of X. Often in practice the random error component U is assumed to have a normal distribution with mean 0, which means the measurement error is unbiased, and variance, σ_U^2. This model is appropriate in situations where an attempt is made to measure X directly, but the measurement is subject to error (Carroll et al. 1995). For instance, this model would be reasonable in the case of an observational study in which the covariates naturally vary from subject to subject and there is no manipulation of the covariate values by the investigator (i.e. an uncontrolled study) (Raboud 1991). Sources of error may include the measurement device and method, the data entry process and even time of day or seasonal variations. In contrast, the *Berkson error model* is appropriate in

controlled studies where the outcome of interest is measured at given levels of the covariate (Carroll et al. 2006). It is of the form:

$$X = W + U, \tag{1.30}$$

where, for a given individual, W is viewed as fixed and U, the measurement error, is viewed as random. This model would be reasonable in a laboratory study in which it was intended to expose subjects to certain fixed levels of a suspected risk factor. There may be error about the intended nominal level of exposure, W.

In the case of discrete covariates, the measurement error process is specified through misclassification probabilities. For dichotomous covariates, taking on $0-1$ values, there are two such probabilities:

- $P(W = 0|X = 1) = 1 - P(W = 1|X = 1) = \pi_{01}$ (1-Sensitivity), and
- $P(W = 1|X = 0) = 1 - P(W = 0|X = 0) = \pi_{10}$ (1-Specificity).

Supplementary information regarding the measurement error or misclassification distributions can be obtained from data either internal or external to the study, but related to the investigation at hand. They can be in the form of validation studies where X is observed directly for some subjects, replication studies where replicate measurements of X (i.e. W) are available, providing information regarding variability in the error process, or instrumental data where information on another variable T is observed in addition to W (Carroll et al. 1995). Another aspect of the measurement error process must be considered at this point. We refer to the error as nondifferential when W provides no information about Y in addition to that provided by (X, Z). Another way of expressing this is to state that the distribution of Y given (X, Z, W) depends only on (X, Z). Nondifferential measurement error is much more straightforward to deal with, as will be clear shortly.

There are two fundamentally different interpretations of the unobserved true values of the covariates, $\boldsymbol{X}$. In functional modeling, the $\boldsymbol{X}$'s are considered as a sequence of fixed unknown vectors, whereas in structural modeling the $\boldsymbol{X}$'s are regarded as random and a model for their joint distribution is assumed. In Carroll et al. (1995), the definition of

functional models is extended to include those for random $\boldsymbol{X}$'s where minimal assumptions about their distributional form are made in addition to fixed $\boldsymbol{X}$. Both modeling approaches will be outlined below.

Structural Modeling

As stated previously, structural modeling views the unobserved true covariate $\boldsymbol{X}$ as random so distributional assumptions are required when using this approach. Likelihood and Bayesian methods fall into this category. Likelihood methods are useful in many situations, including those with misclassified covariates. As is the case with maximum likelihood, the resulting estimators will exhibit the favorable asymptotic properties of consistency and efficiency. It is possible to develop expanded likelihood expressions incorporating an assumed specific form for the measurement error distribution. Often it is required that the distribution of the true covariate $\boldsymbol{X}$ to be known (or assumed) in these formulations (Nakamura 1990). However, Aitkin & Rocci (2002) describe a maximum likelihood approach for generalized linear models and general error models using an EM algorithm. The distribution of X is approximated by a discrete distribution of a finite number of mass-points determined as part of the model. Even though maximum likelihood estimators have such favorable characteristics, the complicated form of the likelihood function, especially in the case of continuous covariates, often makes implementation difficult and computationally intensive. In addition, likelihood methods exhibit a lack of robustness in general to misspecification of the model (Reeves & Cox 1998). However, in some problems dealing with measurement error, they may be the more flexible, efficient or reliable method (Schafer & Purdy 1996).

Likelihood function formulation is very much problem specific (Stefanski & Carroll 1985). As an illustration of this, we will consider the likelihood function for three different situations (Carroll et al. 1995). For the sake of these formulations, we consider continuous covariates, but similar expressions hold for discrete covariates with the integrals replaced by sums.

Case I: The first is the case where the true value of $\boldsymbol{X}$ is unobserved and there are no

validation data accessible to characterize the conditional distribution of $\boldsymbol{W}$ given $\boldsymbol{X}$ and $\boldsymbol{Z}$. Say we are interested in estimating parameters $\boldsymbol{\beta}$ in the model $f_{Y|X,Z}(y|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\beta})$. Since $\boldsymbol{X}$ is measured with error, a model based on the observed data is $f_{Y|W,Z}(y|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters including $\boldsymbol{\beta}$, the parameters of interest. If we are dealing with nondifferential error, then for a particular subject the likelihood function could be expressed as: $\mathcal{L}(\theta) = f_{Y,W|Z}(y,\boldsymbol{w}|\boldsymbol{z};\boldsymbol{\theta})$, where

$$\begin{aligned} f_{Y,W|Z}(y,\boldsymbol{w}|\boldsymbol{z};\boldsymbol{\theta}) &= \int f_{Y,W,X|Z}(y,\boldsymbol{w},\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\theta})d\boldsymbol{x} \\ &= \int f_{Y|X,Z}(y|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\beta})f_{W|X,Z}(\boldsymbol{w}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\delta})f_{X|Z}(\boldsymbol{x}|\boldsymbol{z};\boldsymbol{\lambda})d\boldsymbol{x} \end{aligned}$$

Here, $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are assumed to be known (Reeves & Cox 1998).

Case II: Consider now the situation where $\boldsymbol{X}$ is unobserved and the Berkson error model is appropriate. To obtain an expression for the likelihood in this case for a particular subject, we condition on $\boldsymbol{W}$ to obtain:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= f_{Y|W,Z}(y|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}) \\ &= \int f_{Y|X,Z}(y|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\beta})f_{X|W,Z}(\boldsymbol{x}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\delta})d\boldsymbol{x}. \end{aligned}$$

Appropriate supplementary data can be used to estimate $\boldsymbol{\delta}$ and then the likelihood function will just be in terms of the unknown parameter(s) of interest, $\boldsymbol{\beta}$.

Case III: Finally, consider the case where there is a validation study comprised of subjects for whom in addition to $\boldsymbol{W}$, $\boldsymbol{X}$ is observed. In other words, there are internal validation data available. Analogous to missing data problems, it is critical here that the probability $\boldsymbol{X}$ is measured for a particular subject can depend on $(Y,\boldsymbol{Z},\boldsymbol{W})$, but not $\boldsymbol{X}$ itself (Carroll et al. 1995). To formulate this likelihood function first let

$$\Delta_i = \begin{cases} 1, & \text{if subject } i \text{ is selected for validation study} \\ 0, & \text{otherwise} \end{cases} . \tag{1.31}$$

A likelihood function based on all observed data would have the following form:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\left\{\left[\int f_{Y|X,Z}(y_i|\boldsymbol{x},\boldsymbol{z_i};\boldsymbol{\beta}) f_{W|X,Z}(\boldsymbol{w_i}|\boldsymbol{x},\boldsymbol{z_i};\boldsymbol{\delta}) f_{X|Z}(\boldsymbol{x}|\boldsymbol{w_i};\boldsymbol{\lambda}) d\boldsymbol{x}\right]^{1-\Delta_i}\right. \\ &\quad \left.\cdot \left[f_{Y|X,Z}(y_i|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\beta}) f_{W|X,Z}(\boldsymbol{w_i}|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\delta})\right]^{\Delta_i}\right\}.\end{aligned}$$

There are many more situations that could arise in practice. Regardless of how complex the situation becomes, in the maximum likelihood approach incorporating measurement error, the objective is to express $f_{Y,W|Z}$ in terms of the "true" model, $f_{Y|X,Z}$.

Pan et al. (2006) took a structural modeling approach based on maximum likelihood for a general linear mixed model for a continuous response and for a linear logistic mixed model for a binary response. In their models, the response was permitted to depend on the response at the previous assessment time as is the value of the true covariate, X. They assumed that the classical error model given in (1.29) was appropriate and that the error variance, σ_U^2, was known. They investigated naive models that correctly specified the structure of the response model, but misspecified the structure of the covariate effect model. Asymptotic biases based on the naive model were investigated and a maximum likelihood approach, incorporating measurement error in a continuous covariate was implemented using an EM algorithm. When direct implementation of the likelihood approach is computationally intensive, one may instead base inference on a simpler, approximate function called the pseudo-likelihood function (Yi & Cook 2005).

Bayesian methods assume both the variables and the parameters are random and follow probability distributions. The first step in this approach involves determining the joint probability density function of the data and parameters. From this, the posterior density, or the conditional probability distribution of the parameters given the data, can be obtained based on Bayes Rule. Inference can then be conducted based on this distribution. Computation involving this distribution usually requires high-dimensional numerical integration (Carroll 1998). Therefore, to perform calculations using this distribution often Markov Chain Monte Carlo (MCMC) algorithms such as the Gibbs' Sampler are used. Gustafson (2004) provides a thorough description of the effects of covariate measurement error and misclassification and presents the Bayesian approach of addressing the problem.

Clearly, both likelihood and Bayesian methods require strong distributional assumptions. To relax some of these assumptions, one could instead adopt a functional modeling approach such as those which will be discussed in the next subsection.

Functional Modeling

Functional modeling involves few or no assumptions regarding the distribution of the unknown covariate $\boldsymbol{X}$. For this reason, much of the literature has tended to concentrate on this approach. For general nondifferential error problems, two simple, approximate methods to deal with mismeasured covariates include *regression calibration* and *simulation extrapolation (SIMEX).*

Regression calibration was first suggested by Prentice (1982) for use in survival analysis, specifically for the proportional hazards model. It involves approximating the unknown value of $\boldsymbol{X}$ by the regression of $\boldsymbol{X}$ on $(\boldsymbol{W}, \boldsymbol{Z})$. Ideally, information on the joint behavior of $(\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Z})$ can be obtained through validation data. If validation data are unavailable, then information regarding the value of $\boldsymbol{X}$ can be extracted from replication or instrumental variable data (Carroll et al. 1995). The algorithm proceeds as follows:

- using additional data, whether it be replication, validation or instrumental data, obtain the *calibration function* by estimating the regression of $\boldsymbol{X}$ on $(\boldsymbol{W}, \boldsymbol{Z})$,
- replace $\boldsymbol{X}$ by its approximation from the calibration function and proceed with analysis as if $\boldsymbol{X}$ were measured correctly, and
- adjust the naive standard errors using resampling or asymptotic methods (Carroll 1998).

Carroll et al. (2006) describe the algorithm in detail, giving extensions to the model and providing examples.

Simulation Extrapolation (SIMEX) was first proposed in 1994 by Cook and Stefanski. This procedure is based on the key idea that the effect of measurement error can be investigated and therefore adjusted for using simulation techniques (Carroll et al. 1995).

Estimates are obtained by first inducing bias in parameter estimates by adding additional measurement error using resampling methods, establishing a trend in this induced bias as a function of the error variance and extrapolating back to the case of no measurement error. This method is suitable for use for additive or multiplicative measurement error models and if the model is correctly specified, it will result in improved parameter estimates (Carroll et al. 1995). Implementation of regression calibration and SIMEX is relatively straightforward. However, except in the cases of linear and log-linear models, estimators obtained using these methods are only approximately consistent in general.

Considerable literature involves the use of estimating equations to address the problem of mismeasured covariates. Unbiased estimating equations are often based on fewer distributional assumptions than required for the structural approaches, and computation is generally more straightforward (Yi & Cook 2005). There are three types of estimating equation approaches to deal with the mismeasured covariate problem: *conditional score equations*, *corrected-score equations* and *general unbiased estimating equations* (Carroll et al. 1995). Conditional score equations are derived by conditioning on sufficient statistics. The objective is to reduce the number of parameters that need to be estimated by removing dependence of the estimating equations on nuisance parameters through conditioning. Carroll et al. (1995) illustrate this procedure for distributions which belong to the exponential family. Relatively straightforward results are available using this method for models that belong to this family, although the solution may involve extensive numerical integration or summation (Carroll et al. 1995).

The corrected-score equation method is not restricted to models belonging to the exponential family and is in fact, applicable for most generalized linear models. This method was proposed by Nakamura (1990). In this paper, he described the corrected score function as "one whose expectation with respect to the measurement error distribution coincides with the usual score function based on the unknown true independent variables". Let $U(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, Y)$ denote the score function when all covariates are measured precisely. Now suppose $\boldsymbol{X}$ is mismeasured as $\boldsymbol{W}$. Then the naive score equation is given by $U(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{Z}, Y)$. Use of this naive score equation can result in inconsistent estimates

of $\boldsymbol{\theta}$ since $E\left[U(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{Z}, Y)\right] \neq \mathbf{0}$. Therefore this naive score function should be adjusted to provide the correct estimates. To accomplish this, one must first find an adjusted log-likelihood function, l^*, and provided it is twice differentiable, a corrected-score function, $U^* = \partial l^*/\partial \boldsymbol{\theta}$, and a corrected observed information function, $I^* = \partial U^*/\partial \boldsymbol{\theta}$. These will all be functions of Y, $\boldsymbol{W}$ and $\boldsymbol{Z}$, but not $\boldsymbol{X}$, and provided E^* and $\partial/\partial\boldsymbol{\theta}$ are interchangeable, will satisfy $E^*\left[l^*(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{Z}, Y)\right] = l(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, Y)$, for all $\boldsymbol{\theta}$, where the expectation, E^*, is taken with respect to the distribution of $\boldsymbol{W}|\boldsymbol{X}, Y, \boldsymbol{Z}$. From this, it follows that $E^*\left[U^*(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{Z}, Y)\right] = U(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, Y)$, and $E^*\left[I^*(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{Z}, Y)\right] = I(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, Y)$. Then, the value of $\boldsymbol{\theta}$ which satisfies $U^*(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{Z}, Y) = \mathbf{0}$ with $I^*(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{Z}, Y)$ positive definite is an estimate of $\boldsymbol{\theta}$ accounting for the mismeasured covariate(s). The estimator obtained using this method is asymptotically unbiased (Nakamura 1990). Estimates are usually obtained through numerical iteration using the naive maximum likelihood estimates as initial values (Nakamura 1990). The main disadvantage of this approach is that it is often difficult to determine the appropriate adjusted log-likelihood function, l^*.

Finally, the method of general unbiased estimating equations was described by Robins et al. (1994). This approach is suitable for situations where there are validation data available for a subset of the subjects in the study. The goal of this approach is to incorporate additional information into the analysis without making assumptions regarding the joint distribution of $(\boldsymbol{X}, \boldsymbol{W})$ given $\boldsymbol{Z}$. As is evident from above, there are many possible approaches to choose from when faced with a mismeasured covariate problem, each with advantages and limitations. For the purposes of this research, we will concentrate on the maximum likelihood approach and will also implement the SIMEX method approach for comparison purposes.

Problems with mismeasured covariates involve incomplete data in the sense that the true values of $\boldsymbol{X}$ are not measured. Instead, an error prone version of $\boldsymbol{X}$, $\boldsymbol{W}$, is measured. In the next section, another form of incomplete data will be introduced; *current status data*.

1.3 Current Status Data

An extreme case of interval censoring is current status data. Current status data arise when individual i is examined only once at inspection time $b_i > 0$ so that the event of interest is known to occur in either $(0, b_i]$ or $(b_i, \infty]$. This type of data can arise if the method of observation is destructive or costly. For example, in animal carcinogenicity experiments, animals are sacrificed to obtain information on tumors through autopsy (Lawless 2003). Let b_i be the observation time for subject i and $\delta_i = I(t_i \leq b_i)$. The data for subject i would then be (b_i, δ_i) and the likelihood function for current status data would be:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[F_T(b_i; \boldsymbol{\theta})\right]^{\delta_i} \left[\mathcal{F}_T(b_i; \boldsymbol{\theta})\right]^{1-\delta_i} \tag{1.32}$$

The function that will actually be maximized with respect to the unknown parameters is the log-likelihood function which is given by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{\delta_i \log\left[F_T(b_i; \boldsymbol{\theta})\right] + (1 - \delta_i) \log\left[\mathcal{F}_T(b_i; \boldsymbol{\theta})\right]\right\}. \tag{1.33}$$

There are various approaches one could take in estimation. A parametric approach would involve adopting a parametric form for the distribution of the event time (i.e. assume an exponential or Weibull distribution, for instance). If interest lies in examining covariate effects, one could take either a parametric or semi-parametric approach by characterizing the event time distribution in terms of regression models such as additive hazards models (Shiboski 1998), proportional hazards models or proportional odds models (Jewell & van der Laan 2002). Maximum likelihood techniques could then be applied to make inferences regarding the parameters of interest.

Estimation is simplified through the use of parametric models, where a distributional form is specified. Alternatively, a nonparametric approach avoids parametric assumptions. Let

- m=the number of distinct test or observation times,
- $b_{(j)}$=the j^{th} ordered test time,

- D_j=the set of patients tested at time $b_{(j)}$
- n_j=the number of patients tested at time $b_{(j)}$
- $d_j = \sum_{i \in D_j} \delta_i$

Then, as in Lawless (2003) and Sun (2006), the nonparametric maximum likelihood estimate (*NPMLE*) of F_T is

$$\hat{F_T}(b_{(j)}) = \max_{u \leq j} \min_{v \geq j} \left(\frac{\sum_{l=u}^{v} d_l}{\sum_{l=u}^{v} n_l} \right) \tag{1.34}$$

To implement this, we could use a procedure called *Pool-Adjacent-Violators Algorithm (PAVA)* outlined in Ayer et al. (1955). To proceed with this algorithm we let $p_j^* = d_j/n_j$ for j=1,...,m. Then, the NPMLE of $F_T = 1 - \mathcal{F}_T$ is a step function with up to m jumps and is given by:

$$\text{If } 0 \leq p_1^* \leq \cdots \leq p_m^* \leq 1,\ \hat{F}(t_{(j)}) = p_j^*,\ j = 1, ..., m, \text{ or}$$
$$\text{If } p_k^* > p_{k+1}^* \text{ for some } k = 1, ..., m-1,$$

$$\hat{F}(t_{(k)}) = \frac{d_k + d_{k+1}}{n_k + n_{k+1}},$$
$$\hat{F}(t_{(k+1)}) = \frac{d_k + d_{k+1}}{n_k + n_{k+1}}$$

This algorithm is repeated until a monotone non-decreasing set of ratios is obtained.

1.4 Cure Rate Data

In lifetime data analysis all individuals in the population are assumed to be at risk of experiencing the event of interest and are expected to eventually make the transition between states if they are observed indefinitely (Maller & Zhou 1996). However, there are situations where this may not be the case. Consider a study on the recurrence of cancer in patients who have gone into remission. There are two states in this set-up: cancer-free and recurrence of cancer. Hopefully, most patients will never experience a recurrence. The proportion of immunes or those who will never experience a recurrence, and the effects that

certain covariates (treatments, age at onset, etc.) have on this proportion would be of great interest to investigators. Another example arises in criminology. Consider an investigation of the risk of reoffending for those who have been released from prison. It is reasonable to model the time to the next arrest for these individuals. However, not all ex-convicts will commit another crime and be arrested again (Maller & Zhou 1996). Clearly, there are many situations where it may be appropriate to assume there is an immune component in the population.

In the motivating study for the work in Chapter 4, only about 5% of patients are believed to experience seroconversion, the event of interest. Therefore, it does not make sense to model the time to seroconversion without taking into account that most of the event times will essentially be infinite because the event will not occur for these individuals. Farewell (1977) suggests a way to determine a distribution that allows for immunes in addition to those subject to failure. First, let $X_i \sim BIN(1, \pi)$. X_i is a Bernoulli random variable which represents whether or not individual i will experience the event of interest. When $X_i = 1$, individual i is said to be susceptible or subject to the event of interest and when $X_i = 0$ the individual is immune or will never experience the event of interest. Since we cannot observe the subjects indefinitely, we do not know whether an individual is immune so X_i is unobserved. The individuals who are subject to the event of interest (the susceptibles) have a distribution of the time to the event, T, which is characterized by $F_T(t)$. Assuming this is a proper distribution function, $F_T(0) = 0$ and $\lim_{t\to\infty} F_T(t) = 1$. Those with $X_i = 0$ are considered to have failure times $t_i = \infty$. Therefore, the c.d.f. of T corresponding to this immune group is $G_T(t) = 0$, $0 \leq t < \infty$ since T is degenerate at ∞. Therefore, $F(t)$, the distribution function for the entire population, can be expressed as a mixture of the distributions given by $F_T(t)$ and $G_T(t)$:

$$F(t) = \begin{cases} F_T(t) & \text{with probability } P(X = 1) = \pi \\ 0 & \text{with probability } P(X = 0) = 1 - \pi \end{cases}.$$

for $0 \leq t < \infty$. The event time distribution can be specified parametrically or semi-parametrically and covariate effects on either the event time or the immunity status can be investigated by specifying the appropriate regression model for either the survival distribution or π, respectively.

Cure rate models for right-censored data have received some attention in the literature. Farewell (1982; 1986) took a parametric approach considering a logistic model for the probability of experiencing the event and a Weibull model for the time to event distribution. He notes that nonidentifiability may be an issue because a long-tailed survival curve could mean there is a large cure rate or it could arise simply due to the shape of the true underlying survival curve for the susceptible group. Therefore, under his parametric model it can be difficult to distinguish between the location parameter in the logistic model and the Weibull shape parameter. To relax some of the parametric assumptions under the logistic/Weibull model, Taylor (1995) took a semi-parametric approach by also assuming a logistic model for the probability of experiencing the event but estimating the event time distribution nonparametrically using a Kaplan-Meier estimator. He suggested restricting the survivor function to 0 after a certain time to improve the performance of the estimators under his model. Farewell (1982; 1986) and Taylor (1995) both allowed the incidence to depend on covariates. Other semi-parametric methods have since been investigated that allow the event time distribution rather than the event probability to depend on covariates. Peng & Dear (2000) and Sy & Taylor (2000) proposed logistic/proportional hazards models and used the EM algorithm to obtain maximum likelihood estimates. To investigate the effects of covariates on both the event probability and the time to event, Li & Taylor (2002) assume a logistic/accelerated failure time (AFT) model, fitting covariates to both components. These methods have all been proposed and implemented for right-censored data. Modeling and methodology for cure rate models in the context of current status data will be discussed further in Chapter 4.

1.5 Outline of Thesis

Methodology incorporating mismeasurement for three types of problems involving interval-censored life history data will be investigated in the following chapters. Chapter 2 involves mismeasured covariates with interval-censored survival data. In Chapter 3, this work is extended to include progressive multi-state processes. In both chapters asymptotic biases of naive estimators will be displayed, a naive estimation approach will be compared to two methods incorporating the mismeasurement (i.e. a correct maximum likelihood approach

and SIMEX) via simulation and the methodology will be applied to data arising from a motivating study on Psoriatic Arthritis progression. In Chapter 4, estimation of a cure rate model based on current status data will be explored. The methodology will be applied to orthopedic surgery data. Finally, Chapter 5 will briefly summarize overall findings and outline future work.

Chapter 2

Interval-censored Lifetime Data with Mismeasured Covariates

2.1 Overview

It has been well established that the presence of measurement error or misclassification in covariates affects the properties of estimators for many different types of models. Consider the Cox model with the form (1.6):

$$\lambda(t; \boldsymbol{x}, \boldsymbol{z}) = \lambda_0(t) \exp\left(\boldsymbol{\beta}_x' \boldsymbol{x} + \boldsymbol{\beta}_z' \boldsymbol{z}\right). \tag{2.1}$$

True values of the covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$ are required to use partial likelihood methods to obtain accurate estimates. Prentice (1982) investigated the effect of errors that follow the Berkson model, (1.30), on relative risk estimates under the proportional hazards model. This approach was later extended to the case of classical error models, (1.29). Pepe et al. (1989) give an expression for β^*, the limiting value of the estimator for the regression coefficient based on a naive analysis using the mismeasured version of X in place of its true value under the proportional hazards model and when the classical measurement error model, (1.29), holds. They found β^* has the form:

$$\beta^* = \beta \frac{i(\beta)}{i(\beta) + \sigma_U^2}, \tag{2.2}$$

where $i(\beta) = E\left[-n^{-1}\partial^2 l(\beta; X)/\partial\beta^2\right]$. This result is derived by Raboud (1991) and is similar to the expression of the naive estimator given by Fuller (1987) in the context of simple linear regression with one covariate measured with error under the classical error model (1.29):

$$\beta^* = \beta\frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}.$$

It has also been demonstrated that the magnitude of the bias in coefficients associated with mismeasured covariates may increase if another covariate is included in the model, even when it is error-free. If two or more covariates are mismeasured, then it is difficult to know in which direction the bias will be (Armstrong 1990). Adjustments are therefore appealing when one or more covariates are measured with error.

Likelihood based approaches are useful because they result in consistent estimators whose asymptotic distributions are known. Gong et al. (1990) illustrate how misclassification in discrete covariates can be accounted for based on a likelihood approach using an EM algorithm. Other papers using likelihood approaches are DeGruttola & Tu (1994), Wulfsohn & Tsiatis (1997), Henderson et al. (2000) and Xu & Zeger (2001). Zucker (2005) describes a pseudo-partial likelihood approach where a Breslow-type expression is substituted in place of the baseline cumulative hazard function and the resulting partial log-likelihood function is maximized with respect to all parameters. The measurement error distribution is assumed known or estimated from validation data (internal or external) or from replicate measurements. Disadvantages of the likelihood approach in mismeasured covariate problems are that distributional assumptions for the true unknown covariate and the error must be made and often very complex numerical integration is required. Therefore, researchers have tended to concentrate more on functional methods, relaxing some of the distributional assumptions and easing the computational burden. Semi-parametric likelihood approaches such as those applied in Hu et al. (1998) and Song et al. (2002) relax distributional assumptions on the true underlying covariate. However, these methods can still involve intensive computation (Song & Huang 2005).

Functional methods can be parametric, semi-parametric or nonparametric, depending

on the assumptions made regarding the error distribution. Nakamura (1992) took a parametric approach for which he described an approximate corrected score estimating equation assuming normal errors to obtain estimates of $\boldsymbol{\beta}$ in the proportional hazards models. A similar approach was taken by Buzas (1998). Later, consistency of this estimator under a normal error distribution was demonstrated (Kong & Gu 1999). Huang & Wang (2000) took a nonparametric approach to derive a corrected score estimating equation. In this case, no distributional assumptions were required regarding the true underlying covariate or error distributions but additional information regarding the error was assumed to be available via replication data. Improvements to these parametric and nonparametric corrected score methods were proposed in Song & Huang (2005). These estimators tend to perform better for small sample sizes and large measurement error. Additional data in the form of validation or replication data are required (Song & Huang 2005). Hu & Lin (2002) extended the work in Nakamura (1992) and Huang & Wang (2000) to estimate the baseline cumulative hazard function, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, in addition to $\boldsymbol{\beta}$. A symmetric error distribution is assumed for this method (Hu & Lin 2002). This work, along with that of Huang & Wang (2000) and Xie et al. (2001), have recently been extended to the stratified Cox model, where the baseline hazard is permitted to differ between groups (Gorfine et al. 2004). Yi & Lawless (2007) proposed a method based on the ideas in Nakamura (1990) and a weakly parametric piecewise constant baseline hazard function to estimate all parameters in the model. This method is simpler than others which rely on the partial likelihood function, and as the number of pieces in the baseline hazard increases, the estimator approaches that which results from other more complicated approaches (Yi & Lawless 2007). Augustin (2004) derived an exact corrected log-likelihood function, also based on proportional hazards with piecewise constant baseline hazards and the classical error model. A conditional score estimator was given by Tsiatis & Davidian (2001). The estimators of the corrected and conditional score approaches have been shown to be consistent and equivalent for the case where the errors are normally distributed.

There have been some approximate methods developed to deal with mismeasured covariates in survival analysis as well. As mentioned earlier Prentice (1982) introduced the regression calibration estimator for use in proportional hazards models when the Berkson er-

ror model is employed and validation data are available. In this case $E\{\exp[\boldsymbol{\beta}'_X\boldsymbol{X}+\boldsymbol{\beta}'_Z\boldsymbol{Z}]\}$ is approximated by $\exp\left[\boldsymbol{\beta}'_{X|W,Z}E(\boldsymbol{X}|\boldsymbol{W},\boldsymbol{Z})+\boldsymbol{\beta}'_Z\boldsymbol{Z}\right]$ (Kalbfleisch & Prentice 2002). Later, asymptotic results were developed for the regression calibration estimator in Wang et al. (1997). Xie et al. (2001) extended this method to the setting where the classical error model is used and replicate data are available to estimate parameters of the measurement error model. Although their method introduced some small asymptotic bias, based on simulation studies, their approach was shown to be robust to some misspecification of the true underlying covariate and error distributions when they are symmetric (Gorfine et al. 2004). Zucker & Speigelman (2004) proposed a method to deal with misclassified discrete covariates when there are validation data available. Their method first estimates the survival function to obtain information regarding the parameters of interest. The estimator involves least squares analysis of weighted averages of transformed Kaplan-Meier curves for the different possible values of $\boldsymbol{W}$. Other approximate methods involve estimation of the partial likelihood. Zhou & Pepe (1995) took this approach in the presence of misclassified discrete covariates with a validation sample. Later, Zhou & Wang (2000) extended this to the situation where there was measurement error in continuous covariates. These approximate methods are successful in reducing bias in the estimates for $\boldsymbol{\beta}$ but the estimators may not be consistent in general. Most of the work to date has concentrated on mismeasured covariates with right-censored data. This thesis addresses the mismeasured covariate problem with interval-censored lifetime data. The study discussed in the next section was the motivation for the work described in this chapter as well as that on multi-state progressive models presented in the next chapter.

2.2 Motivating Study

Psoriasis is a chronic disease that causes scaling and swelling of the skin. Unfortunately about $10-42\%$ of those who suffer from this disease also develop *psoriatic arthritis* (PsA) which is characterized by pain, stiffness, swelling and tenderness in and around the joints. This secondary disease was first described in 1818 by Alibert, a French physician, but it was not until the 1950's when it was recognized as a distinct form of arthritis. In the past,

aggressive treatment regimes have been avoided due to potential adverse effects. However, PsA is a progressive disease in the sense that without treatment, it can increase in severity causing disability through deformity and destruction of the joints (Gladman et al. 1995). There are five types of PsA:

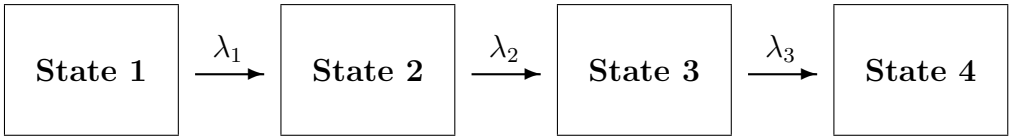
- Symmetric Arthritis, which can affect multiple symmetric pairs of joints and behaves similarly to a mild form of Rheumatoid Arthritis,
- Asymmetric Arthritis, which can involve any number of joints, but does not necessarily involve symmetric pairs,
- Distal Interphalangeal Predominant (DIP), which involves the joints closest to the nails on the fingers and toes,
- Spondylitis, which involves inflammation of the spinal column, impairing movement, and
- Arthritis Mutilans, which is the most severe form, involving deformity and destruction of the joints (Kelley et al. 1981; National Psoriasis Foundation 2004).

It is of interest to determine prognostic factors that relate to disease severity (Gladman et al. 1995). The objective would be to treat individuals who are considered more likely to develop severe PsA early to help prevent or slow progression of the disease. It has been found that early indicators of disease severity include young age at onset, spinal involvement and having a large number of joints affected (National Psoriasis Foundation 2004).

Gladman et al. (1995) concluded that high numbers of joints having an accumulation of fluid (effusions) and high past medications predict disease progression. Their analysis was based on data obtained from the University of Toronto PsA clinic at the Toronto Western Hospital which was established in 1978 and is currently the largest prospective cohort of PsA patients (Husted et al. 2005). In this cohort, patients are scheduled to be assessed every six months at which point extensive information is recorded regarding clinical and laboratory tests. The data used in their analysis consisted of 143 women and 162 men; the average age being 42.2 years and the average duration of PsA, 6.9 years at

clinic entry. They assumed a multi-state Markov model with four states defined by the number of damaged joints determined by clinical assessment. The rationale behind this state structure was that larger numbers of damaged joints reflect disease severity. Figure 2.1 illustrates this model. The states represent 0, 1-4, 5-9 and 10 or more damaged joints, respectively. A proportional hazards model similar to (1.13) was adopted with constant

Figure 2.1: *Progressive model for PsA based on number of damaged joints assumed in Gladman et al. (1995).*



baseline hazards. Covariates were discretized and grouped so that they could be coded as binary variables to apply the likelihood method of Kalbfleisch & Lawless (1985). Investigated covariates included functional class, number of actively inflamed joints, number of effused joints, Lansbury index, rheumatoid factor, erythrocyte sedimentation rate (ESR) and whether or not the patient was on medication in the past and if so, the medication level. It was assumed that the covariates had common coefficients across the three transitions. This assumption, along with the time-homogeneity assumption, was assessed using likelihood ratio tests. Based on this analysis, it was concluded that high numbers of effusions, actively inflamed joints and some past medications were associated with progression of PsA. ESR level appeared to have a protective effect on PsA progression in the sense that those with a low ESR were less likely to progress through the states to develop severe PsA (Gladman et al. 1995).

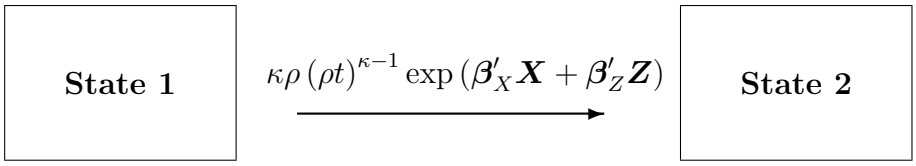
Values of the covariates were obtained through clinical, radiological and serological tests performed during patient assessments, however, only baseline covariate values were considered in their analysis. The covariates were treated as error-free but it is quite reasonable to suspect that there is some degree of error present in some of these measurements. Recorded values may vary between physicians, serological tests are known to be prone

to error in general, and information on medications appears to be based primarily upon patient recall. A more prudent analysis would take this potential uncertainty into account. Information regarding the extent of measurement variability in some of these clinical measurements has recently been gathered by way of reliability studies carried out on patients from this clinic. The results are described in Gladman et al. (1990) and Gladman et al. (2004). These studies demonstrate that there are often imperfect covariate assessments in patients with PsA so valuable information regarding the measurement process in these predictors can be used to improve the analyses. We consider an analysis which accounts for measurement error later in this chapter and compare the results to a naive maximum likelihood approach. We begin by considering a simpler two state model in this chapter to demonstrate the effects of measurement error with interval-censored lifetime data. The outcome that will be considered is the development of the first damaged joint identified by way of clinical examination.

2.3 Impact of Ignoring Error in Covariates

Assume that the true underlying process is represented by Figure 2.2. We consider tran-

Figure 2.2: *Time homogeneous two-state progressive model.*



sition times which follow a proportional hazards Weibull regression model with hazard function,

$$\lambda_T\left(t|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}\right)=\kappa\rho\left(\rho t\right)^{\kappa-1}\exp\left(\boldsymbol{\beta}'_X\boldsymbol{X}+\boldsymbol{\beta}'_Z\boldsymbol{Z}\right)$$

and hence, survivor function,

$$\mathcal{F}_T\left(t|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}\right)=\exp\left\{-\left(\rho t\right)^{\kappa}\exp\left(\boldsymbol{\beta}'_X\boldsymbol{X}+\boldsymbol{\beta}'_Z\boldsymbol{Z}\right)\right\}.$$

The following notation will be used throughout this chapter.

- $i = 1, 2, ..., n$ indexes subjects in the study,
- the assessment times for subject i are u_{ij}, $j = 1, 2, ..., m_i$,
- t_i is the transition time for subject i which is unobserved,
- if $u_{i,j-1} < t_i \leq u_{ij}$ then $c_i = u_{i,j-1}$, $d_i = u_{ij}$ and $\delta_i = 1$ to indicate the transition time is interval-censored,
- if $t_i > u_{im_i}$ then $c_i = u_{im_i}$ and $\delta_i = 0$ to indicate the transition time is right-censored,
- $\boldsymbol{w_i}$ is a mismeasured version of the true unobserved $(p_x \times 1)$ fixed covariate vector, $\boldsymbol{x_i}$, and
- $\boldsymbol{z_i}$ is a perfectly measured $(p_z \times 1)$ covariate vector.

If a standard Weibull regression model were fit to the data, the naive likelihood function would be:

$$\mathcal{L}(\boldsymbol{\theta}^*) = \prod_{i=1}^{n} \left[\mathcal{F}_T^*(c_i|\boldsymbol{w_i}, \boldsymbol{z_i}; \boldsymbol{\theta}^*) - \mathcal{F}_T^*(d_i|\boldsymbol{w_i}, \boldsymbol{z_i}; \boldsymbol{\theta}^*)\right]^{\delta_i} \left[\mathcal{F}_T^*(c_i|\boldsymbol{w_i}, \boldsymbol{z_i}; \boldsymbol{\theta}^*)\right]^{1-\delta_i}, \tag{2.3}$$

A "$*$" is attached to the parameters in this model to emphasize that they differ from the true model parameters in Figure 2.2. In this formulation, we assume the assessment scheme is noninformative and the structure of the hazard function is specified correctly.

Maximization of (2.3) will result in estimates for $\boldsymbol{\theta}^*$, not the parameters of interest, $\boldsymbol{\theta}$. Since the estimators for $\boldsymbol{\theta}^*$ are based on mismeasured covariates, we would expect them to be inconsistent for $\boldsymbol{\theta}$. Determination of the limiting values can provide insight into the effects of mismeasured covariates and illustrate their impact. White (1982) described the asymptotic properties of maximum likelihood estimators under misspecified models which we now briefly review.

In the current setting, the response, $\boldsymbol{Y}$ consists of a vector of the observed states (i.e. 1 or 2) at each of the assessment times. Let $f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta})$ be the distribution from which

the data are generated (i.e. the true distribution) and let $f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}^*)$ be the assumed distribution. In practice, naive maximum likelihood estimates are obtained based on the naive probability distribution and hence the naive likelihood function given by

$$l_{naive}(\boldsymbol{\theta}^*) = \sum_{i=1}^{n} \log f(\boldsymbol{y_i}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*).$$

Estimates are obtained by solving

$$S_{naive}(\boldsymbol{\theta}^*) = \frac{\partial l_{naive}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} = \mathbf{0},$$

and White (1982) showed that the resulting naive "maximum likelihood estimator" is a strongly consistent estimator for $\boldsymbol{\theta}^*$, the parameter value which minimizes the Kullback-Leibler Information Criterion (KLIC) given by

$$I(\boldsymbol{\theta},\boldsymbol{\theta}^*) = E_{Y,W,X,Z}\left\{\log\left[\frac{f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}^*)}\right]\right\} = E_{W,X,Z}\left\{E_{Y|W,X,Z}\left[\log\left(\frac{f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}^*)}\right)\right]\right\}, \tag{2.4}$$

where the expectation is taken with respect to the true underlying distribution. Assuming nondifferential measurement error, the inside expectation of (2.4) can be rewritten as

$$\int \log f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})\,\mathrm{d}F(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) - \int \log f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}^*)\,\mathrm{d}F(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}).$$

Intuitively, the KLIC is a measure of ignorance about the true structure of the distribution when $f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}^*)$ is used to model data generated from $f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})$ (note that when model is correctly specified the KLIC is 0). To obtain the value of $\boldsymbol{\theta}^*$ which minimizes (2.4), we first take its derivative with respect to $\boldsymbol{\theta}^*$ and set it to $\mathbf{0}$:

$$\frac{\partial I(\boldsymbol{\theta},\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} = \frac{\partial E_{Y,W,X,Z}\left[\log \frac{f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}^*)}\right]}{\partial \boldsymbol{\theta}^*} = \mathbf{0}. \tag{2.5}$$

Given certain regularity conditions hold (White 1982), (2.5) is equivalent to

$$E_{W,X,Z}\left[E_{Y|W,X,Z}\left(\frac{\partial\left(\log \frac{f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}^*)}\right)}{\partial \boldsymbol{\theta}^*}\right)\right] \;=\; \mathbf{0},$$

giving

$$E_{W,X,Z}\left\{\frac{\partial\int\log f\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}\right)\mathrm{d}F\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}^{*}}-\frac{\partial\int\log f\left(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}^{*}\right)\mathrm{d}F\left(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{z};\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}^{*}}\right\} \;=\; \mathbf{0}.$$

Since the expectation of the first term is zero ($\int\log f\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}\right)\mathrm{d}F\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}\right)$ does not depend on $\boldsymbol{\theta}^{*}$), we note that

$$E_{Y,W,X,Z}\left[S_{naive}\left(\boldsymbol{\theta^{*}}\right);\boldsymbol{\theta}\right]=\mathbf{0}. \tag{2.6}$$

As can be seen from the expression, (2.6) implies a relation $\boldsymbol{\theta}^{*}=g\left(\boldsymbol{\theta}\right)$ but g may be a very complicated function. Turnbull et al. (1997) use this idea to develop adjustments to the naive maximum likelihood estimators in the presence of measurement error in covariates for a mixed effects Poisson regression model for data involving recurrent events. If it is difficult to derive an explicit expression for $\boldsymbol{\theta}$ in terms of $\boldsymbol{\theta}^{*}$ (or vice versa), (2.6) can be solved numerically.

Here we apply a similar approach to investigate the asymptotic bias in the case of measurement error with interval-censored failure time data. Assuming the mild regularity conditions outlined in White (1982) hold, the derivative and the expectation operators in (2.6) can be interchanged such that

$$\frac{\partial E_{Y,W,X,Z}\left[l_{naive}\left(\boldsymbol{\theta^{*}}\right)\right]}{\partial\boldsymbol{\theta^{*}}}=\mathbf{0},$$

where $\boldsymbol{Y}$ represents the states (1 or 2) occupied at each assessment time. For the purpose of this investigation, all subjects were assumed to enter the study at time 0 in state 1 and to be assessed at five equally spaced times in addition to the baseline assessment (i.e. m=5 and there are a total of six assessments for each subject) . The study duration, τ, was selected such that $P\left(T<\tau\right)$ was at least 0.6 or 0.8 for all values of (X,Z) (binary covariates) or such that $P\left(T<\tau\right)$ was 0.6 or 0.8 at $\boldsymbol{\mu}'=\left(\mu_{X},\mu_{Z}\right)'=\left(0,0\right)'$ (continuous (X,Z)). The vectors $\left(\boldsymbol{Y},X,W,Z\right)'$ were assumed to be independent and identically distributed across individuals so that we could focus on the contributions from a single individual. Suppose $\mathcal{F}_{T}\left(t|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}\right)$ represents the survivor function for the true underlying distribution, considered to be a Weibull regression model here, and $\mathcal{F}_{T}^{*}\left(t|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^{*}\right)$ represents the survivor

distribution for naive model. The naive survivor function may be equal to $\mathcal{F}_T\left(t|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*\right)$ if the model is specified correctly aside from using $\boldsymbol{W}$ in place of $\boldsymbol{X}$. Denote the state occupied by individual i at time u_{ij}, $i=1,2,...,m$, as $y_i(u_{ij})$ and let

$$P^*_{y_i(u_{i,j-1}),y_i(u_{i,j-1})}\left(u_{ij}-u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right)=P^*\left(Y_{ij}=y_i(u_{ij})|Y_{i,j-1}=y_i(u_{i,j-1}),\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right),$$

so

$$\begin{aligned}
P^*_{1,1}\left(u_{ij}-u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right) &= \frac{\mathcal{F}^*_T(u_{ij}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*)}{\mathcal{F}^*_T(u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*)}, \qquad (2.7)\\
P^*_{1,2}\left(u_{ij}-u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right) &= \frac{\mathcal{F}^*_T(u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*)-\mathcal{F}^*_T(u_{ij}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*)}{\mathcal{F}^*_T(u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*)}, \text{ and}\\
P^*_{2,2}\left(u_{ij}-u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right) &= 1.
\end{aligned}$$

Based on this notation the likelihood given in (2.3) can be re-expressed as:

$$\mathcal{L}\left(\boldsymbol{\theta}^*\right)=\prod_{i=1}^{n}\prod_{j=1}^{m}P^*_{y_i(u_{i,j-1}),y_i(u_{i,j-1})}\left(u_{ij}-u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right). \qquad (2.8)$$

Subject i contributes $\mathcal{F}^*_T\left(\tau|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right)$ to the naive likelihood if T_i is right-censored (i.e. $T_i>\tau$) and $\mathcal{F}^*_T\left(u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right)-\mathcal{F}^*_T\left(u_{ij}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta}^*\right)$ if the transition time occurs between $u_{i,j-1}$ and u_{ij} for some $j=1,2,...,m$. Let $\mathcal{P}$ represent the set of all possible values of $\boldsymbol{Y}$ (i.e. all possible state paths) and $\boldsymbol{v}$ be a six dimensional vector. Then the function equivalent to (2.6) that should be maximized with respect to θ^* in this setting is $E_{Y,W,X,Z|Y_0}\left[l_{naive}\left(\boldsymbol{\theta^*}\right)\right]$ which is given by:

$$
\begin{aligned}
& E_{Y,W,X,Z|Y_0}\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\log\left[P^*_{v_{j-1},v_j}\left(u_{ij}-u_{i,j-1}|\boldsymbol{W_i},\boldsymbol{Z_i};\boldsymbol{\theta}^*\right)\right]^{I(\boldsymbol{Y_i}=\boldsymbol{v})}\right\} \\
= \; & E_{Y,W,X,Z|Y_0}\left\{\sum_{i=1}^{n}\sum_{j=1}^{5} I(\boldsymbol{Y_i}=\boldsymbol{v})\log\left[P^*_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{W_i},\boldsymbol{Z_i};\boldsymbol{\theta}^*\right)\right]\right\} \\
= \; & nE_{W,X,Z}\left\{E_{Y|W,X,Z,Y_0}\left[\sum_{j=1}^{5} I(\boldsymbol{Y}=\boldsymbol{v})\log\left(P^*_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*\right)\right)\right]\right\} \\
= \; & nE_{W,X,Z}\left\{\sum_{\boldsymbol{v}\in\mathcal{P}} P(\boldsymbol{Y}=\boldsymbol{v}|\boldsymbol{W},\boldsymbol{X},\boldsymbol{Z},Y_0;\boldsymbol{\theta})\sum_{j=1}^{5}\log\left(P^*_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*\right)\right)\right\} \\
= \; & nE_{W,X,Z}\left\{\sum_{\boldsymbol{v}\in\mathcal{P}}\prod_{j=1}^{5} P_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}\right)\sum_{j=1}^{5}\log\left(P^*_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*\right)\right)\right\}.
\end{aligned}
$$

In the above expression, $P_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}\right)$ represents the true model version of (2.7). Specifically

$$
\begin{aligned}
P_{1,1}\left(u_{ij}-u_{i,j-1}|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\theta}\right) &= \frac{\mathcal{F}_T(u_{ij}|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\theta})}{\mathcal{F}_T(u_{i,j-1}|\boldsymbol{w_i},\boldsymbol{z_i};\boldsymbol{\theta})}, \\
P_{1,2}\left(u_{ij}-u_{i,j-1}|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\theta}\right) &= \frac{\mathcal{F}_T(u_{i,j-1}|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\theta})-\mathcal{F}_T(u_{ij}|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\theta})}{\mathcal{F}_T(u_{i,j-1}|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\theta})}, \text{ and} \\
P_{2,2}\left(u_{ij}-u_{i,j-1}|\boldsymbol{x_i},\boldsymbol{z_i};\boldsymbol{\theta}\right) &= 1.
\end{aligned} \tag{2.9}
$$

Naive models given by a Weibull regression model (correctly specified aside from the mismeasured covariate) and a robust piecewise constant baseline hazards model will be considered in the following sections. Graphical displays are presented which illustrate the bias in naive maximum likelihood estimators as functions of misclassification or measurement error for the two-state models discussed here.

2.3.1 Binary Covariates

The function maximized with respect to $\boldsymbol{\theta}^*$ is

$$
\begin{aligned}
&\textstyle\sum_{x=0}^{1}\sum_{w=0}^{1}\sum_{z=0}^{1} P\left(X=x,W=w,Z=z\right) \\
&\cdot\left\{\textstyle\sum_{\boldsymbol{v}\in\mathcal{P}}\prod_{j=1}^{5} P_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}\right)\sum_{j=1}^{5}\log\left[P^*_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*\right)\right]\right\},
\end{aligned} \tag{2.10}
$$

where $P_{v_{j-1},v_j}(\cdot)$ and $P^*_{v_{j-1},v_j}(\cdot)$ are given by (2.9) and (2.7), respectively, the true underlying model is

$$\mathcal{F}_T(t|X,Z;\boldsymbol{\theta}) = \exp\left[-(\rho t)^{\kappa}\exp(\beta_X X + \beta_Z Z)\right]$$

and the naive fitted model based on a Weibull regression model is

$$\mathcal{F}^*_T(t|W,Z;\boldsymbol{\theta}^*) = \exp\left[-(\rho^* t)^{\kappa^*}\exp(\beta^*_X W + \beta^*_Z Z)\right].$$

The piecewise constant baseline hazards (PCBH) model is

$$\mathcal{F}^*_T(t|W,Z;\boldsymbol{\theta}^*) = \exp\left[-\lambda_0(t)\exp(\beta^*_X W + \beta^*_Z Z)\right],$$

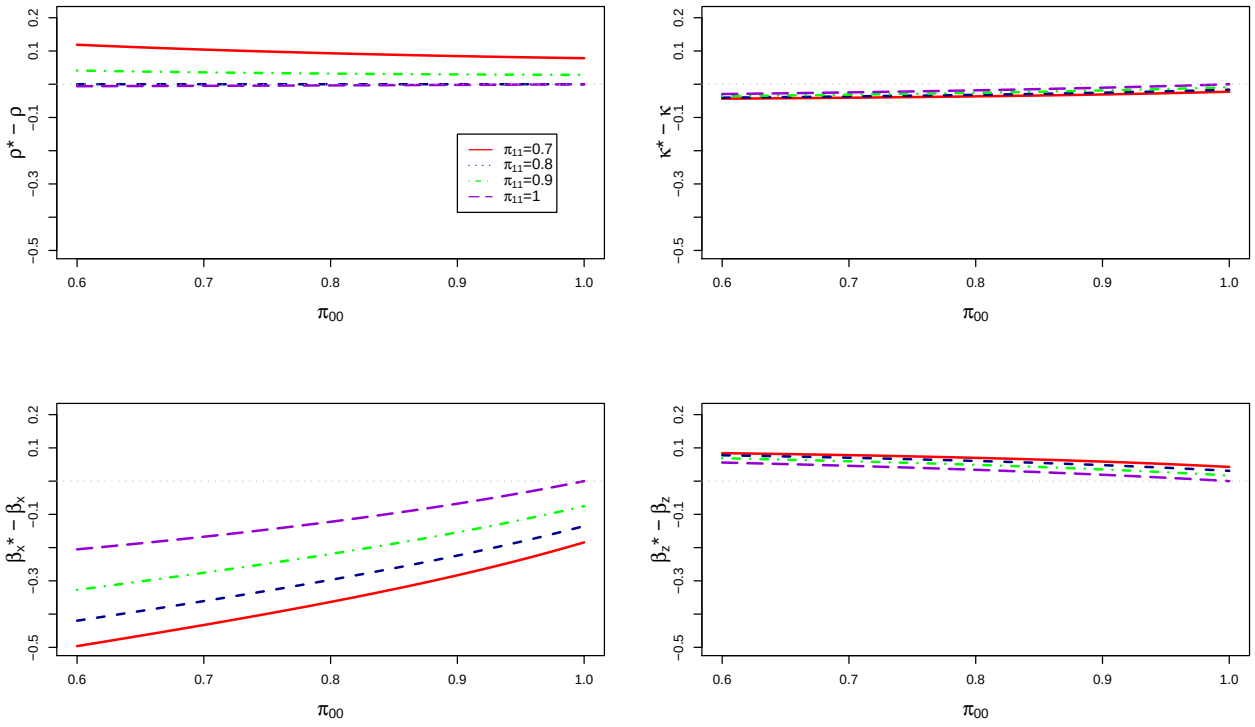
with

$$\lambda_0(t) = \begin{cases} \lambda_{01}, & 0 \le t < \tau/4 \\ \lambda_{02}, & \tau/4 \le t < \tau/2 \\ \lambda_{03}, & \tau/2 \le t < 3\tau/4 \\ \lambda_{04}, & 3\tau/4 \le t < \tau \end{cases}.$$

Misclassification of X is characterized by the misclassification probabilities, $\pi_{01} = P(W=0|X=1)$ and $\pi_{10} = P(W=1|X=0)$ or equivalently, by the specificity (i.e. $\pi_{11} = 1-\pi_{01}$) and the sensitivity (i.e. $\pi_{00} = 1-\pi_{10}$). Optimization of (2.10) was carried out via PROC NLP in SAS based on a quasi-Newton algorithm. Figure 2.3 contains a plot of the asymptotic bias of the four naive estimators in the Weibull regression model for a representative parameter configuration. Similar trends were observed for the other parameter configurations investigated. In practice, concern often lies in the covariate effects rather than the parameters associated with the baseline hazard, so Figures 2.4 to 2.8 display and compare the asymptotic bias in β_X and β_Z estimators based on Weibull regression and PCBH models.

It is clear from Figure 2.3 that even if the structure of the model is specified correctly, using a misclassified version of the true covariate in the model leads to asymptotic bias in the four estimators. As expected, the magnitude of the bias increases as the degree of misclassification present increases and it appears to be greatest for the estimator associated with the misclassified covariate. Values of κ investigated were 0.5, 1 and 2 to represent a

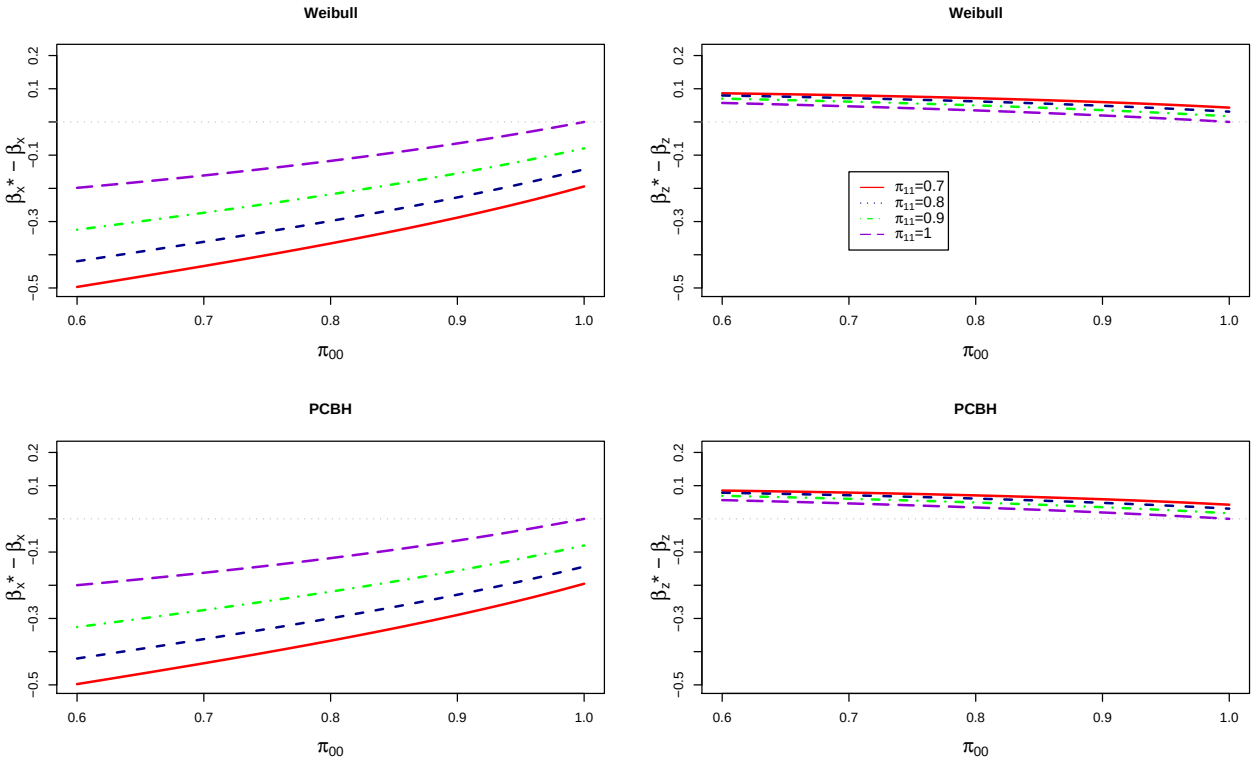
Figure 2.3: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\kappa = 0.5$, $\beta_X = \log(2)$, $\beta_Z = \log(1.25)$*; maximum right censoring rate at* τ *is 20%;* $P(Z = 1) = 0.5$ *and* $\text{logit}[P(X = 1|Z = z)] = \log(2)\, z$.



range of plausible values. Since the plots appeared quite similar for different values of κ, plots corresponding to $\kappa = 1$ are presented in Figures 2.4 to 2.8.

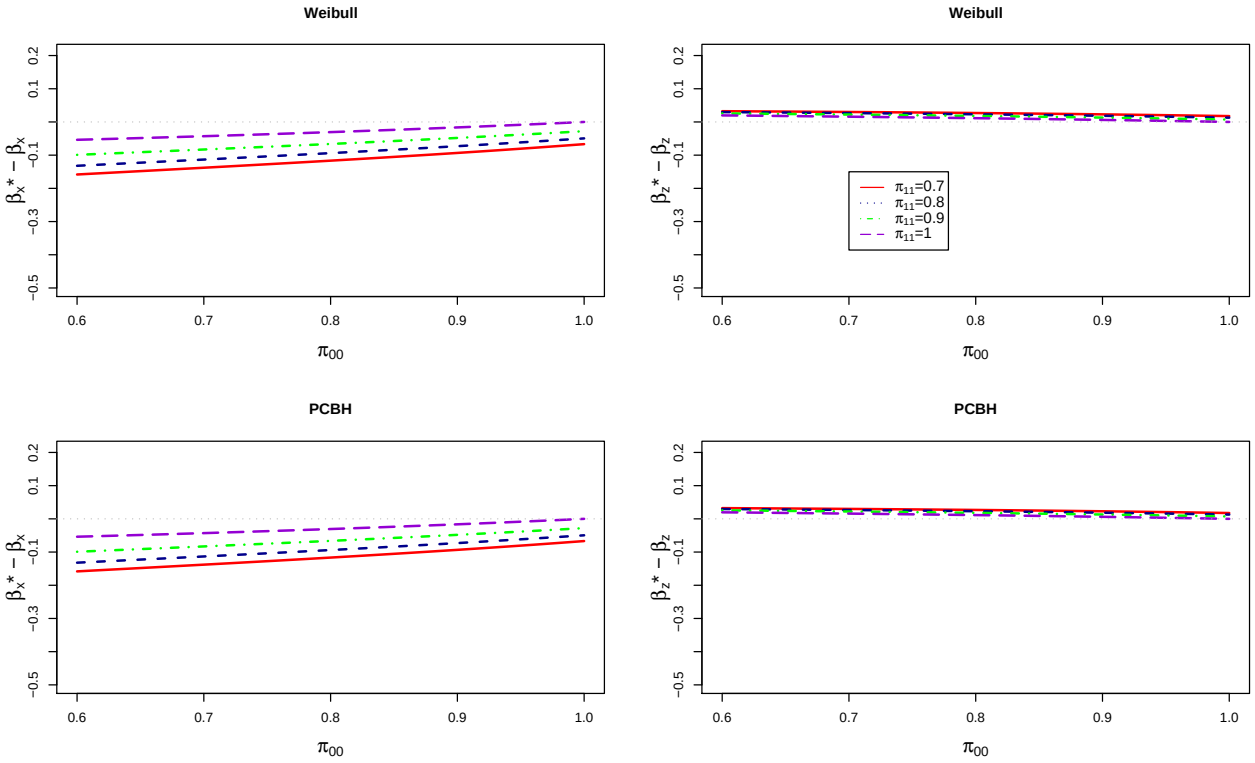
Based on these figures, the asymptotic biases of the naive estimators of the covariate effects appear to be similar for the Weibull and PCBH models. This suggests that the PCBH model provides a robust approach for structural model misspecification but a similar effect of covariate misclassification can be expected. Its performance for finite samples, and for use in methods accounting for misclassification, will be examined in the simulation study summarized in Section 2.5.1.

Figure 2.4: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\boldsymbol{\kappa} = \mathbf{1}$, $\beta_X = \log(2)$, $\beta_Z = \log(1.25)$*; maximum right censoring rate at* τ *is 20%;* $P(Z = 1) = 0.5$ *and* $\text{logit}[P(X = 1|Z = z)] = \log(2)\, z$.



In addition to the apparent effects of the misclassification rates, the magnitude of the bias in estimators for both β_X and β_Z appears to be driven by the true underlying value of β_X. It seems to increase as the true underlying effect of X increases in magnitude and based on the parameter configurations investigated here, the estimator for β_Z appears to exhibit smaller asymptotic bias even when β_X and β_Z are the same (see Figure 2.5). This is possibly because X and Z are positively correlated for Figures 2.3 to 2.7. When they are uncorrelated, as in Figure 2.8, there appears to be negligible asymptotic bias in the estimator for β_Z. Upon comparison of Figure 2.6 to the other plots, it appears that the sign of the true underlying X effect can impact the direction of the asymptotic bias.

Figure 2.5: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\kappa = 1$, $\boldsymbol{\beta_X = \log(1.25)}$, $\beta_Z = \log(1.25)$*; maximum right censoring rate at* τ *is 20%;* $P(Z = 1) = 0.5$ *and* $\text{logit}[P(X = 1|Z = z)] = \log(2)\, z$.

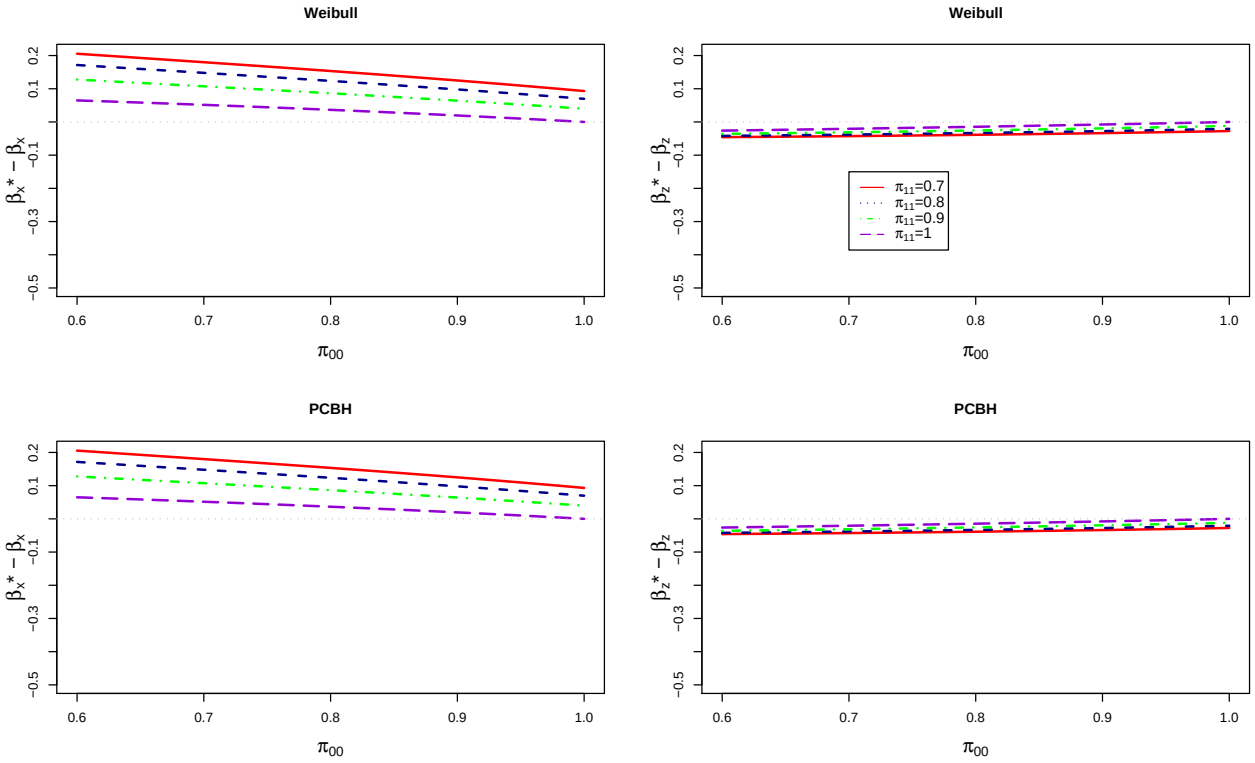


Asymptotically, the naive estimator for β_X underestimates the magnitude of the true effect (i.e. there appears to be an attenuation effect) and although the true value for β_Z remains unchanged, its asymptotic bias is in the other direction.

Now we consider current status data which is a special case of interval-censored lifetime data when a single inspection time is available. The equation to be maximized in this setting is

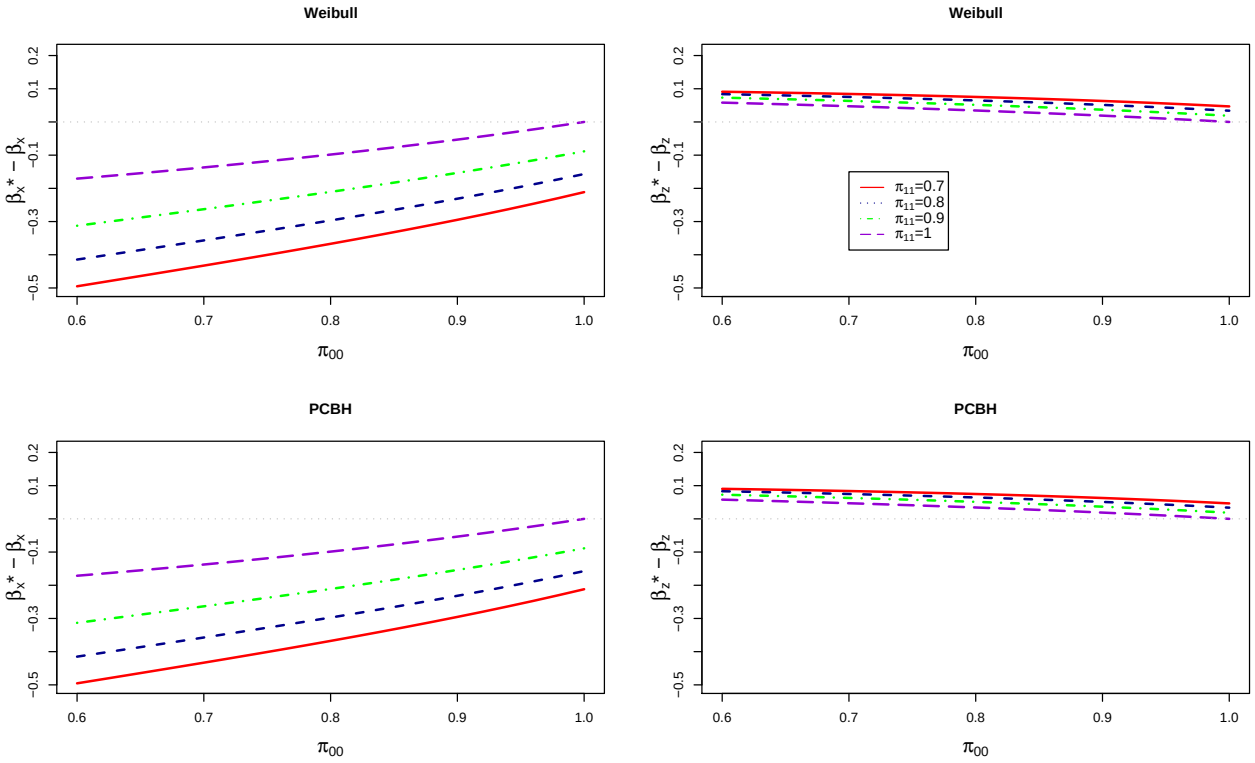
$$\sum_{x=0}^{1}\sum_{w=0}^{1}\sum_{z=0}^{1} P(X = x, W = w, Z = z)\left\{\mathcal{F}_T(b|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}) \log\left[\mathcal{F}_T^*(b|\boldsymbol{W}, \boldsymbol{Z}; \boldsymbol{\theta}^*)\right]\right.$$
$$\left. + \left[1 - \mathcal{F}_T(b|\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})\right] \log\left[1 - \mathcal{F}_T^*(b|\boldsymbol{W}, \boldsymbol{Z}; \boldsymbol{\theta}^*)\right]\right\},$$

Figure 2.6: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\kappa = 1$, $\boldsymbol{\beta_X = \log(0.75)}$, $\beta_Z = \log(1.25)$*; maximum right censoring rate at* τ *is 20%;* $P(Z = 1) = 0.5$ *and* $\text{logit}[P(X = 1|Z = z)] = \log(2)\, z$.



where b is the assessment time. Suppose that τ is determined as in the general interval-censored situation above (i.e. such that $\min(T < \tau|x, z) = 0.8$), and that patients are observed once at assessment time 0.75τ. Figures 2.9 to 2.10 illustrate the asymptotic bias in the estimators for a representative parameter configuration. Only two plots were included here since the asymptotic bias exhibited in all plots created appeared to be pretty much consistent with the general interval censoring context. Since more information can be ascertained about the transition time as the number of assessments increases, it seems reasonable to suspect that there will be a difference in the performance of current status data versus general interval censoring for finite samples. This will be examined in the

Figure 2.7: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\kappa = 1$, $\boldsymbol{\beta_X = \log(2)}$, $\beta_Z = \log(1.25)$*; maximum right censoring rate at* τ *is* **40%**; $P(Z = 1) = 0.5$ *and* $\text{logit}\left[P(X = 1|Z = z)\right] = \log(2)\, z$.



supplementary simulation results presented later in this chapter for binary covariates.

Figure 2.8: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\kappa = 1$, $\beta_X = \log(2)$, $\beta_Z = \log(1.25)$*; maximum right censoring rate at* τ *is 40%;* $P(Z = 1) = 0.5$ *and* $\mathbf{logit}\left[\boldsymbol{P}\left(\boldsymbol{X} = \mathbf{1}|\boldsymbol{Z} = \boldsymbol{z}\right)\right] = \mathbf{0.5}$.

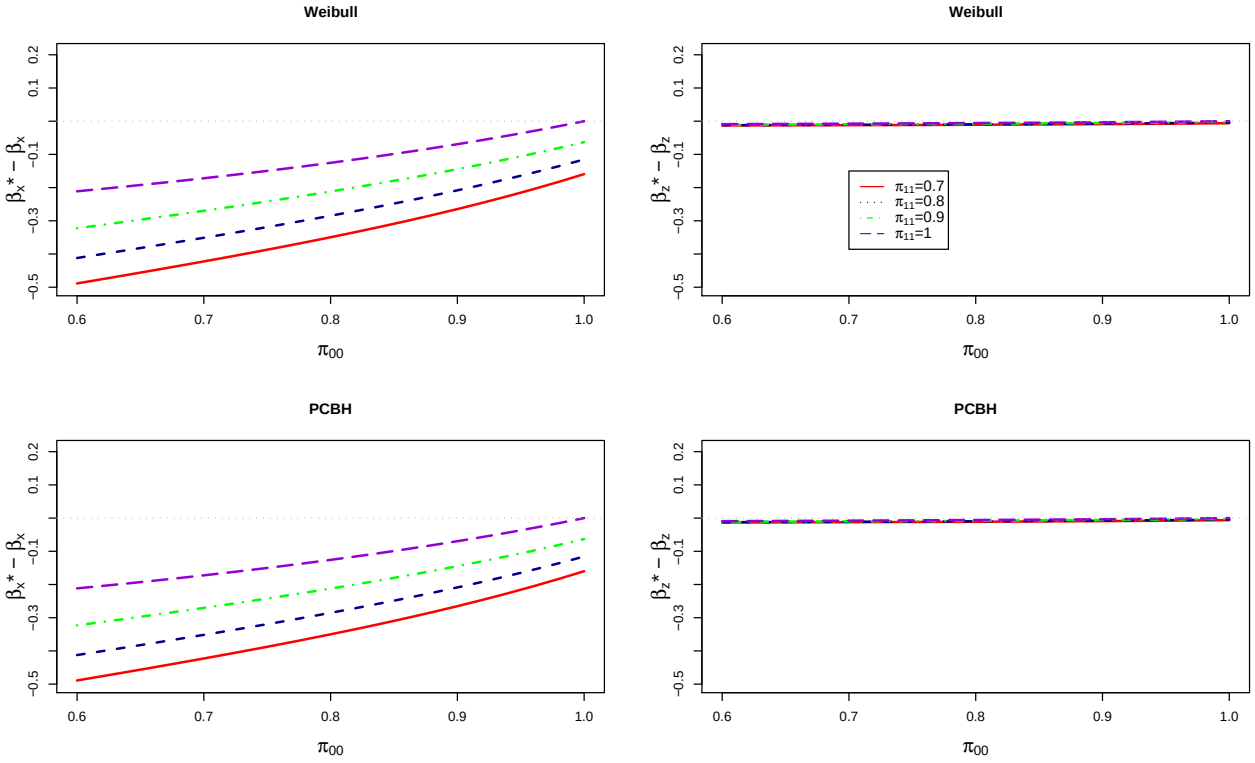


Figure 2.9: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model with a misclassified binary covariate based on current status data; assessment time 0.75τ; $\rho = 0.2$, $\kappa = 0.5$, $\beta_X = \log(2)$, $\beta_Z = \log(1.25)$; maximum right censoring rate at τ is 40%; $P(Z = 1) = 0.5$ and* $\text{logit}\left[P\left(X = 1|Z = z\right)\right] = log\left(2\right) z$.

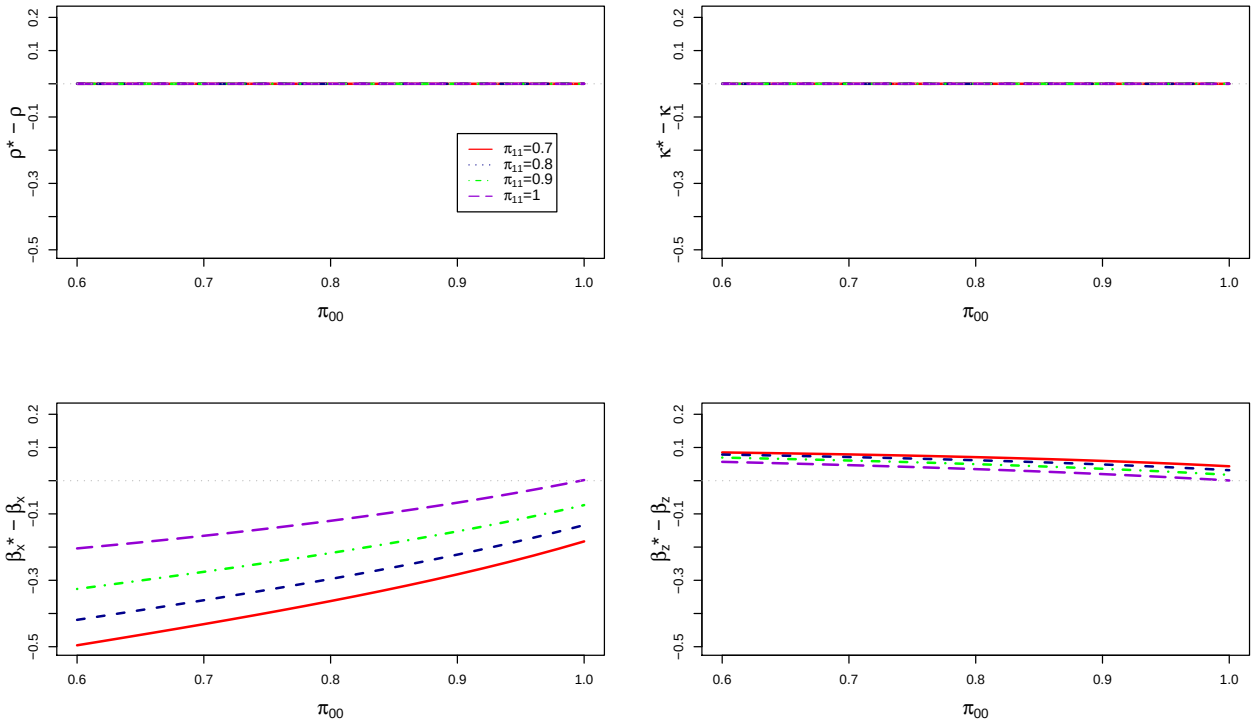
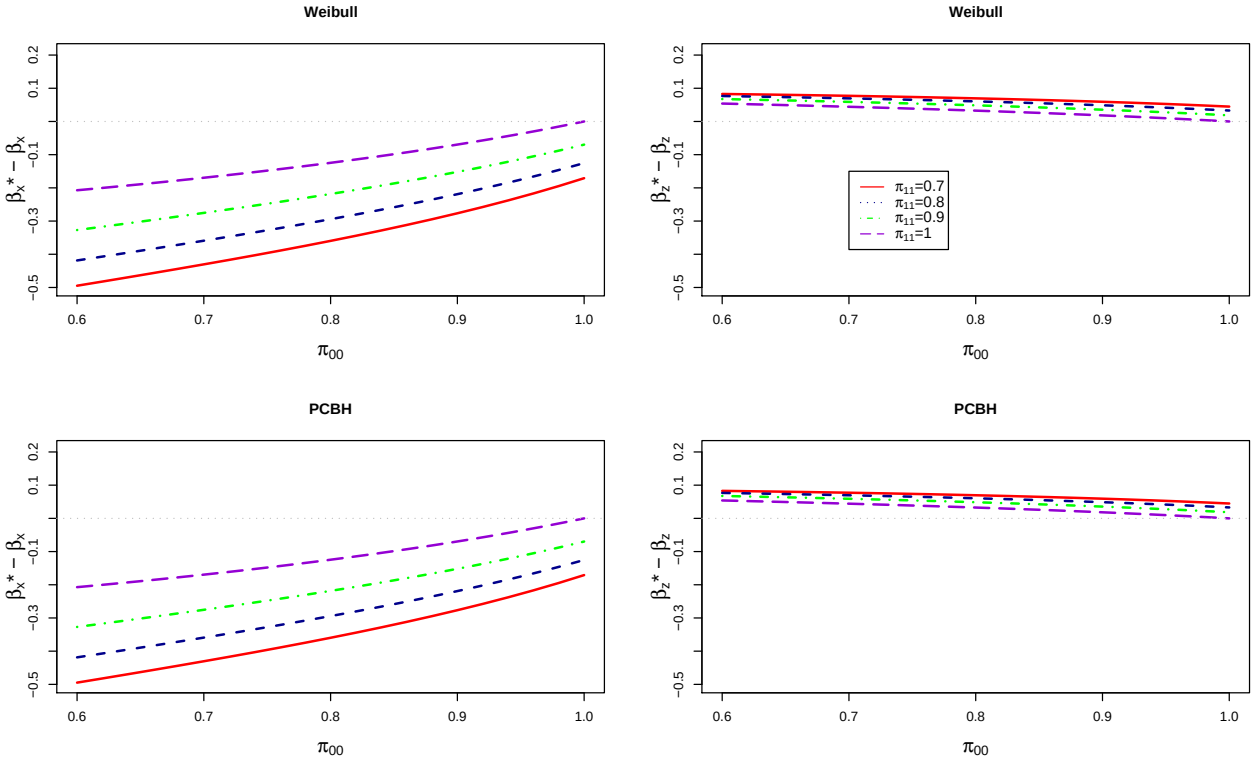


Figure 2.10: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a misclassified binary covariate based on current status data; assessment time 0.75τ; ρ = 0.2, κ = 0.5,* $\beta_X = \log(2)$, $\beta_Z = \log(1.25)$*; maximum right censoring rate at τ is 40%;* $P(Z=1) = 0.5$ *and* $\text{logit}[P(X=1|Z=z)] = \log(2)\,z$.



2.3.2 Continuous Covariates

The function maximized with respect to $\boldsymbol{\theta}^*$ here is

$$\begin{array}{l}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,W,Z}\left(x,w,z\right) \\ \cdot\left\{\sum_{\boldsymbol{v}\in\mathcal{P}}\prod_{j=1}^{5} P_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta}\right)\sum_{j=1}^{5}\log\left[P^*_{v_{j-1},v_j}\left(\tau/5|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*\right)\right]\right\}\mathrm{d}w\mathrm{d}x\mathrm{d}z,\end{array} \tag{2.11}$$

where $P_{v_{j-1},v_j}\left(\cdot\right)$ and $P^*_{v_{j-1},v_j}\left(\cdot\right)$ are given by (2.9) and (2.7), respectively. In addition, $f_{W|X,Z}\left(w|x,z\right)$ is the probability density function (p.d.f.) of a $N\left(x,\sigma_U^2\right)$ distribution (σ_U^2 is the measurement error variance), $f_{X|Z}\left(x|z\right)$ is the p.d.f. of a $N\left(\xi_Z Z,\sigma^2_{X|Z}\right)$ distribution, $f_Z\left(z\right)$ is the p.d.f. of a $N\left(0,\sigma_Z^2\right)$ distribution, and

$$f_{X,W,Z}\left(x,w,z\right)=f_{W|X,Z}\left(w|x,z\right)f_{X|Z}\left(x|z\right)f_Z\left(z\right).$$

As in the binary case, the true underlying model was specified as

$$\mathcal{F}_T\left(t|X,Z;\boldsymbol{\theta}\right)=\exp\left[-\left(\rho t\right)^{\kappa}\exp\left(\beta_X X+\beta_Z Z\right)\right]$$

and the naive fitted model was either

$$\mathcal{F}^*_T\left(t|W,Z;\boldsymbol{\theta}^*\right)=\exp\left[-\left(\rho^* t\right)^{\kappa^*}\exp\left(\beta^*_X W+\beta^*_Z Z\right)\right]$$

or

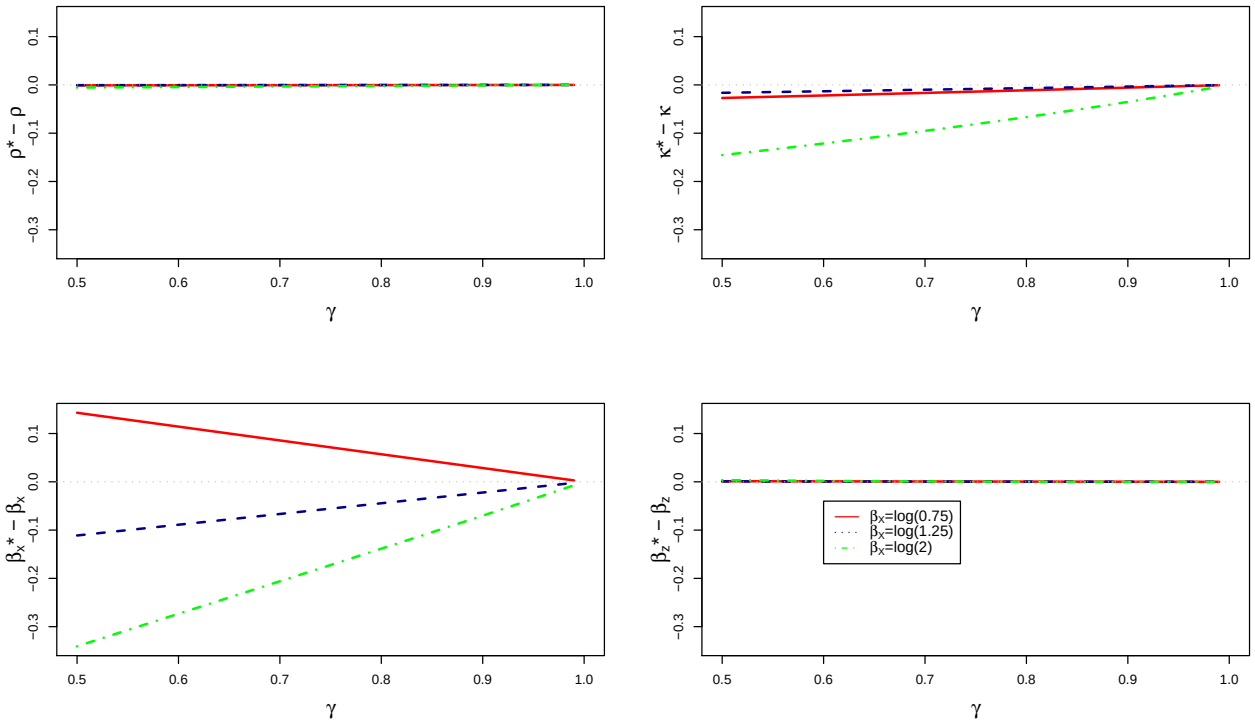
$$\mathcal{F}^*_T\left(t|W,Z;\boldsymbol{\theta}^*\right)=\exp\left[-\lambda_0(t)\exp\left(\beta^*_X W+\beta^*_Z Z\right)\right],$$

with

$$\lambda_0(t)=\begin{cases}\lambda_{01}, & 0\le t<\tau/4\\ \lambda_{02}, & \tau/4\le t<\tau/2\\ \lambda_{03}, & \tau/2\le t<3\tau/4\\ \lambda_{04}, & 3\tau/4\le t<\tau\end{cases},$$

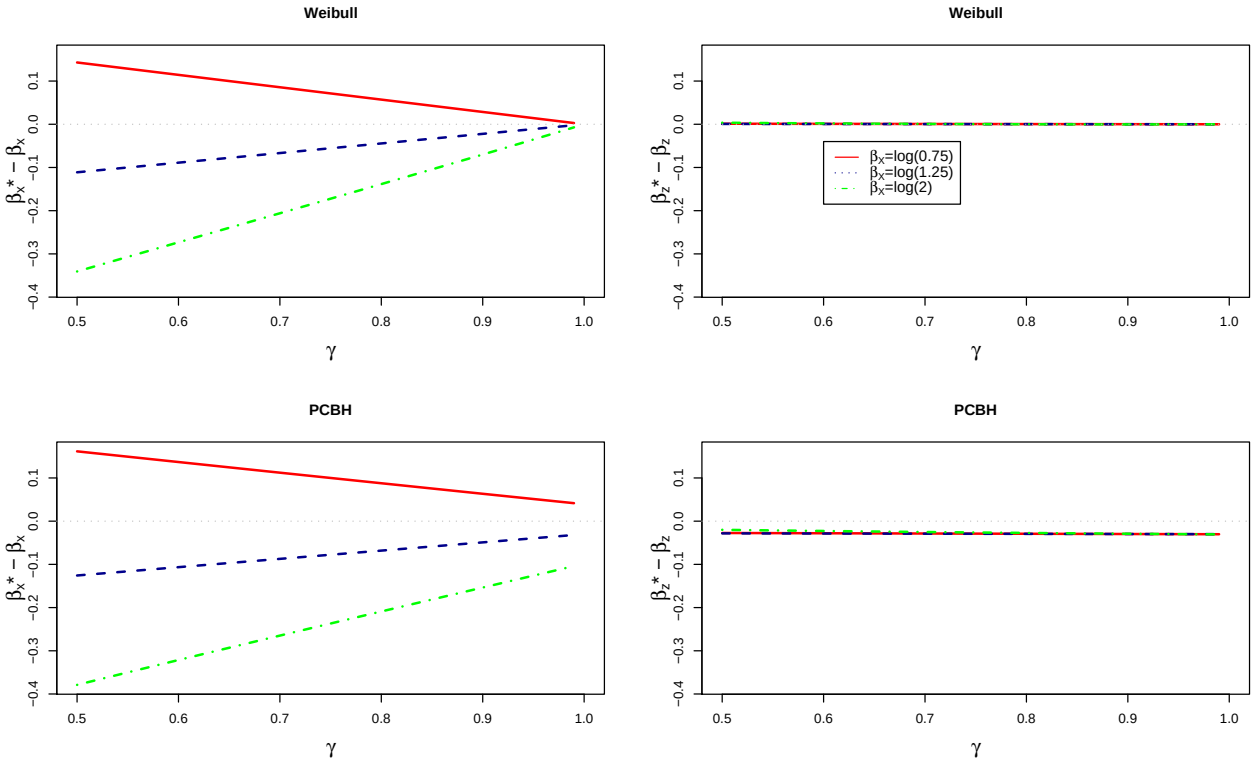
depending on whether a Weibull regression model or a piecewise constant baseline hazards (PCBH) model was assumed. This optimization was conducted in PROC NLMIXED in SAS based on a quasi-Newton algorithm. Numerical integration of the integrals in (2.11) was conducted using adaptive Gaussian quadrature based on the default settings in PROC NLMIXED.

Figure 2.11: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model with a mismeasured continuous covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\kappa = 2$, $\beta_Z = \log(1.25)$*; right censoring rate at* τ *is 20% when evaluated at the means of* X *and* Z*;* $Z \sim N(0,1)$ *and* $X|Z \sim N(0,1)$ *such that* $\rho_{XZ} = 0$.



Figures 2.11 to 2.14 summarize the asymptotic bias for a couple of representative parameter configurations. Measurement error is characterized by the reliability ratio, $\gamma = \sigma^2_{X|Z}/(\sigma^2_{X|Z} + \sigma^2_U)$, with γ ranging from 0.5 to 1 to represent varying degrees of measurement error. Based on these plots, bias tends to increase as the measurement error becomes more severe (i.e. as γ decreases). The bias in the estimators for ρ appears to be negligible, at least for the parameter configuration considered in Figure 2.11. However, estimation of κ seems to be affected and the resulting bias appears to depend on the magnitude of the regression coefficient corresponding to the error-prone covariate. The estimator for β_Z does not appear to exhibit bias if X and Z are uncorrelated, but does when they are correlated. As in the binary covariate setting, the magnitude of the true

Figure 2.12: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a mismeasured continuous covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\kappa = 2$, $\beta_Z = \log(1.25)$*; right censoring rate at* τ *is 20% when evaluated at the means of* X *and* Z*;* $Z \sim N(0,1)$ *and* $X|Z \sim N(0,1)$ *such that* $\rho_{XZ} = 0$.



underlying value for β_X seems to impact the asymptotic bias in the estimators for β_X and β_Z. The asymptotic bias based on the PCBH model looks to be shifted downward slightly from the bias based on a Weibull model when $\kappa = 2$. Since this does not appear to be the case when $\kappa = 1$, it may be due to the piecewise constant approximation to the baseline hazard. Although not presented here, the asymptotic biases based on current status data seemed to exhibit similar trends as the general interval-censored data as was the case in the binary covariate setting.

Figure 2.13: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a mismeasured continuous covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\kappa = 2$, $\beta_Z = \log(1.25)$*; right censoring rate at* τ *is 20% when evaluated at the means of* X *and* Z*;* $Z \sim N(0,1)$ *and* $\boldsymbol{X|Z \sim N(1.33Z, 1)}$ *such that* $\boldsymbol{\rho_{XZ} = 0.8}$.

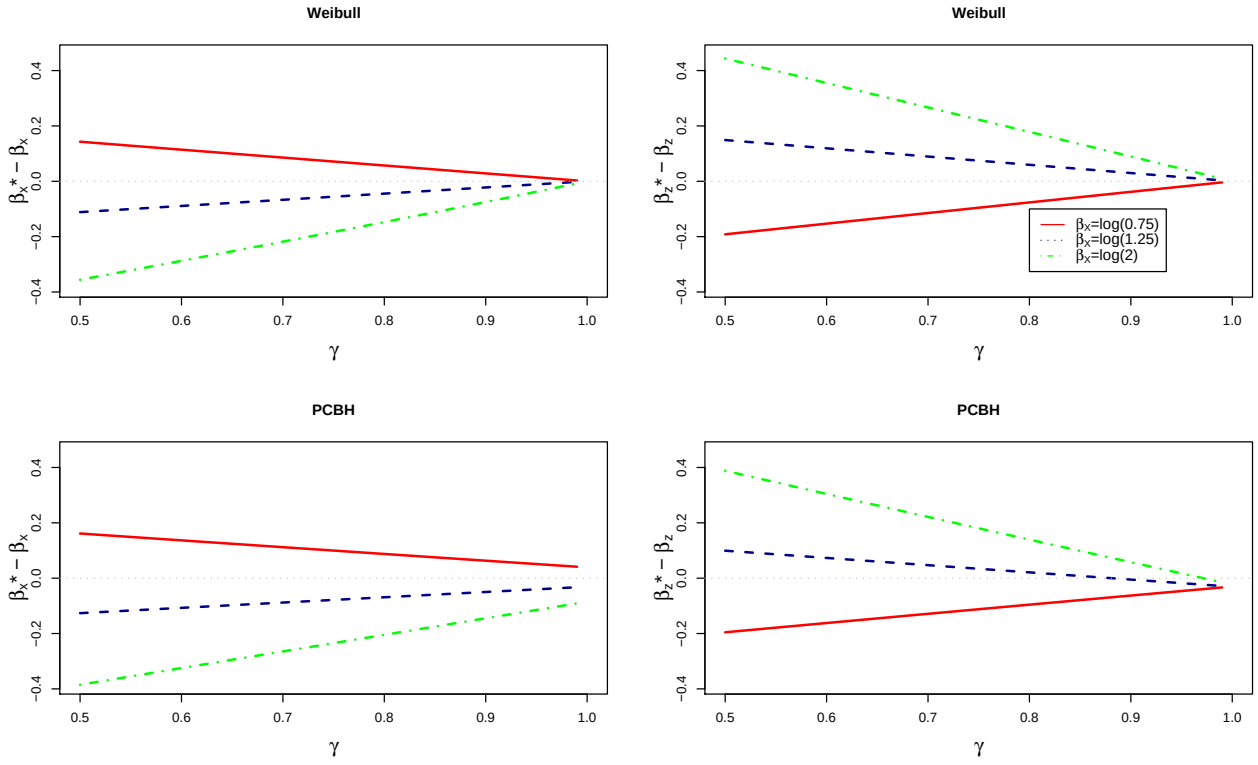
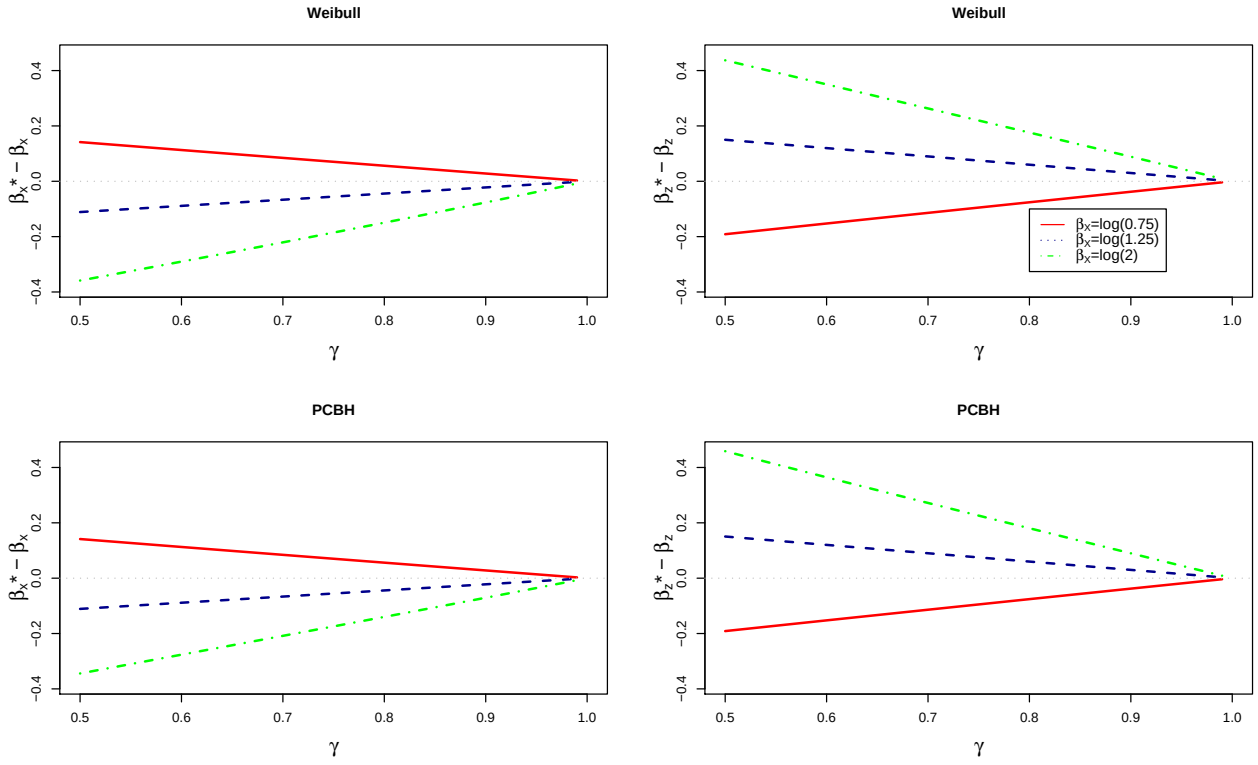


Figure 2.14: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards Weibull regression model and a piecewise baseline hazard (PCBH) model with a mismeasured continuous covariate;* $m = 5$ *equally spaced assessments;* $\rho = 0.2$, $\boldsymbol{\kappa} = \mathbf{1}$, $\beta_Z = \log(1.25)$*; right censoring rate at* τ *is 20% when evaluated at the means of* X *and* Z*;* $Z \sim N(0,1)$ *and* $X|Z \sim N(1.33Z, 1)$ *such that* $\rho_{XZ} = 0.8$.



2.4 Correcting for Mismeasured Covariates

It has been demonstrated that mismeasured covariates induce bias in parameter estimators even when the model is specified correctly otherwise. We now describe and evaluate methods accounting for this error. SIMEX and likelihood approaches will be investigated both for Weibull regression models and models with piecewise constant baseline hazards. These approaches are applicable and can be implemented in a similar way for other lifetime data models although models with piecewise constant baseline hazards are broadly applicable due to their robustness. First we introduce some additional notation.

Continuous X

In the case of continuous covariates, we will also assume the following:

- an error model similar to the classical additive error model in (1.29) is appropriate so that the conditional distribution of $\boldsymbol{W_i}$ given $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ is $MVN(\boldsymbol{\mu}_{W|X,Z}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_{W|X,Z} = \zeta_0 + \boldsymbol{\zeta}'_X \boldsymbol{X}_i + \boldsymbol{\zeta}'_Z \boldsymbol{Z}_i$, and the $\boldsymbol{W_i}$ are conditionally independent given $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ for $i = 1, 2, ..., n$,
- $\boldsymbol{\Sigma}$ is known or information has been obtained regarding this via supplementary data consisting of repeated measurements on $\boldsymbol{W_i}$ or validation data, and
- we are dealing with nondifferential measurement error, that is, the distributions of $\boldsymbol{Y}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z}$ and $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ are equivalent.

Binary X

For the sake of illustration, assume that $\boldsymbol{X_i} = X_i$ and $\boldsymbol{W_i} = W_i$ are fixed one-dimensional binary covariates and that

- $\pi_{10} = P(W_i = 1|X_i = 0, \boldsymbol{z_i})$, or $\pi_{00} = 1 - \pi_{10}$ is the so-called *specificity*,
- $\pi_{01} = P(W_i = 0|X_i = 1, \boldsymbol{z_i})$, or $\pi_{11} = 1 - \pi_{01}$ is the so-called *sensitivity*,
- π_{10} and π_{01} are known or can be estimated from supplementary data, and

- we are dealing with nondifferential misclassification, that is, the distributions of $\boldsymbol{Y}|W, X, \boldsymbol{Z}$ and $\boldsymbol{Y}|X, \boldsymbol{Z}$ are equivalent.

Based on the notation and model setup outlined above, we will describe two inference procedures accounting for misclassification or measurement error in covariates that can be used in the case of a progressive multi-state model with interval-censored data. The first approach that will be introduced, Simulation Extrapolation (SIMEX), is a functional modeling approach; whereas, the second, maximum likelihood, is a structural modeling approach.

2.4.1 SIMEX

As mentioned in Chapter 1 SIMEX is a simulation-based method of dealing with mismeasured covariates. Estimates are obtained by first inducing more bias in parameter estimates by adding measurement error using simulation, establishing a trend in this induced bias as a function of the induced error variance, and then extrapolating back to the case of no measurement error (Cook & Stefanski 1994). This method is suitable for use for additive or multiplicative measurement error models and if this model is correctly specified, will result in improved parameter estimates (Carroll et al. 2006). Since this method was originally developed for continuous covariates subject to error, the SIMEX algorithm and variance estimation will first be described for the case of measurement error in a continuous covariate and then for the situation involving a misclassified binary covariate.

Measurement error in continuous covariates:

Rather than considering a vector of mismeasured continuous covariates, for purposes of this description, consider the simpler case where $\boldsymbol{X}$ is one-dimensional. Assume that the classical error model given in (1.29) holds such that the random variable U representing measurement error is $N(0, \sigma_U^2)$. It follows that $W|X, \boldsymbol{Z} \sim N\left(\mu_{W|X,Z}, \sigma_U^2\right)$, where $\mu_{W|X,Z} = X$. It is important to note however, that normality is not required in order to apply SIMEX and in fact, this method can be easily extended to more complex error models (Carroll et al. 2006). Let $\boldsymbol{\theta}$ be the vector of the parameters of interest. Assume that σ_U^2 is known,

or at least a good estimate is available from supplementary data, and an estimator which is consistent in the absence of measurement error is available. Then, for a given dataset, the SIMEX algorithm would proceed as outlined in Carroll et al. (1995) which we now summarize.

Simulation Step

- Choose M constants, ν_m, $i = 1, 2, ..., M$, such that $0 = \nu_1 < \nu_2 < \cdots < \nu_M$. Common choices for these constants include $\{0, 0.5, 1, 1.5, 2\}$ (Cook & Stefanski 1994; Li & Lin 2003; Wang et al. 1998) and
  $\{0, 0.0625, 0.125, 0.25, 0.375, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$ (Greene & Cai 2004).

- Generate M sets of B datasets from the original, each time modifying the error-prone covariate values by including additional variability in the form of

  $$VAR_m\left(W|X, \boldsymbol{Z}\right) = \left(1 + \nu_m\right)\sigma_U^2$$

  for $m = 1, 2, ..., M$. All data will remain the same except for the revised w_i's which are generated from the original w_i's according to $W_{bi}(\nu_m) = W_i + \nu_m^{\frac{1}{2}} U_{bi}$, where $i = 1, 2, ..., n$, $b = 1, 2, ..., B$ and $m = 1, 2, ..., M$. The U_{bi}'s are mutually independent, independent of $\{\boldsymbol{Y}_i, X_i, W_i, \boldsymbol{Z}_i\}$ for all b and i and are generated from a $N(0, \sigma_U^2)$ distribution. Values for B that have been suggested in the literature include 50 (Li & Lin 2003), 100 (Cook & Stefanski 1994; Li & Lin 2003), and 200 (Greene & Cai 2004).

- For each of the $M \times B$ datasets, estimate $\boldsymbol{\theta}$ using $W_{bi}(\nu_m)$, for $i = 1, 2, ..., n$ based on a naive method, ignoring the measurement error, to obtain $\hat{\boldsymbol{\theta}}_b(\nu_m)$,

- Calculate the average of the naive parameter estimates for each of the M sets of data as

  $$\hat{\boldsymbol{\theta}}(\nu_m) = \frac{\sum_{b=1}^{B} \hat{\boldsymbol{\theta}}_b(\nu_m)}{B}, \tag{2.12}$$

  for $m = 1, 2, ..., M$. Wang et al. (1998) suggest using the median of the B estimates for $m = 1, 2, ..., M$ rather than the mean to calculate $\hat{\boldsymbol{\theta}}(\nu_m)$.

- Plot $(\nu_m, \hat{\boldsymbol{\theta}}(\nu_m))$ for $m = 1, 2, ..., M$. Wang et al. (1998) refer to this plot as a *partial bias plot* since the part of the relationship between the parameter estimates and $\nu < \nu_1$ (or equivalently, the error variance less than σ_U^2) is hidden.

Extrapolation Step

- Model each component of the estimated parameter vector, $\hat{\boldsymbol{\theta}}(\nu_m)$, as a function of ν. The shape of the partial bias plot will provide insight into the type of model which may be appropriate. Typical extrapolation functions that may be fit include:
  - linear models (Cook & Stefanski 1994; Li & Lin 2003),
  - quadratic models (Cook & Stefanski 1994; Li & Lin 2003; Wang et al. 1998),
  - rational linear extrapolant models such as $\theta(\nu) = a + \frac{b}{c+\nu}$ (Carroll et al. 2006; Li & Lin 2003), and
  - cubic models (Li & Lin 2003).

  These models could be fit using least squares regression methods (Carroll et al. 1996).

- The SIMEX estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{SIMEX}$, is obtained by extrapolating the fitted models back to the case where $\nu = -1$ for each component of $\boldsymbol{\theta}$. This represents the situation where X is error-free.

Carroll et al. (1996) develop asymptotic distribution theory for SIMEX estimators based on unbiased estimating equations. They show that they are unbiased when the extrapolation function is known and give an expression for the asymptotic variance, both for the case where the measurement error variance is known and for the case when it is estimated. Most of the time, however, the extrapolant is not known exactly; it is an approximation. Therefore, the resulting SIMEX estimator is only approximately consistent in general. By approximately consistent, we mean that it converges in probability to a constant that is only approximately equal to the true value of the parameter (Cook & Stefanski 1994).

When the measurement error variance is known, or a good estimate of it is available, Stefanski & Cook (1995) describe a simple method to obtain SIMEX standard errors that is related to Tukey's jackknife variance estimation. Carroll et al. (2006) indicate that this variance estimation procedure is valid for large samples and small measurement error. Let

- $T(\cdot)$ be an estimator for $\boldsymbol{\theta}$,
- $\hat{\boldsymbol{\theta}}_b(\nu) = T(\boldsymbol{Y}, W_b(\nu), \boldsymbol{Z})$, where $W_b(\nu)$ is the b^{th} W generated with measurement error variance given by $(1+\nu)\sigma_U^2$,
- $\tau_b^2(\nu) = VAR(\hat{\boldsymbol{\theta}}_b(\nu))$.

The following identity will be used to derive an estimate of the variance of the SIMEX estimator:

$$\hat{\boldsymbol{\theta}}(\nu) = E_{U|Y,W,Z}\left[\hat{\boldsymbol{\theta}}_b(\nu)\right]. \tag{2.13}$$

Here $\hat{\boldsymbol{\theta}}_b(\nu)$ depends on $\{\boldsymbol{Y}, W_b(\nu), \boldsymbol{Z}\}$, where $W_b(\nu)$ involves the random variables U_b. Therefore, this expectation is taken with respect to the distribution of U. The SIMEX estimator for $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_{SIMEX} = \hat{\boldsymbol{\theta}}(-1)$. From (2.13) and as indicated in Stefanski & Cook (1995), it follows that

$$VAR(\hat{\boldsymbol{\theta}}_b(\nu) - \hat{\boldsymbol{\theta}}(\nu)) \approx VAR(\hat{\boldsymbol{\theta}}_b(\nu)) - VAR(\hat{\boldsymbol{\theta}}(\nu)),$$

which will be used to approximate $VAR(\hat{\boldsymbol{\theta}}_{SIMEX})$:

$$\begin{aligned}
VAR(\hat{\boldsymbol{\theta}}_{SIMEX}) &= VAR(\hat{\boldsymbol{\theta}}(-1)) \\
&= \lim_{\nu \to -1} VAR(\hat{\boldsymbol{\theta}}(\nu)) \\
&= \lim_{\nu \to -1} \left\{ VAR(\hat{\boldsymbol{\theta}}(\nu)) + VAR(\hat{\boldsymbol{\theta}}_b(\nu)) - VAR(\hat{\boldsymbol{\theta}}_b(\nu)) \right\} \\
&= \lim_{\nu \to -1} \left\{ VAR(\hat{\boldsymbol{\theta}}_b(\nu)) - \left[VAR(\hat{\boldsymbol{\theta}}_b(\nu)) - VAR(\hat{\boldsymbol{\theta}}(\nu)) \right] \right\} \\
&\approx \lim_{\nu \to -1} \left\{ VAR(\hat{\boldsymbol{\theta}}_b(\nu)) - VAR\left[\hat{\boldsymbol{\theta}}_b(\nu) - \hat{\boldsymbol{\theta}}(\nu) \right] \right\}.
\end{aligned}$$

The first term in this expression represents sampling variability in $\hat{\boldsymbol{\theta}}(\nu)$ and can be estimated by $\boldsymbol{\tau}^2(\nu) = \frac{\sum_{b=1}^{B} \boldsymbol{\tau}_b^2(\nu)}{B}$, where $\boldsymbol{\tau}_b^2(\nu)$ is estimated by the naive model-based variance

of $\hat{\boldsymbol{\theta}}_b(\nu)$. The second term represents the variability due to the presence of measurement error. An unbiased estimator for this term is given by $\boldsymbol{s}^2(\nu)$, the sample covariance matrix calculated based on the B estimates of $\boldsymbol{\theta}$ for a given ν. The variance estimates for the SIMEX estimators can then be obtained by fitting a model to the components of the differences, $\boldsymbol{\tau}^2(\nu) - \boldsymbol{s}^2(\nu)$, and extrapolating back to $\nu = -1$. These variance estimates are referred to as the *SIMEX Information* when the naive model-based variances, $\boldsymbol{\tau}_b^2(\nu)$, are estimated by the inverse of the information matrix (Carroll et al. 2006).

When SIMEX is based on an estimate of the measurement error variance and the variation associated with this estimator is suspected to be substantial, bootstrap or jackknife resampling methods or a sandwich-type estimator based on unbiased estimating equation theory can be used to estimate the standard errors of the SIMEX estimators (Carroll et al. 2006). The resampling methods tend to be computationally burdensome due to the nested nature of the required resampling. The unbiased estimating equation approach requires additional programming, but less computation. A detailed description of this variance estimation approach is given in Carroll et al. (2006).

Misclassification in dichotomous covariates:

Küchenhoff et al. (2005) extended this approach to misclassified discrete covariates by introducing the "*Misclassification SIMEX*". Consider the situation where we are dealing with one misclassified covariate and for simplicity, assume that it can take on two values, 0 and 1, as outlined in the assumptions in the previous section. Let the misclassification be represented by matrix $\boldsymbol{\Pi}$,

$$\boldsymbol{\Pi} = \begin{pmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{pmatrix}, \tag{2.14}$$

where $\pi_{00} = P(W = 0|X = 0, \boldsymbol{Z})$ is the specificity and $\pi_{11} = P(W = 1|X = 1, \boldsymbol{Z})$ is the sensitivity. Let $\boldsymbol{\theta}$ be the vector of parameters of interest and assume both π_{00} and π_{11} are known or can be estimated from supplementary data. The naive estimator $\hat{\boldsymbol{\theta}}^*$, ignoring misclassification has a limit which depends on the degree of misclassification present which is characterized by the misclassification matrix, $\boldsymbol{\Pi}$. This limit is denoted by $\boldsymbol{\theta}^*(\boldsymbol{\theta}, \boldsymbol{\Pi})$. If

$\hat{\boldsymbol{\theta}}^*$ is a consistent estimator in the absence of misclassification, then $\boldsymbol{\theta}^*(\boldsymbol{\theta}, I_{2\times 2}) = \boldsymbol{\theta}$ where $I_{2\times 2}$ is a 2×2 identity matrix. For a given dataset, the Misclassification SIMEX algorithm would proceed as follows:

Simulation Step

- Choose M constants, ν_m, such that $0 = \nu_1 < \nu_2 < \cdots < \nu_M$.
- For each m, $m = 1, 2, ..., M$, generate B datasets from the original data, each time modifying the already misclassified W_i's by adding misclassification given by $\boldsymbol{\Pi}^{\nu_m}$ to generate new $W_{bi}(\nu_m)$'s for $i = 1, 2, ..., n$. Using matrix decomposition, $\boldsymbol{\Pi}^{\nu_m}$ can be rewritten as $\boldsymbol{\Pi}^{\nu_m} = \boldsymbol{E}\boldsymbol{\Lambda}^{\nu_m}\boldsymbol{E}$, where $\boldsymbol{\Lambda} = \text{diag}(e_1, e_2)$ and $\boldsymbol{E} = (\boldsymbol{E}_1, \boldsymbol{E}_2)$, with e_1, e_2, the eigenvalues of $\boldsymbol{\Pi}$, and $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$, their associated eigenvectors (Küchenhoff et al. 2005). To ensure $\boldsymbol{\Pi}^{\nu_m}$ is a well-defined misclassification matrix, $\text{Det}(\boldsymbol{\Pi}) = \pi_{00}+\pi_{11}-1$ must be greater than 0. This is true for $\pi_{00} > 0.5$ and $\pi_{11} > 0.5$. These values make sense as any sensitivities and specificities 0.5 or less would suggest W is not a very reasonable measurement of X (Küchenhoff et al. 2005). Küchenhoff et al. (2005) suggest using $B = 100$.
- For each of the $M \times B$ datasets, estimate $\boldsymbol{\theta}$ using $W_{bi}(\nu_m)$, for $i = 1, 2, ..., n$, based on the naive method which ignored misclassification to obtain $\hat{\boldsymbol{\theta}}_b(\nu_m)$.
- Calculate the average of the naive parameter estimates for each of the M sets of data as

$$\hat{\boldsymbol{\theta}}(\nu_m) = \frac{\sum_{b=1}^{B} \hat{\boldsymbol{\theta}}_b(\nu_m)}{B}, \tag{2.15}$$

  for $m = 1, 2, ..., M$.
- Plot $\left(\nu_m, \hat{\boldsymbol{\theta}}(\nu_m)\right)$ for $m = 1, 2, ..., M$.

Extrapolation Step

- Model each component of the estimated parameter vector, $\hat{\boldsymbol{\theta}}(\nu_m)$, as a function of ν. The shape of the partial bias plot will provide insight into the type of model which may be appropriate. Candidate extrapolant functions that may be fit include:

  - linear models,
  - quadratic models, and
  - log-linear models (Küchenhoff et al. 2005).

- The SIMEX estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{SIMEX}$, is obtained by extrapolating the fitted models back to the case where $\nu = -1$ for each component of $\boldsymbol{\theta}$. This approximates the situation where X is error-free. If the fitted model is a good approximation to the true underlying extrapolation function, this SIMEX procedure produces approximately consistent estimators.

Küchenhoff et al. (2005) also applied the approximate method (described in Stefanski & Cook (1995) for continuous measurement error) in their simulations for misclassified binary covariates, and it appeared to perform well. This method was outlined in further detail earlier in this section for the case of continuous X and W. It gives approximate standard errors in the case that the misclassification probabilities are known or are estimated reasonably well in the sense that the sampling variability is presumed to be negligible. Using this variance approximation, they conducted simulations based on logistic regression with a misclassified covariate and permitted the misclassification to be differential as well as nondifferential. They demonstrated that the SIMEX approach performs well in both situations. For the case where the misclassification probabilities are estimated, Küchenhoff et al. (2006) describe a variance estimator based on unbiased estimating equation approach that parallels the approach for continuous measurement error given in Carroll et al. (1996). Their approach assumes the availability of an independent validation study to estimate the misclassification matrix.

The SIMEX method can be readily extended to the situation where $\boldsymbol{X}$ is a vector for both continuous and binary covariates. In this case, vectors $\boldsymbol{W}_{bi}(\nu_m)$, rather than the scalars described above, can be generated in the simulation step for $b = 1, 2, ..., B$ and $i = 1, 2, ..., n$. The main advantage of SIMEX over other methods available to deal with mismeasured covariates is the relative ease of implementation. Since it involves repeated analysis of a dataset, existing software can be used to obtain estimates. Also, there is no need to specify a distribution for the underlying true covariate, X, and it involves a

built-in simulation study which demonstrates the effect of measurement error on parameter estimation for a given set of data. Since it is a generally applicable method, it is often useful in cases for which methodology has not been yet fully developed to deal with measurement error or misclassification (Cook & Stefanski 1994). Computation, however, may become burdensome if M and B are chosen to be large and estimation for the problem at hand is complicated. Disadvantages include the requirement that the error variance or misclassification matrix be known and the potential of obtaining inaccurate results due to poor extrapolation (Gustafson 2004). In some cases, the variance of SIMEX estimators can actually be much larger than that of the correct maximum likelihood estimators (Küchenhoff & Carroll 1997).

Further Remarks on SIMEX

Since SIMEX estimators are only approximately consistent in general, SIMEX may seem like a somewhat ad-hoc method of addressing covariate measurement error. However there is some theoretical support for its use. Provided the true extrapolation function is known, SIMEX estimators have been shown to be consistent and asymptotically normally distributed (Carroll et al. 1996; Küchenhoff et al. 2006). The difficulty lies in the identification of the true extrapolation function. Cook & Stefanski (1994) identified exact extrapolants for several models assuming normally distributed measurement errors. The SIMEX estimator in these settings is consistent. Consider the simple case of estimation of the variance of X based on W. The linear extrapolant is exact in this case. Extrapolation to $\nu = -1$ gives the methods-of-moments estimator (i.e. $\theta_{SIMEX} = s^2_W - \sigma^2_U$, where s^2_W is the sample variance and σ^2_U is the measurement error variance). They also showed that estimators for the regression coefficients in multiple linear regression models and log-linear mean models are consistently estimated by SIMEX estimators based on the rational linear extrapolant $\theta(\nu) = a + \frac{b}{c+\nu}$ (Cook & Stefanski 1994).

For general problems such as those considered in this chapter and the next, the true extrapolation function is unknown. Carroll et al. (2006) suggest that the rational linear or quadratic extrapolation functions are usually adequate for small measurement error. How-

ever, nonconvergence is often an issue with the rational linear extrapolant. Küchenhoff et al. (2005) conclude that the quadratic and exponential extrapolants are adequate in the case of misclassified covariates. Greene & Cai (2004) investigate SIMEX using linear, quadratic, cubic and rational linear extrapolation functions in the context of a marginal hazards model for multivariate failure time data. Based on their simulations, they observe that the quadratic, cubic and rational linear extrapolants perform pretty well; although as is the case for other problems, convergence problems can be encountered when fitting the rational linear model. Regardless of whether you are dealing with measurement error in continuous covariates or misclassified covariates, standard model building techniques and diagnostics (e.g. residual analyses) should be conducted to help with the selection of an extrapolant. Even if the model is carefully selected in this way, it is difficult to extrapolate to $\nu = -1$ based on data simulated for $\nu \in (0, 2]$. This is a disadvantage of the SIMEX approach.

The SIMEX approach as outlined earlier in this section treats the estimated measurement error variance or misclassification probabilities as known even though they are estimated based on supplementary data. The variance approximation of Stefanski & Cook (1995) assumes that the measurement error variance is known, so it may tend to underestimated standard errors, especially if the size of the supplementary dataset is small. In the case of misclassified covariates, Küchenhoff et al. (2005) suggest using a two-stage bootstrap procedure to estimate the variance of the SIMEX estimator in the case where $\mathbf{\Pi}$ is estimated. In the first stage, a bootstrap sample is drawn from a validation study to estimate $\mathbf{\Pi}$. Then, using this estimate, the above procedure is performed on a bootstrap sample from the primary data to obtain a SIMEX estimate. This is repeated a large number of times and the variance of the SIMEX estimator is estimated by the sample variance of the bootstrapped SIMEX estimates. The similar bootstrap procedure could be conducted in the case of continuous covariates to incorporate uncertainty in the measurement error variance estimator based on supplementary data. However, this approach can be computationally burdensome, so it is difficult to investigate the performance of these standard error estimators via simulation.

2.4.2 Correct Likelihood Approach

Misclassification in dichotomous covariates:

We will first discuss the likelihood formulation when the true covariate, X, is a dichotomous, one-dimensional variable. We are taking a structural approach so we will assume a distribution for X. Let $X_i|\boldsymbol{Z_i} \sim BIN\left(1, p(\boldsymbol{z}_i)\right)$, where $p(\boldsymbol{z}_i) = e^{\phi_0+\boldsymbol{\phi}_Z'\boldsymbol{z}_i} / \left[1 + e^{\phi_0+\boldsymbol{\phi}_Z'\boldsymbol{z}_i}\right]$ so when $\boldsymbol{\phi}_Z = \boldsymbol{0}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are uncorrelated. In addition, assume the X_i are conditionally independent given $\boldsymbol{Z_i}$ for all $i = 1, 2, ..., n$. Let $\boldsymbol{\theta}$ represent the unknown parameters to be estimated. Then the contribution to the likelihood by the i^{th} subject is

$$\mathcal{L}_i(\boldsymbol{\theta}) = f_{Y,W|Z}\left(\boldsymbol{y_i}, \boldsymbol{w_i}|\boldsymbol{z_i}\right). \tag{2.16}$$

Assuming that we are dealing with nondifferential misclassification and noninformative assessment times, this contribution becomes

$$\begin{aligned}\mathcal{L}_i(\boldsymbol{\theta}) &= f_{Y,W|Z}\left(\boldsymbol{y_i}, \boldsymbol{w_i}|\boldsymbol{z_i}\right) \\ &= \sum_x f_{Y|X,Z}\left(\boldsymbol{y_i}|x, \boldsymbol{z_i}; \boldsymbol{\theta}\right) f_{W|X,Z}\left(w_i|x, \boldsymbol{z_i}; \boldsymbol{\Pi}\right) f_{X|Z}\left(x|\boldsymbol{z_i}; \boldsymbol{\phi}\right),\end{aligned}$$

where

- $f_{Y|X,Z}\left(\boldsymbol{y_i}|x_i, \boldsymbol{z_i}\right)$ is based on either a Weibull regression model or a piecewise constant hazards model (Note that the distribution of Y, or the state path, conditional on X and Z can be thought of in terms of the distribution of T given X and Z. The probability of a transition occurring between assessment times c and d, such that $y_c = 1$ and $y_d = 2$, is $\mathcal{F}_T\left(c|x, \boldsymbol{z}; \boldsymbol{\theta}\right) - \mathcal{F}_T\left(d|x, \boldsymbol{z}; \boldsymbol{\theta}\right)$.),

- the W_i's are conditionally independent given X_i and $\boldsymbol{Z_i}$ and $f_{W|X,Z}$ is specified by the misclassification probabilities as

$$f_{W|X,Z}\left(w_i|x_i, \boldsymbol{z_i}; \boldsymbol{\Pi}\right) = \left[\pi_{10}^{w_i}(1-\pi_{10})^{(1-w_i)}\right]^{(1-x_i)} \left[\pi_{01}^{(1-w_i)}(1-\pi_{01})^{w_i}\right]^{x_i}, \text{ and}$$

- $f_{X|Z}\left(x_i|\boldsymbol{z_i}; \boldsymbol{\phi}\right) \propto p(\boldsymbol{z}_i)^{x_i}\left(1 - p(\boldsymbol{z}_i)\right)^{1-x_i}$.

Then the full likelihood function is given by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\sum_{x=0}^{1} f_{Y|X,Z}\left(\boldsymbol{y_i}|x,\boldsymbol{z_i}\right) f_{W|X,Z}\left(w_i|x,\boldsymbol{z_i}\right) f_{X|Z}\left(x|\boldsymbol{z_i}\right) \\
&= \prod_{i=1}^{n}\sum_{x=0}^{1} \left[\mathcal{F}_T\left(c_i|x,\boldsymbol{z_i};\boldsymbol{\theta}\right) - \mathcal{F}_T\left(d_i|x,\boldsymbol{z_i};\boldsymbol{\theta}\right)\right]^{\delta_i} \left[\mathcal{F}_T\left(c_i|x,\boldsymbol{z_i};\boldsymbol{\theta}\right)\right]^{1-\delta_i} \\
&\quad \left[\pi_{10}^{w_i}(1-\pi_{10})^{(1-w_i)}\right]^{(1-x)} \left[\pi_{01}^{(1-w_i)}(1-\pi_{01})^{w_i}\right]^{x} \frac{\left[\mathrm{e}^{\phi_0+\boldsymbol{\phi}_Z'\boldsymbol{z}_i}\right]^{x}}{1+\mathrm{e}^{\phi_0+\boldsymbol{\phi}_Z'\boldsymbol{z}_i}}. \qquad (2.17)
\end{aligned}$$

Inference for $\boldsymbol{\theta}$ can then be conducted based on maximization of the above likelihood function with respect to the unknown parameters. This can be readily extended to categorical covariates with more than two levels.

Measurement error in continuous covariates:

Now we consider the likelihood formulation for continuous covariates, $\boldsymbol{X}$. In addition to the assumptions previously outlined, since this maximum likelihood approach is based on structural modeling, we must make assumptions regarding the distribution of the true underlying covariates, $\boldsymbol{X}$. For the sake of this illustration, we will consider one-dimensional X and allow its distribution to depend on $\boldsymbol{Z}$ as follows: $X_i|\boldsymbol{Z}_i \sim N(\mu_{X|Z}, \sigma^2_{X|Z})$, where $\mu_{X|Z}$ and $\sigma_{X|Z}$ are known (or can be readily estimated using supplementary data). We will also assume that the X_i are conditionally independent given $\boldsymbol{Z}_i$ for $i = 1, 2, ..., n$. To construct the likelihood, we need to consider the observed data. The contribution to the likelihood function from subject i would be as follows:

$$\begin{aligned}
\mathcal{L}_i(\boldsymbol{\theta}) &= f_{Y,W|Z}\left(\boldsymbol{y_i},\boldsymbol{w_i}|\boldsymbol{z_i}\right) \\
&= \int_{-\infty}^{\infty} f_{Y|W,X,Z}\left(\boldsymbol{y_i}|\boldsymbol{w_i},x,\boldsymbol{z_i};\boldsymbol{\theta}\right) f_{W|X,Z}\left(\boldsymbol{w_i}|x,\boldsymbol{z_i};\boldsymbol{\theta}_{W|X,Z}\right) f_{X|Z}\left(x|\boldsymbol{z_i};\boldsymbol{\theta}_{X|Z}\right) dx \\
&= \int_{-\infty}^{\infty} f_{Y|X,Z}\left(\boldsymbol{y_i}|x,\boldsymbol{z_i};\boldsymbol{\theta}\right) f_{W|X,Z}\left(\boldsymbol{w_i}|x,\boldsymbol{z_i};\boldsymbol{\theta}_{W|X,Z}\right) f_{X|Z}\left(x|\boldsymbol{z_i};\boldsymbol{\theta}_{X|Z}\right) dx.
\end{aligned}$$

The third line follows from the assumption of nondifferential error. The functions in the last line above are all known, at least up to some unknown parameters, due to the assumptions presented earlier. The functions appearing in the above likelihood are:

- $f_{Y|X,Z}\left(\boldsymbol{y_i}|x,\boldsymbol{z_i};\boldsymbol{\theta}\right)$, which is based on a Weibull regression model, or a piecewise constant baseline hazards model that is a function of unknown parameters $\boldsymbol{\theta} = (\lambda_{0k}, \boldsymbol{\beta}_{xk}, \boldsymbol{\beta}_{zk})$ for $k = 1, 2, ..., K$,

- $f_{W|X,Z}\left(\boldsymbol{w_i}|x,\boldsymbol{z_i};\boldsymbol{\theta}_{W|X,Z}\right)$, which is given by the classical error model, (1.29), with the measurement error variance assumed known or estimated via supplementary data, and

- $f_{X|Z}\left(x|\boldsymbol{z_i};\boldsymbol{\theta}_{X|Z}\right)$, which is assumed known since we have taken a structural approach and assumed $X_i|\boldsymbol{Z}_i \sim N\left(\mu_{X|Z}, \sigma^2_{X|Z}\right)$; information regarding this distribution can be obtained from prior knowledge or data collected on X and $\boldsymbol{Z}$ in the current investigation.

Then the full likelihood function is

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{\theta}\right) &= \prod_{i=1}^{n}\int_{-\infty}^{\infty} f_{Y|X,Z}\left(\boldsymbol{y_i}|x,\boldsymbol{z_i};\boldsymbol{\theta}\right) f_{W|X,Z}\left(\boldsymbol{w_i}|x,\boldsymbol{z_i};\boldsymbol{\theta}_{W|X,Z}\right) f_{X|Z}\left(x|\boldsymbol{z_i};\boldsymbol{\theta}_{X|Z}\right) dx \\
&= \prod_{i=1}^{n}\int_{-\infty}^{\infty} \Big\{ \left[\mathcal{F}_T\left(c_i|x,\boldsymbol{z}_i;\boldsymbol{\theta}\right) - \mathcal{F}_T\left(d_i|x,\boldsymbol{z}_i;\boldsymbol{\theta}\right)\right]^{\delta_i} \left[\mathcal{F}_T\left(c_i|x,\boldsymbol{z}_i;\boldsymbol{\theta}\right)\right]^{1-\delta_i} \\
&\quad \frac{1}{\sqrt{2\pi}\sigma_U} \mathrm{e}^{\frac{-(w_i-x)^2}{2\sigma_U^2}} \frac{1}{\sqrt{2\pi}\sigma_{X|Z}} \mathrm{e}^{\frac{-\left(x-\mu_{X|Z}\right)^2}{2\sigma^2_{X|Z}}} \Big\} dx. \qquad (2.18)
\end{aligned}$$

Due to the potential complexity of the integrand, the integrals in the above likelihood function can be approximated numerically. One strategy involves the use of numerical techniques such as *Monte Carlo Methods*. Let N be a large prespecified number. Then this method proceeds in the following way. Given $\boldsymbol{z}_i$, for each i, simulate N values of x_i from the $N\left(\mu_{X|Z}, \sigma^2_{X|Z}\right)$ distribution to obtain $\left(x_i^{(1)}, x_i^{(2)}, ..., x_i^{(N)}\right)$. Then

$$\begin{aligned}
\hat{\mathcal{L}}_i &\approx \hat{E}_{X|Z}\left(f_{Y|X,Z}\left(\boldsymbol{y_i}|x_i,\boldsymbol{z_i};\boldsymbol{\theta}\right) f_{W|X,Z}\left(\boldsymbol{w_i}|x_i,\boldsymbol{z_i};\boldsymbol{\theta}_{W|X,Z}\right)\right) \\
&= \frac{1}{N}\sum_{r=1}^{N} f_{Y|X,Z}\left(\boldsymbol{y_i}|x_i^{(r)},\boldsymbol{z_i};\boldsymbol{\theta}\right) f_{W|X,Z}\left(\boldsymbol{w_i}|x_i^{(r)},\boldsymbol{z_i};\boldsymbol{\theta}_{W|X,Z}\right),
\end{aligned}$$

and inference regarding $\boldsymbol{\theta}$ can be conducted based on $\hat{\mathcal{L}} = \prod_{i=1}^{n}\hat{\mathcal{L}}_i$.

Another way to numerically approximate the integrals in (2.17) is via *Gaussian Quadrature*. For a given integer N, we can determine abscissas, x_j and a set of weights, w_j such that

$$\int_a^b W(x) f(x) dx \approx \sum_{j=0}^{N-1} w_j f(x_j) \tag{2.19}$$

For low dimensional $\boldsymbol{X}$, this approach is a recommended numerical integration technique by Evans & Swartz (2000). Throughout this thesis when X is continuous, $X|\boldsymbol{Z}$ is assumed to follow a normal distribution. Therefore, Gauss-Hermite quadrature will be used. In this case, the integral to be approximated has the form $\int_{-\infty}^{\infty} f(x) \exp(-x^2) dx$. The abscissas and weights can be determined based on a recurrence relation involving Hermite polynomials which have the form $H_{j+1} = 2xH_j - 2jH_{j-1}$. These polynomials are the solutions to the differential equation $y'' - 2xy' + 2n_H y = 0$, $n_H = 0, 1, 2, ...$ (Press et al. 2002). This numerical integration approach was used in the simulation studies and the application presented later in this chapter. The maximum likelihood approach would proceed in a similar manner for higher dimensional $\boldsymbol{X}$ or for more complex distributions for the error and the true underlying covariates. However, the numerical integration approach would need to be revisited and revised accordingly.

2.4.3 Estimation of Mismeasurement and Covariate Distribution Parameters

To implement both the SIMEX and the correct likelihood approaches, supplementary data is required to estimate parameters associated with distributions other than those for $\boldsymbol{Y}|X, \boldsymbol{Z}$, the distribution of interest. Internal validation data, where X is recorded in addition to W, for a subset of the study participants is ideal. It provides information on the structure of the error distribution and often leads to greater precision in estimation (Carroll et al. 2006). However, reliability data and external validation data can still be used to collect information on the error distribution, but the assumption of "transportability" must be made when using external data. A model is transportable if it and its associated parameters can be applied in the context of another problem without introducing bias

(Carroll et al. 2006). It is common in practice to assume that the same classical error model holds across populations. However, it is important to keep in mind that using a model which is not transportable in an errors-in-variables analysis may actually introduce bias (Carroll et al. 2006). Measurement error and misclassification can have a substantial impact on parameter estimation. Therefore, if the associated parameters are assumed or estimated via external data, it is good practice to augment the analysis with a sensitivity study to demonstrate uncertainty of departures from the assumed values in the estimates and investigate the impact of departures from the assumed values on parameter estimation (Aitkin & Rocci 2002). For SIMEX, we need to estimate the parameters associated with the mismeasurement distribution (i.e. $W|X$, or possibly $W|X, \boldsymbol{Z}$ if the mismeasurement distribution also depends on the error-free covariates) and for the correct likelihood approach, we also require estimates of the parameters of the conditional covariate distribution (i.e. $X|\boldsymbol{Z}$).

First consider the case where X, and therefore W, are one-dimensional binary variables. For both the SIMEX and maximum likelihood approaches, π_{01} and π_{10} must be estimated. With validation data, maximizing the likelihood function

$$\mathcal{L}(\pi_{01}, \pi_{10}) = \pi_{01}^{n_{01}} (1 - \pi_{01})^{n_1 - n_{01}} \pi_{10}^{n_{10}} (1 - \pi_{10})^{n_0 - n_{10}}$$

results in the maximum likelihood estimates $\hat{\pi}_{01} = n_{01}/n_1$ and $\hat{\pi}_{10} = n_{10}/n_0$, where n_{01} is the number of subjects in the validation study with $X = 1$ and $W = 0$, n_{10} is the number of subjects in the study with $X = 0$ and $W = 1$, and n_1 and n_0 are the number of subjects with $X = 1$ and $X = 0$, respectively.

When X is not observed, the misclassification probabilities can also be estimated with reliability data by latent class analysis (Goodman 1974). If there were r_i replicated observations for subject i, a likelihood contribution from the i^{th} subject would be based on $\sum_{X_i} P(X_i)P(\boldsymbol{W}_i|X_i)$ and given by

$$\mathcal{L}_i(\pi, \pi_{01}, \pi_{10}) = \pi \prod_{j=1}^{r_i} \left\{ (1 - \pi_{01})^{w_{ij}} \pi_{01}^{1 - w_{ij}} \right\} + (1 - \pi) \prod_{j=1}^{r_i} \left\{ (1 - \pi_{10})^{1 - w_{ij}} \pi_{10}^{w_{ij}} \right\}, \quad (2.20)$$

where $\pi = P(X_i = 1)$.

We also need to specify a distribution for $X|\boldsymbol{Z}$ to proceed with the maximum likelihood approach. If external validation data are available with $\boldsymbol{Z}$ measured in addition to X and W, a logistic regression of X on $\boldsymbol{Z}$ can be performed to estimate this distribution.

If there are internal validation data available and $\Delta_i = 1$ when subject i is in the validation study, the likelihood could be specified as follows:

$$\mathcal{L}_i(\boldsymbol{\theta}) = \begin{cases} \sum_{x=0}^{1} f_{Y|X,Z}(\boldsymbol{y}_i|x, \boldsymbol{z}_i; \boldsymbol{\theta}) f_{W|X}(w_i|x; \boldsymbol{\Phi}) f_{X|Z}(x|\boldsymbol{z}_i; \boldsymbol{\Psi}), & \Delta_i = 0 \\ f_{Y|X,Z}(\boldsymbol{y}_i|x_i, \boldsymbol{z}_i; \boldsymbol{\theta}) f_{W|X}(w_i|x_i; \boldsymbol{\Phi}), & \Delta_i = 1 \end{cases}, \tag{2.21}$$

where Φ and Ψ are the parameters associated with the measurement error and conditional covariate distributions, respectively. Then the misclassification and conditional covariate distribution parameters can be estimated along with the parameters of interest. With reliability data, if the misclassification probabilities do not depend on $\boldsymbol{Z}$, an estimate of the $X|\boldsymbol{Z}$ distribution could be obtained by the logistic regression of W on $\boldsymbol{Z}$. If the assumption is made that X does not depend on $\boldsymbol{Z}$, an estimate of $\pi = P(X = 1)$ could be obtained directly from (2.20).

Now consider the case where X, and therefore W, are one-dimensional continuous variables. Reliability data, or data consisting of repeated measurements of W, can be used to estimate σ_U^2 when the classical error model, (1.29), is appropriate. Suppose there are n_r subjects in the reliability study and there are r_i replicate measurements of W_{ij}, $j = 1, 2, ..., r_i$, for each subject, i. Based on a component of variance analysis, this measurement error variance can then be estimated by

$$\sigma_U^2 = \frac{\sum_{i=1}^{n_r} \sum_{j=1}^{r_i} (w_{ij} - \bar{w}_{i\cdot})^2}{\sum_{i=1}^{n_r} (r_i - 1)},$$

where $\bar{w}_{i\cdot} = \frac{1}{r_i} \sum_{j=1}^{r_i} w_{ij}$. Validation data, or data containing measurements of both X and W on the same subjects, could also be used to estimate σ_U^2. A simple linear regression analysis could be used both to verify the reasonableness of the classical error model assumption as well as to provide an estimate for σ_U^2 by the estimated residual variance.

If the measurement error distribution depends on $\boldsymbol{Z}$ in addition to X these variables can be included in the regression analysis. However, if the validation data is external to the primary data, it would have to contain measurements on the error-free covariates, $\boldsymbol{Z}$ in addition to those on X and Z.

For the SIMEX approach, estimation of σ_U^2 would follow in a similar manner regardless of whether the supplementary data arose from external data or were included in the primary data. If the variation in these estimates is not negligible, the variability associated with these estimates can then be incorporated in the SIMEX variance estimation based on resampling methods or unbiased estimating equations (Carroll et al. 2006). When no supplementary data are available to characterize the measurement error distribution, a sensitivity analysis could be performed to investigate the impact of varying degrees of measurement error (Li & Lin 2003). For the correct likelihood approach, σ_U^2 could be estimated in the manner outlined above if we were dealing with external supplementary data. However, with internal supplementary data the likelihood function can be expressed in terms of σ_U^2 and it can be estimated along with the other parameters rather than simply imputing an estimate into the likelihood function.

Parameter estimates associated with the distribution of $X|\boldsymbol{Z}$ are also needed to implement the likelihood approach. Again, ideally the data would consist of an internal validation subset. With internal validation data and with $\Delta_i = 1$ when subject i is in the validation study, the contribution of the i^{th} subject to likelihood function would be

$$\mathcal{L}_i(\boldsymbol{\theta}) = \begin{cases} \int_{-\infty}^{\infty} f_{Y|X,Z}\left(\boldsymbol{y}_i|x, \boldsymbol{z}_i; \boldsymbol{\theta}\right) f_{W|X}\left(w_i|x; \boldsymbol{\Phi}\right) f_{X|Z}\left(x|\boldsymbol{z}_i; \boldsymbol{\Psi}\right) dx, & \Delta_i = 0 \\ f_{Y|X,Z}\left(\boldsymbol{y}_i|x_i, \boldsymbol{z}_i; \boldsymbol{\theta}\right) f_{W|X}\left(w_i|x_i; \boldsymbol{\Phi}\right) & \Delta_i = 1 \end{cases} \tag{2.22}$$

We can also obtain information about the distribution of $X|\boldsymbol{Z}$ via reliability data or external validation data. With an external validation subset that includes measurements on $\boldsymbol{Z}$ in addition to X and W, a simple linear regression of X on $\boldsymbol{Z}$ could be used to estimate the conditional covariate distribution and the estimated parameters can be used in the likelihood function for $\boldsymbol{Y}, W|Z$. With reliability data, either external or internal, we have no measurements of X, just repeated measurements of W. However, if we assume the classical error model (1.29) is appropriate, we can still estimate the measurement error

variance as above. Then, considering the case where there are no error-free covariates, $\boldsymbol{Z}$, measured or the case where the measurement error and the X distributions do not depend on $\boldsymbol{Z}$, it follows from $W = X + U$, $U \sim N(0, \sigma_U^2)$ that $\mu_X = \mu_W$ and $\sigma_X^2 = \sigma_W^2 - \sigma_U^2$, so estimates could be obtained from $\hat{\mu}_X = \hat{\mu}_W$ and $\hat{\sigma}_X^2 = \hat{\sigma}_W^2 - \hat{\sigma}_U^2$.

2.5 Simulation Study

The objective of these simulations is to compare the performance of the naive and correct estimation approaches in the presence of measurement error and misclassification. Two-state models (Figure 2.2) were investigated. Values for the hazard function parameters were selected so that the simulations represent situations encountered in practice and so that they would be consistent with those used in the next chapter on three state models. In these simulations, W is the mismeasured version of X that will be used to fit models and Z is a perfectly measured covariate. Parameters associated with $\lambda_T(t|x, z; \boldsymbol{\theta})$ are denoted by $\boldsymbol{\theta} = (\rho, \kappa, \beta_X, \beta_Z)$. All simulations were conducted in SAS using PROC NLP and PROC IML.

2.5.1 Binary Covariates

DATA GENERATION

Data were generated based on the true models and the joint distribution of (X, W) as follows:

- Number of datasets: $N = 500$,
- Number of subjects per dataset: $n = 500$,
- Years of follow-up: τ was selected such that the probability of transition to state 2 from state 1 by time τ (i.e. $P_{1,2}(\tau|\, X, Z)$) was at least 0.8 based on all possible values of (X, W),

- Average number of follow-ups: $\mu = 5$ (20 was also investigated for a small number of parameter configurations),
- Baseline hazards: $\rho = 0.2$ and $\kappa = 0.5, 1, 2$,
- Covariate effects: $e^{\beta_X} = 1.25, 2$ and $e^{\beta_Z} = 1.25$, and
- SIMEX parameters: $M = 5$ with $\{\nu_1, \nu_2, \nu_3, \nu_4, \nu_5\} = \{0, 0.5, 1, 1.5, 2\}$ and $B = 100$.

For each subject, first the number of follow-up times were generated as $m_i \sim POI(\mu)$. The assessment times, $u_{ij}, j = 1, 2, ..., m_i$ were then generated from m_i independent $UNIF(0, \tau)$ random variables. The transition times for each individual were simulated by generating values of $T_i \sim EXP(\lambda_T(t|x_i, z_i, \boldsymbol{\theta}))$. The transition times were then compared to the assessment times. If the transition time was interval-censored and fell between $u_{i,j-1}$ and u_{ij} for some $j = 1, ..., m_i$, then $c_i = u_{i,j-1}$, $d_i = u_{ij}$ and $\delta_i = 1$. Otherwise, if the transition time was right-censored, $c_i = u_{im_i}$ and $\delta_i = 0$.

Misclassification was characterized by the probabilities, $\pi_{01} = 1 - \pi_{11}$ and $\pi_{10} = 1 - \pi_{00}$, or equivalently, by π_{00} and π_{11} (i.e. specificity and sensitivity). Covariate values were generated by the following steps:

- $Z \sim BIN(1, p_Z)$, with $p_Z = 0.5$.
- $X|Z \sim BIN(1, \frac{e^{\xi_0 + \xi_Z Z}}{1 + e^{\xi_0 + \xi_Z Z}})$, with $\xi_0 = 0$ for a 50% baseline probability X=1, and $\xi_Z = -\log(2), \log(2)$, which represent negative and positive effects of Z on X.
- $\pi_{11} = P(W = 1|X = 1) = 0.7, 1$ (sensitivity), and
- $\pi_{00} = P(W = 0|X = 0) = 0.7, 0.9, 1$ (specificity).
  These values were selected to represent minor to severe misclassification. The parameter configurations also allow us to investigate the situations when only false negatives are possible ($\pi_{11} = 1$ and $\pi_{00} < 1$) and only false positive are possible ($\pi_{00} = 1$ and $\pi_{01} < 1$). As is clear from these expressions, in these simulations, we are assuming the misclassification probabilities do not depend on Z.

Two validation samples (one of size 50 and one of size 200) were randomly selected from the $n = 500$ subjects to estimate the misclassification probabilities and for the correct maximum likelihood approach, the $X|Z$ distribution.

ESTIMATION

Estimates of π_{01} and π_{10} were obtained by fitting a logistic regression of W on X based on the validation data:

$$\hat{\pi}_{01} = \frac{1}{1 + \mathrm{e}^{\hat{\phi}_0 + \hat{\phi}_X}} \text{ and } \hat{\pi}_{10} = \frac{\mathrm{e}^{\hat{\phi}_0}}{1 + \mathrm{e}^{\hat{\phi}_0}}.$$

These estimates of the misclassification probabilities were used to generate inflated misclassification in the SIMEX approach and were used in the likelihood function for the maximum likelihood approach (i.e. ignoring the sampling variability). A logistic regression of X on Z provided estimates of ξ_0 and ξ_Z to provide an estimate for $P(X = 1|z)$ also for use in the correct likelihood function:

$$\hat{p}_{X|Z} = \frac{\mathrm{e}^{\hat{\xi}_0 + \hat{\xi}_Z Z}}{1 + \mathrm{e}^{\hat{\xi}_0 + \hat{\xi}_Z Z}}.$$

Both Weibull models and piecewise constant baseline hazard (PCBH) models were fit to the data.

SIMEX involved repeated estimation based on the naive likelihood function. For a multiple of the original misclassification given by ν_m, $m = 2, 3, 4, 5$, $B = 100$ revised $\boldsymbol{w}_b$'s were generated and each time, $\hat{\boldsymbol{\theta}}_b(\nu_m) = \left(\hat{\rho}_{0b}(\nu_m), \hat{\kappa}_{0b}(\nu_m), \hat{\beta}_{Xb}(\nu_m), \hat{\beta}_{Zb}(\nu_m)\right)'$ was obtained by maximizing the following likelihood function:

$$\mathcal{L}_{naive}\left(\boldsymbol{\theta}(\nu_m)\right) = \prod_{i=1}^{500} \left[\mathcal{F}_T\left(c_i|w_i, z_i; \boldsymbol{\theta}(\nu_m)\right) - \mathcal{F}_T\left(d_i|w_i, z_i; \boldsymbol{\theta}(\nu_m)\right)\right]^{\delta_i} \left[\mathcal{F}_T\left(c_i|w_i, z_i; \boldsymbol{\theta}(\nu_m)\right)\right]^{1-\delta_i}. \tag{2.23}$$

Since ρ and κ must be larger than 0, they were reparametrized as $\rho = \mathrm{e}^r$ and $\kappa = \mathrm{e}^k$ to avoid imposing constraints in the optimization procedure. Then the log-likelihood function was maximized with respect to $\boldsymbol{\theta} = (r, k, \beta_X, \beta_Z)$. By the invariance property of maximum likelihood estimators, estimates of ρ and κ were obtained by $\hat{\rho} = \mathrm{e}^{\hat{r}}$ and $\hat{\kappa} = \mathrm{e}^{\hat{k}}$. Their

respective variances were estimated by

$$\begin{aligned}\widehat{\mathrm{VAR}}\left(\rho\right) &= \left(\mathrm{e}^{\hat{r}},\mathrm{e}^{\hat{k}},0,0\right)^{t} I^{-1}\left(\hat{\theta}\right)\left(\mathrm{e}^{\hat{r}},\mathrm{e}^{\hat{k}},0,0\right)_{[1,1]}, \text{ and}\\ \widehat{\mathrm{VAR}}\left(\kappa\right) &= \left(\mathrm{e}^{\hat{r}},\mathrm{e}^{\hat{k}},0,0\right)^{t} I^{-1}\left(\hat{\theta}\right)\left(\mathrm{e}^{\hat{r}},\mathrm{e}^{\hat{k}},0,0\right)_{[2,2]},\end{aligned}$$

where $I^{-1}\left(\hat{\theta}\right)$ was the inverse of the observed information function evaluated at the maximum likelihood estimate, $\hat{\boldsymbol{\theta}} = \left(\hat{r},\hat{k},\hat{\beta}_X,\hat{\beta}_Z\right)'$. Although SAS's PROC NLP was used to conduct the maximum likelihood estimation and the likelihood function was coded, existing software (such as PROC LIFEREG in SAS) could have been employed to maximize (2.23).

At each ν_m, $\hat{\boldsymbol{\theta}}(\nu_m)$ was obtained by taking the average of the $B = 100$ parameter estimates. The estimate $\hat{\boldsymbol{\theta}}(\nu_1)$ is simply the original naive maximum likelihood estimate. Then, a model was fit to these five values and the SIMEX estimates were obtained by extrapolating back to the case where $\nu = -1$ based on this model. The simple variance approximation approach as described in Stefanski & Cook (1995) for continuous measurement error and used in Küchenhoff et al. (2005) for misclassification was applied here. Therefore, variance estimates for the SIMEX estimators were obtained by first fitting a model to the differences, $\tau^2\left(\nu_m\right) - s^2\left(\nu_m\right)$, $m = 1, ..., 5$, where $\tau^2\left(\nu_m\right)$ is the average of the $B = 100$ model-based variance estimates at each ν_m (which were based on the inverse of the information matrix here) and $s^2\left(\nu_m\right)$ is the sample variance of the $B = 100$ parameter estimates at ν_m. The SIMEX variance estimates were then obtained by extrapolating this relationship back to $\nu = -1$. As in Küchenhoff et al. (2005), quadratic ($\theta = a + b\nu + c\nu^2$) and exponential ($\theta = a\mathrm{e}^{b\nu}$) extrapolation functions were considered and fit using least squares in SAS (PROC REG and PROC NLIN, respectively). For the purposes of these simulations the same extrapolation function was used to obtain the SIMEX parameter and variance estimates. However, there is no requirement that the parameter and variance estimate extrapolants be the same. In practice, extrapolant function selection would not be automated in this way. Usual model building techniques would be used and diagnostics based on residuals would provide information regarding the adequacy of the models. It is difficult to implement this in simulation studies. However, in an attempt to automate

this model building process, two other approaches were considered in the simulation based on quadratic and exponential functions. First, the optimal model of the two based on adjusted R^2 (i.e. $R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$) was selected to estimate the parameters and the variances. Second, since both extrapolation models appeared to perform reasonably well, the average of the estimates arising from the two models was considered.

The maximum likelihood approach accommodating misclassification was based on the following likelihood function (see (2.17)).

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\sum_{x=0}^{1}\left[\mathcal{F}_T\left(c_i|x,z_i;\boldsymbol{\theta}\right)-\mathcal{F}_T\left(d_i|x,z_i;\boldsymbol{\theta}\right)\right]^{\delta_i}\left[\mathcal{F}_T\left(c_i|x,z_i;\boldsymbol{\theta}\right)\right]^{1-\delta_i} \\ &\quad \cdot \left[\frac{\left(\mathrm{e}^{\hat{\phi}_0+\hat{\phi}_X x}\right)^{w_i}}{1+\mathrm{e}^{\hat{\phi}_0+\hat{\phi}_X x}}\right]\left[\frac{\left(\mathrm{e}^{\hat{\xi}_0+\hat{\xi}_Z z_i}\right)^{x}}{1+e^{\hat{\xi}_0+\hat{\xi}_Z z_i}}\right].\end{aligned}$$

Again, as in the SIMEX case, ρ and κ were reparameterized as $\rho = \mathrm{e}^r$ and $\kappa = \mathrm{e}^k$ and the log-likelihood function was maximized to obtain estimates for $\boldsymbol{\theta} = (r, k, \beta_X, \beta_Z)$. Even though existing software could have been used for the SIMEX analyses, for the sake of consistency in implementation, for both the SIMEX and the maximum likelihood approaches, the objective functions were maximized based on a quasi-Newton algorithm using PROC NLP in SAS. Quasi-Newton approaches require computation of the first derivative of the log-likelihood function, but not the second derivative, which is approximated. This reduces the computing required compared to Newton's method where the second derivatives must be computed in addition to the first derivatives. Typically though, quasi-Newton approaches require more iterations than Newton's method. In developing the code used to conduct the simulations throughout this thesis, both optimization approaches were tried and it was found that the quasi-Newton approach seemed to be much more efficient than Newton's method in terms of computation time. All simulations are based on the default QUANEW procedure in SAS's PROC NLP and initial values for the parameters were randomly generated by PROC NLP. The default quasi-Newton procedure in SAS is based on a dual quasi-Newton algorithm that updates the Cholesky factor of the approximate Hessian based on the BFGS update (Broyden 1969; Fletcher 1970; Goldfarb 1970; Shanno 1970). According to the SAS documentation (SAS 9.1.3 OnlineDoc 2006), the default line-search

method is based on quadratic interpolation and cubic extrapolation functions which are used to obtain a step length that adheres to Goldstein's conditions. Further details are available in the SAS 9.1.3 online documentation, Martinez (2000) and Schoenberg (2001).

Fitting the models involving piecewise constant baseline hazards followed in a similar way. However, cut-points had to be selected before the log-likelihood could be maximized. Models with four pieces were considered in these simulations. The cut-points were chosen to be the quartiles of the true underlying distribution. In practice, these cut-points could be selected based on empirical distribution quantiles or they could be equally spaced over the length of the study. Representative results are displayed in Tables 2.1 to 2.4. These tables summarize the results corresponding to effects on X and Z compare results based on the Weibull model and the piecewise constant baseline hazards (PCBH) model. In practice, it is usually the covariate effects that are of interest rather than the baseline hazards. "Sample I" refers to estimation with supplementary data in the form of a validation sample of size 50 while "Sample II" refers to a validation sample of size 200. The "known" results under the correct maximum likelihood approach were based on full knowledge of the misclassification and conditional covariate distributions. This is meant to represent the best case scenario when X is unobserved for all subjects in the study. Comparison of these results to those using validation data can provide an indication of how well the method performs when parameters associated with those distributions need to be estimated in addition to the model parameters of interest. SIMEX results based on a small validation sample and using quadratic and exponential extrapolation functions are also presented in the tables.

DISCUSSION

There did not seem to be any major problems with convergence (convergence rates ranged from about 97% to 100% of the simulation replications). Upon examination of the tabulated results for the parameter configurations investigated, the following general observations can be made.

- For both models the naive maximum likelihood biases and coverage probabilities (based on the model-based standard errors) exhibit poorer performance as the mis-

Table 2.1: *Empirical performance of estimators of the regression parameters associated with binary X and Z; Number of assessments are POI* (5)*;* $\rho = 0.2$, $\kappa = 1$, $\beta_X = \beta_Z = \log(1.25)$, $\pi_{11} = 0.7$, $p_z = 0.5$ *and logit* $\left(p_{x|z}\right) = \log(2)\, z$.

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.1350	-0.1356	-0.1064	-0.0924	0.0320	0.0337	0.0270	0.0239
	SE_1	0.1042	0.1043	0.1043	0.1046	0.1040	0.1042	0.1043	0.1044
	SE_2	0.1041	0.1043	0.1082	0.1057	0.1018	0.1017	0.1021	0.1035
	ECP	0.7460	0.7470	0.8533	0.8567	0.9460	0.9458	0.9500	0.9517
Likelihood									
Known	Bias	0.0073	0.0078	0.0034	-0.0003	0.0043	0.0047	0.0004	0.0005
	SE_1	0.2749	0.2789	0.1794	0.1979	0.1144	0.1146	0.1084	0.1084
	SE_2	0.2830	0.2826	0.1801	0.1821	0.1135	0.1137	0.1068	0.1078
	ECP	0.9499	0.9580	0.9525	0.9497	0.9579	0.9620	0.9576	0.9549
Sample I	Bias	0.0268	0.0356	0.0807	0.0944	0.0019	-0.0023	-0.0107	0.0944
	SE_1	0.3078	0.3295	0.2439	0.2565	0.1278	0.1318	0.1180	0.1199
	SE_2	0.3695	0.4650	0.3195	0.4715	0.1399	0.1534	0.1310	0.1369
	ECP	0.9319	0.9220	0.9280	0.9296	0.9539	0.9500	0.9365	0.9433
Sample II	Bias	0.0182	0.0251	0.0072	0.0039	0.0025	0.0012	-0.0010	-0.0015
	SE_1	0.2944	0.3030	0.1840	0.1844	0.1203	0.1218	0.1095	0.1095
	SE_2	0.3266	0.3452	0.1916	0.1924	0.1300	0.1310	0.1103	0.1118
	ECP	0.9419	0.9540	0.9457	0.9437	0.9639	0.9640	0.9559	0.9542
SIMEX									
Quadratic	Bias	-0.0464	-0.0509	-0.0359	-0.0512	0.0196	0.0203	0.0276	0.0120
	SE_1	0.2025	0.1633	0.2866	0.1435	0.1317	0.1040	0.2503	0.1045
	SE_2	0.2317	0.2109	0.3799	0.1626	0.1051	0.1026	0.3665	0.1028
	ECP	0.8922	0.8804	0.8747	0.8801	0.9578	0.9505	0.9575	0.9532
Exponential	Bias	0.0637	0.0524	-0.0384	-0.0377	0.0326	0.0322	0.0243	0.0249
	SE_1	0.3224	0.3239	0.1709	0.1696	0.1044	0.1042	0.1044	0.1046
	SE_2	0.5545	0.6094	0.1768	0.1752	0.1023	0.1034	0.1021	0.1021
	ECP	0.9244	0.9185	0.9139	0.9221	0.9511	0.9464	0.9553	0.9533

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 2.2: *Empirical performance of estimators of the regression parameters associated with binary X and Z; Number of assessments are POI* (5)*;* $\rho = 0.2$, $\boldsymbol{\kappa = 0.5}$, $\beta_X = \beta_Z = \log(1.25)$, $\pi_{11} = 0.7$, $p_z = 0.5$ *and logit* $\left(p_{x|z}\right) = \log(2)\, z$.

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.1400	-0.1404	-0.0937	-0.0931	0.0319	0.0331	0.0395	0.0334
	SE_1	0.1027	0.1029	0.1030	0.1033	0.1025	0.1027	0.1028	0.1031
	SE_2	0.1069	0.1056	0.1062	0.1064	0.1007	0.1010	0.1946	0.1016
	ECP	0.7120	0.7160	0.8280	0.8353	0.9480	0.9440	0.9420	0.9438
Likelihood									
Known	Bias	-0.0075	0.0027	-0.0024	0.0014	0.0048	0.0044	0.0100	0.0106
	SE_1	0.2716	0.2796	0.1764	0.1777	0.1129	0.1136	0.1069	0.1070
	SE_2	0.2794	0.2976	0.1812	0.1838	0.1115	0.1132	0.1050	0.1055
	ECP	0.9578	0.9620	0.9500	0.9500	0.9598	0.9560	0.9560	0.9580
Sample I	Bias	0.0386	0.0439	0.0705	0.0663	-0.0032	-0.0032	0.0015	0.0024
	SE_1	0.3187	0.3486	0.2360	0.2423	0.1271	0.1310	0.1155	0.1172
	SE_2	0.4383	0.5507	0.3501	0.3870	0.1411	0.1541	0.1258	0.1312
	ECP	0.9260	0.9096	0.9376	0.9198	0.9460	0.9398	0.9497	0.9479
Sample II	Bias	-0.0053	0.0210	0.0029	0.0080	0.0049	0.0026	0.0090	0.0095
	SE_1	0.2850	0.2976	0.1794	0.1812	0.1172	0.1186	0.1078	0.1079
	SE_2	0.3575	0.3583	0.2068	0.2169	0.1255	0.1287	0.1075	0.1083
	ECP	0.9440	0.9478	0.9540	0.9540	0.9440	0.9398	0.9600	0.9560
SIMEX									
Quadratic	Bias	-0.0733	-0.0643	-0.0395	-0.0347	0.0048	0.0207	0.0076	0.0163
	SE_1	0.2134	0.1621	0.1669	0.1428	0.2346	0.1621	0.1693	0.1032
	SE_2	0.2267	0.2114	0.1681	0.1657	0.1639	0.2114	0.1385	0.1022
	ECP	0.8367	0.8381	0.9091	0.9091	0.9532	0.8381	0.9514	0.9515
Exponential	Bias	-0.0062	0.0193	-0.0221	-0.0183	0.0259	0.0322	0.0290	0.0304
	SE_1	0.3257	0.3159	0.1747	0.1766	0.1023	0.1028	0.1031	0.1032
	SE_2	0.3463	0.3802	0.1799	0.1802	0.1022	0.1011	0.1015	0.1017
	ECP	0.9111	0.9133	0.9212	0.9286	0.9502	0.9433	0.9398	0.9414

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 2.3: *Empirical performance of estimators of the regression parameters associated with binary X and Z; Number of assessments are POI* (5)*;* $\rho = 0.2$, $\boldsymbol{\kappa = 2}$, $\beta_X = \beta_Z = \log(1.25)$, $\pi_{11} = 0.7$, $p_z = 0.5$ *and logit* $\left(p_{x|z}\right) = \log(2)\, z$.

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.1294	-0.1308	-0.0855	-0.0861	0.0426	0.0402	0.0222	0.0209
	SE_1	0.1113	0.1114	0.1114	0.1115	0.1112	0.1113	0.1114	0.1115
	SE_2	0.1184	0.1130	0.1105	0.1100	0.1191	0.1142	0.1176	0.1170
	ECP	0.7840	0.7892	0.9060	0.9038	0.9300	0.9257	0.9320	0.9299
Likelihood									
Known	Bias	0.0158	0.0093	0.0026	0.0103	0.0105	0.0107	-0.0132	-0.0027
	SE_1	0.2932	0.2903	0.1916	0.1915	0.1219	0.1215	0.1153	0.1157
	SE_2	0.2973	0.2872	0.1911	0.1873	0.1242	0.1229	0.1246	0.1208
	ECP	0.9519	0.9618	0.9467	0.9558	0.9519	0.9538	0.9316	0.9357
Sample I	Bias	0.0539	0.0236	0.0735	0.0577	-0.0062	0.0138	-0.0232	-0.0117
	SE_1	0.3248	0.3382	0.2540	0.2629	0.1364	0.1383	0.1285	0.1315
	SE_2	0.3886	0.3992	0.3132	0.3594	0.1521	0.1513	0.1546	0.1816
	ECP	0.9399	0.9498	0.9400	0.9538	0.9459	0.9478	0.9220	0.9217
Sample II	Bias	0.0251	0.0143	0.0188	0.0139	0.0082	0.0093	-0.0028	-0.0031
	SE_1	0.3058	0.3048	0.1971	0.1966	0.1263	0.1257	0.1168	0.1167
	SE_2	0.3306	0.3073	0.1974	0.1926	0.1332	0.1282	0.1236	0.1225
	ECP	0.9460	0.9639	0.9519	0.9580	0.9440	0.9518	0.9359	0.9400
SIMEX									
Quadratic	Bias	-0.0438	-0.0433	-0.0196	-0.0227	0.0280	0.0270	0.0063	0.0024
	SE_1	0.1819	0.1735	0.2130	0.1517	0.1226	0.1110	0.1974	0.1116
	SE_2	0.2281	0.2256	0.2015	0.1695	0.1164	0.1144	0.1773	0.1179
	ECP	0.8648	0.8727	0.9320	0.9316	0.9362	0.9394	0.9339	0.9296
Exponential	Bias	0.0082	0.0220	-0.0070	-0.0077	0.0407	0.0391	0.0159	0.0163
	SE_1	0.3574	0.3323	0.1894	0.1802	0.1116	0.1113	0.1159	0.1117
	SE_2	0.5474	0.4978	0.1896	0.1897	0.1148	0.1142	0.1184	0.1190
	ECP	0.9189	0.9140	0.9496	0.9431	0.9265	0.9253	0.9372	0.9356

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 2.4: *Empirical performance of estimators of the regression parameters associated with binary X and Z; Number of assessments are POI* (5)*;* $\rho = 0.2$, $\boldsymbol{\kappa = 1}$, $\boldsymbol{\beta_X = \log(2)}$, $\beta_Z = \log(1.25)$, $\pi_{11} = 0.7$, $p_z = 0.5$ *and logit* $\left(p_{x|z}\right) = \log(2)\, z$.

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.4327	-0.4339	-0.2882	-0.2892	0.0893	0.0881	0.0601	0.0592
	SE_1	0.1021	0.1022	0.1026	0.1028	0.1017	0.1018	0.1021	0.1022
	SE_2	0.1023	0.1019	0.1000	0.1001	0.1015	0.1010	0.1061	0.1056
	ECP	0.0120	0.0100	0.1940	0.1864	0.8640	0.8657	0.9180	0.9178
Likelihood									
Known	Bias	-0.0029	0.0058	0.0017	0.0028	0.0103	0.0103	-0.0011	-0.0006
	SE_1	0.2688	0.2938	0.1769	0.1812	0.1153	0.1160	0.1101	0.1102
	SE_2	0.2624	0.2889	0.1720	0.1771	0.1133	0.1142	0.1141	0.1142
	ECP	0.9559	0.9660	0.9500	0.9618	0.9519	0.9500	0.9420	0.9378
Sample I	Bias	-0.0020	0.0608	0.0785	0.1195	0.0113	0.0073	-0.0036	-0.0034
	SE_1	0.2705	0.3098	0.2086	0.2287	0.1217	0.1258	0.1151	0.1165
	SE_2	0.3445	0.4389	0.2648	0.3764	0.1574	0.1697	0.1417	0.1465
	ECP	0.8818	0.8820	0.9116	0.9095	0.9018	0.9020	0.9056	0.8974
Sample II	Bias	-0.0026	0.0288	0.0075	0.0105	0.0167	0.0140	0.0000	-0.0000
	SE_1	0.2691	0.3010	0.1788	0.1845	0.1177	0.1200	0.1107	0.1109
	SE_2	0.2867	0.3545	0.1827	0.1886	0.1251	0.1340	0.1189	0.1194
	ECP	0.9360	0.9218	0.9479	0.9580	0.9400	0.9359	0.9359	0.9320
SIMEX									
Quadratic	Bias	-0.1931	-0.1940	-0.0956	-0.0940	0.0604	0.0594	0.0176	0.0202
	SE_1	0.1620	0.1601	0.2411	0.1421	0.1051	0.1022	0.2020	0.1037
	SE_2	0.2067	0.2058	0.1708	0.1584	0.1064	0.1059	0.1289	0.1111
	ECP	0.6822	0.6809	0.8660	0.8717	0.8966	0.9024	0.9380	0.9399
Exponential	Bias	0.0089	0.0052	-0.0530	-0.0530	0.0868	0.0849	0.0518	0.0512
	SE_1	0.3324	0.3195	0.1740	0.1719	0.1017	0.1018	0.1022	0.1023
	SE_2	0.4606	0.4553	0.1990	0.1993	0.1015	0.1010	0.1084	0.1076
	ECP	0.8674	0.8611	0.8600	0.8617	0.8653	0.8699	0.9212	0.9218

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

classification increases in severity. This is especially the case for estimates of β_X.

- Consistent with the asymptotic bias results, the performance of the naive approach seems to depend on the value of β_X. When $\beta_X = \log(2)$ the magnitude of the biases are much greater and the empirical coverage probabilities are much lower than when $\beta_X = \log(1.25)$. The impact on estimation of β_Z also appears to be greater for $\beta_X = \log(2)$ versus $\beta_X = \log(1.25)$. These results are based on positively correlated X and Z. It is plausible that the impact would change if the correlation between X and Z were different.

- For both models the correct maximum likelihood approach results in estimated biases close to 0 and empirical coverage probabilities close to the nominal value of 0.95. Maximum likelihood based on a large validation sample tends to perform better than with a small validation sample.

- The SIMEX approach appears to perform much better for minor misclassification. It only provides a partial correction for misclassification in the presence of severe misclassification. SIMEX is an approximate method in general since the exact extrapolation function is unknown. Also, the same extrapolation function is used in these simulations to obtain both the parameter and variance estimates. An exponential extrapolation function appears to perform better for estimation of the parameters associated with the misclassified variable, X and the quadratic extrapolation function seems to work well for estimation of the other parameters. Interestingly, the vast majority of the optimal extrapolation functions chosen based on adjusted R^2 were quadratic for κ, ρ, and β_Z; however, an exponential extrapolant was selected more frequently in the estimation for β_X. However, neither the selection of the "optimal" extrapolant based on adjusted R^2 or the average of the exponential and quadratic estimates appeared to consistently perform better than the quadratic or exponential models. For that reason these results are not presented here.

- There does not appear to be much of a difference between the results based on a Weibull model and the piecewise constant baseline hazards model regardless of the model used or for the value of κ.

- The results based on $\kappa = 0.5$, $\kappa = 1$ and $\kappa = 2$ were fairly consistent in terms of empirical biases, estimated standard errors and empirical coverage probabilities. The similarity in the biases is not surprising due to the trends observed in the asymptotic bias plots in the previous section.

- Two standard error estimates were provided in the tables. SE_1 is the average model-based standard error and SE_2 is the empirical standard deviation of the parameter estimates. For the most part, these two values are close. This suggests that the actual variation in the parameter estimates obtained based on these likelihood functions is what we would expect based on the model-based standard errors estimated by the inverse of the observed information matrix. However, there appears to be a difference between these two values for the correct likelihood approach based on validation data. The empirical standard errors tend to be larger than the average model-based ones and this difference is greater for the small validation sample than for the large validation sample. This is likely due to the excess variability introduced when parameters associated with misclassification and the $X|Z$ distribution are estimated and then treated as known in the likelihood function. This did not appear to be as much of an issue for the standard errors associated with the estimator for the Z effect. The SIMEX results presented are based on an estimated misclassification matrix using a small validation sample. The empirical standard errors for the β_X estimator also appear to be larger than the extrapolated model-based ones.

- The standard errors based on the naive approach are smaller than those calculated based on both of the approaches accounting for misclassification. This is consistent with the findings in other mismeasured covariate contexts that a naive approach leads to underestimated standard errors. This difference is especially dramatic for the standard errors associated with the estimator for β_X. For both the β_X and β_Z estimators, this difference appears to shrink as the misclassification decreased. Based on these results, it is difficult to draw conclusions regarding the relative size of the SIMEX or the correct likelihood standard errors. The relative size of the standard errors appear to depend on which SIMEX extrapolant is used and which parameter is being considered.

- The results summarized in Tables 2.1 to 2.4 were based on data generated such that the average number of assessments was 5 (i.e. $\mu = 5$). A small number of simulations were conducted with this mean set to 20. However, this did not appear to have much of an impact on the empirical biases, empirical coverage probabilities and standard errors (the results closely resembled those presented here).

A small number of simulations involving current status data where all subjects were only observed once were also performed. Here, a sample size of $n = 2000$ was considered rather than the sample size of $n = 500$ that was used for the general censoring scheme simulations above. It is not uncommon for large clinical databases that collect information on patients prospectively to contain more than 500 patients. Therefore, n=2000 is likely a reasonable sample size to investigate here. The maximum assessment time τ was first selected such that $P(T < \tau)$ was 0.6. Then, the individual assessment times, b_i, were generated according to an $EXP(\lambda_B)$ distribution, where λ_B was the solution to $p(Y = 1) = P(T < \min(B, \tau)) = 0.6$. This ensured that there would be an ample number of transitions prior to assessment or the end of the study to provide information on the parameters in the model depicted in Figure 2.2 (or in the parameters associated with piecewise constant baseline hazards). Table 2.5 summarizes the simulation results for one parameter configuration. Upon comparison to the results in Table 2.1, the empirical biases appear to be similar. However, the estimated standard errors appear to be much smaller for the current status data simulations. This is likely due to the sample size ($n = 2000$ in Table 2.5 versus the $n = 500$ used in the simulations which are summarized in Table 2.1). The empirical coverage probabilities are slightly less in Table 2.5. Otherwise, similar trends can be observed for the current status data situation. The naive parameter estimates are biased and the empirical coverage probabilities are less than the nominal 0.95. The correct likelihood approach seems to be successful in reducing bias and bringing the ECP's closer to the nominal levels and SIMEX provides only a partial correction in general for the misclassified covariate. The PCBH model resulted in several extreme estimates for β_X (i.e. $4 < \hat{\beta}_X < 5$) in the current status case. The empirical biases, coverage probabilities and standard errors reflect this. This did not appear to be an issue for the general interval censoring situation.

Table 2.5: *Empirical performance of estimators of the regression parameters associated with binary X and Z based on current status data; Assessment times $B \sim EXP(\lambda_B)$ where λ_B is chosen such that $P(T < min(B,\tau)) = 0.6$; $\rho = 0.2$, $\kappa = 1$, $\beta_X = \beta_Z = \log(1.25)$, $\pi_{11} = 0.7$, $p_z = 0.5$ and logit $\left(p_{x|z}\right) = \log(2)\, z$.*

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$		$\pi_{00} = 0.7$		$\pi_{00} = 0.9$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.1388	-0.1402	-0.0900	-0.0919	0.0285	0.0282	0.0276	0.0272
	SE_1	0.0606	0.0609	0.0610	0.0614	0.0606	0.0609	0.0606	0.0610
	SE_2	0.0638	0.0633	0.0661	0.0664	0.0605	0.0606	0.0610	0.0612
	ECP	0.3660	0.3622	0.6580	0.6646	0.9240	0.9256	0.9320	0.9329
Likelihood									
Known	Bias	-0.0058	0.0108	0.0001	0.0005	-0.0002	-0.0005	0.0042	0.0049
	SE_1	0.1582	0.1591	0.1030	0.1034	0.0660	0.0661	0.0631	0.0632
	SE_2	0.1644	0.3195	0.1114	0.1123	0.0674	0.0683	0.0636	0.0634
	ECP	0.9355	0.9333	0.9374	0.9425	0.9456	0.9454	0.9556	0.9548
Sample I	Bias	0.0317	0.0853	0.0093	0.0250	-0.0108	-0.0144	0.0032	0.0034
	SE_1	0.1925	0.2121	0.1070	0.1070	0.0755	0.0781	0.0645	0.0643
	SE_2	0.2556	0.4879	0.1277	0.2904	0.0956	0.1057	0.0669	0.0670
	ECP	0.9034	0.9165	0.9058	0.8998	0.9195	0.9145	0.9399	0.9305
Sample II	Bias	-0.0015	0.0097	0.0019	0.0109	-0.0015	-0.0020	0.0029	0.0036
	SE_1	0.1626	0.1643	0.1040	0.1040	0.0674	0.0675	0.0635	0.0635
	SE_2	0.1775	0.2715	0.1134	0.2268	0.0716	0.0725	0.0651	0.0654
	ECP	0.9336	0.9331	0.9359	0.9351	0.9477	0.9473	0.9499	0.9432
SIMEX									
Quadratic	Bias	-0.0595	-0.0614	-0.0177	-0.0276	0.0305	0.0204	0.0288	0.0160
	SE_1	0.1436	0.0960	0.0963	0.0844	0.1199	0.0612	0.0899	0.0616
	SE_2	0.1343	0.1273	0.1135	0.1036	0.1160	0.0625	0.0832	0.0616
	ECP	0.8388	0.8111	0.8654	0.8649	0.9076	0.9240	0.8953	0.9480
Exponential	Bias	0.0228	0.0179	-0.0021	-0.0186	0.0456	0.0270	0.0427	0.0247
	SE_1	0.1875	0.1900	0.0970	0.0987	0.0611	0.0609	0.0619	0.0611
	SE_2	0.3312	0.2893	0.1178	0.1157	0.0551	0.0611	0.0566	0.0604
	ECP	0.9107	0.8917	0.8980	0.8815	0.9024	0.9261	0.9173	0.9376

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

2.5.2 Continuous Covariates

DATA GENERATION

To be consistent with the simulations involving binary covariates, data were generated based on the true models and the joint distribution of (X, W) as outlined in Section 2.5.1. The true measurement error and covariate distributions had the following forms:

- $Z \sim N(\mu_Z, \sigma_Z^2)$, where
  - $\mu_Z = 0$ without loss of generality since Z could be a centered version of the covariate of interest, and
  - $\sigma_Z^2 = \begin{cases} 0.1 \\ 1 \end{cases}$ to represent low and high variability in Z.
- $X|Z \sim N\left(\mu_{X|Z}, \sigma_{X|Z}^2\right)$, where
  - $\mu_{X|Z} = \xi_0 + \xi_Z Z$, where $\xi_0 = 0$ and $\xi_Z = 0, 1.33$ to represent two plausible relationships between X and Z (i.e. when $\xi_Z = 0$, X and Z are independent and when $\xi_Z = 1.33$, $CORR(X, Z) = \rho_{XZ} = \xi_Z \sigma_Z / \sqrt{\sigma_{X|Z}^2 + \xi_Z^2 \sigma_Z^2} = 0.8$), and
  - $\sigma_{X|Z}^2 = 0.1, 1$ to represent low and high variability in X given Z. Note that we are making the simplifying assumption that the distribution of X only depends on Z through its mean. In other words, $\sigma_{X|Z}^2$ does not depend on Z.
- $W|X, Z \sim N\left(\mu_{W|X,Z}, \sigma_{W|X,Z}^2\right)$, where
  - $\mu_{W|X,Z} = \zeta_0 + \zeta_X X$, where $\zeta_0 = 0$ and $\zeta_X = 1$ (this is the classical error model (1.29)), and
  - $\sigma_{W|X,Z}^2$, which is the measurement error variance σ_U^2, will be selected to give values of 0.5 and 0.8 for the *reliability ratio*, which is defined as $\gamma = \sigma_{X|Z}^2 / \left(\sigma_{X|Z}^2 + \sigma_u^2\right)$. These values of γ were selected to represent low and moderate reliability of W as a measure for X. Values for σ_U^2 are summarized in the following table based on the selected simulation values for $\sigma_{X|Z}^2$ and γ:

	σ_U^2	
	$\sigma_{X\|Z}^2 = 0.1$	$\sigma_{X\|Z}^2 = 1$
$\gamma = 0.5$	0.1	1
$\gamma = 0.8$	0.025	0.25

When working with measurement error problems, the above hierarchical distributions are usually considered separately. However, given that each of these distributions are Normal here, we can consider the joint distribution of $(W, X, Z)'$ as follows:

$$\begin{pmatrix} X \\ W \\ Z \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_X \\ \mu_W \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{WX} & \sigma_{XZ} \\ \sigma_{WX} & \sigma_W^2 & \sigma_{WZ} \\ \sigma_{XZ} & \sigma_{WZ} & \sigma_Z^2 \end{pmatrix}\right). \tag{2.24}$$

To generate the covariate data in the simulations all parameters in (2.24) were expressed in terms of the simulation parameters from the hierarchical distribution specification described previously.

$$\begin{aligned}
\mu_X &= \xi_0 + \xi_Z \mu_Z \\
\mu_W &= \zeta_0 + \zeta_X \xi_0 + \mu_Z \left(\zeta_X \xi_Z + \zeta_Z\right) \\
\mu_Z &= \mu_Z \text{ (as specified above)} \\
\sigma_X^2 &= \sigma_{X|Z}^2 + \xi_Z^2 \sigma_Z^2 \\
\sigma_W^2 &= \sigma_U^2 + \zeta_X^2 \sigma_X^2 + \zeta_Z^2 \sigma_Z^2 - 2\zeta_X \zeta_Z \sigma_{XZ} \\
\sigma_Z^2 &= \sigma_Z^2 \text{ (as specified above)} \\
\sigma_{WX} &= \left(\zeta_X \xi_Z + \zeta_Z\right) \sigma_Z^2 \\
\sigma_{XZ} &= \xi_Z \sigma_Z^2 \\
\sigma_{WZ} &= \zeta_X \sigma_{X|Z}^2 + \left(\zeta_X \xi_Z^2 + \zeta_Z \xi_Z\right) \sigma_Z^2
\end{aligned}$$

As in the binary case, two validation samples were randomly selected from the 500 subjects in each dataset to estimate the measurement error and conditional covariate distributions. The small validation study was of size 50 and the large, of size 200.

ESTIMATION

Based on the validation data, the measurement error distribution was modeled as $W = \zeta_0 + \zeta_X X + \zeta_Z Z$ and estimates for ζ_0, ζ_X and ζ_Z were obtained using least squares. The model $X = \xi_0 + \xi_Z Z$ was also fit using least squares to obtain $\hat{\xi_0}$ and $\hat{\xi_Z}$ to substitute into the likelihood function for the correct maximum likelihood approach. The lifetime models fit to the data were of the same structure as the models used to generate the data so there was no model misspecification other than the mismeasurement in X. However, robust models with piecewise constant baseline hazards (weakly parametric) models were also considered here.

The SIMEX approach was implemented in the same way as described for the case of binary covariates. It involved repeated maximization of the likelihood function given in (2.23). The same simple variance approximation approach was used as described in Stefanski & Cook (1995) for continuous measurement error. Linear ($\theta(\nu) = a + b\nu$), quadratic ($\theta(\nu) = a + b\nu + c\nu^2$), cubic ($\theta(\nu) = a + b\nu + c\nu^2 + d\nu^3$), exponential ($\theta(\nu) = ae^{b\nu}$) and rational linear ($\theta(\nu) = a + \frac{b}{c+\nu}$) extrapolation functions were considered and fit using least squares in SAS with PROC REG and PROC NLIN. Again, for the purposes of these simulations the same extrapolant was used to obtain the SIMEX parameter and variance estimates, although there is no requirement that the parameter and variance estimate extrapolants be the same. However, as for the binary covariate case, in an attempt to automate this model building process, two other approaches were considered in this simulation study. First, the optimal model of the five fitted based on adjusted R^2 (i.e. $R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$) was selected to estimate the parameters and the variances. Second, since both the quadratic and exponential extrapolation models appeared to perform reasonably well, the average of the estimates arising from the two models were considered. Again, PROC LIFEREG could have been used to obtain the naive maximum likelihood estimates, however, PROC NLP was used here to be consistent with the other simulations.

The correct maximum likelihood approach was based on the following likelihood function (see (2.18)):

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left[\mathcal{F}_T\left(c_i|x, z_i; \boldsymbol{\theta}\right) - \mathcal{F}_T\left(d_i|x, z_i; \boldsymbol{\theta}\right)\right]^{\delta_i} \left[\mathcal{F}_T\left(c_i|x, z_i; \boldsymbol{\theta}\right)\right]^{1-\delta_i} \frac{1}{\sqrt{2\pi}\hat{\sigma_U}} \mathrm{e}^{\frac{-(w_i-x)^2}{2\hat{\sigma}_U^2}} \frac{1}{\sqrt{2\pi}\hat{\sigma}_{X|Z}} \mathrm{e}^{\frac{-\left(x-\hat{\mu}_{X|Z}\right)^2}{2\hat{\sigma}_{X|Z}^2}} dx.$$

This function was maximized with respect to $\boldsymbol{\theta} = (r, k, \beta_X, \beta_Z)'$, again with $\rho = \mathrm{e}^r$ and $\kappa = \mathrm{e}^k$. Gaussian quadrature was used to numerically approximate the integrals. In line with practice, the abscissas and weights were determined based on 20 points (Aitkin & Rocci 2002; Zucker 2005). These were generated in SAS based on code adapted from C^{++} code (Press et al. 2002). For both SIMEX and the maximum likelihood approaches the objective functions were maximized based on a quasi-Newton algorithm using PROC IML in SAS. Representative results from this simulation study are displayed in Tables 2.6 to 2.11.

DISCUSSION

As in the binary simulation results summarized in Section 2.5.1, the convergence rates were high (approximately 95% to 100%). Upon examination of the results, based on the parameter configurations considered, the following observations can be made regarding the finite sample behavior of the naive and adjusted covariate effect estimators.

- For all parameter configurations investigated, bias in the naive estimators appears to be greater in magnitude and the empirical coverage probabilities tend to be farther away from the nominal value of 0.95 for severe measurement error (i.e. $\gamma = 0.5$) versus moderate measurement error (i.e. $\gamma = 0.8$).

- The naive estimator for β_Z exhibits bias and low empirical coverage probabilities when X and Z are correlated (see Tables 2.6, 2.8 and 2.10). However, when they are uncorrelated, there appears to be negligible bias associated with the naive Z effect estimator and the associated empirical coverage probabilities are much closer to 0.95 (see Tables 2.7, 2.9 and 2.11). Interestingly however, when $\beta_X = \log(2)$ and $\gamma = 0.5$

Table 2.6: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are POI (5); $\rho = 0.2$, $\kappa = 1$, $\beta_X = \beta_Z = \log(1.25)$, $Z \sim N(0,1)$ and $X|Z \sim N(1.33Z, 1)$ such that $\rho_{XZ} = 0.8$.*

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\gamma = 0.5$		$\gamma = 0.8$		$\gamma = 0.5$		$\gamma = 0.8$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.1137	-0.1138	-0.0435	-0.0437	0.1433	0.1437	0.0575	0.0571
	SE_1	0.0393	0.0394	0.0500	0.0500	0.0776	0.0777	0.0868	0.0868
	SE_2	0.0377	0.0377	0.0523	0.0526	0.0768	0.0773	0.0948	0.0949
	ECP	0.1760	0.1809	0.8240	0.8255	0.5480	0.5447	0.8860	0.8850
Likelihood									
Known	Bias	-0.0015	-0.0017	0.0016	0.0018	0.0008	0.0011	0.0003	-0.0004
	SE_1	0.0821	0.0822	0.0636	0.0636	0.1184	0.1183	0.0997	0.0997
	SE_2	0.0793	0.0790	0.0664	0.0664	0.1170	0.1168	0.1089	0.1085
	ECP	0.9663	0.9641	0.9512	0.9485	0.9600	0.9599	0.9214	0.9227
Sample I	Bias	0.0086	0.0090	0.0055	0.0055	-0.0137	-0.0141	-0.0045	-0.0046
	SE_1	0.0862	0.0864	0.0646	0.0646	0.1234	0.1235	0.1008	0.1008
	SE_2	0.0964	0.0966	0.0710	0.0711	0.1392	0.1395	0.1121	0.1121
	ECP	0.9428	0.9388	0.9424	0.9378	0.9364	0.9283	0.9232	0.9185
Sample II	Bias	0.0011	0.0012	0.0023	0.0023	-0.0025	-0.0025	-0.0008	-0.0008
	SE_1	0.0832	0.0834	0.0637	0.0638	0.1197	0.1197	0.0998	0.0997
	SE_2	0.0827	0.0826	0.0668	0.0670	0.1226	0.1228	0.1092	0.1093
	ECP	0.9557	0.9578	0.9507	0.9528	0.9515	0.9494	0.9208	0.9270
SIMEX									
Quadratic	Bias	-0.0639	-0.0610	-0.0093	-0.0021	0.0672	0.0658	0.0027	0.0021
	SE_1	0.0779	0.0549	0.0753	0.0959	0.1200	0.0908	0.1019	0.0959
	SE_2	0.0639	0.0566	0.0667	0.1050	0.0977	0.0936	0.1057	0.1050
	ECP	0.7844	0.7799	0.9413	0.9259	0.8785	0.8763	0.9234	0.9259
Cubic	Bias	-0.0543	-0.0460	-0.0161	-0.0099	0.0271	0.0248	-0.0174	-0.0181
	SE_1	0.1229	0.0616	0.0931	0.0613	0.1390	0.0969	0.1089	0.0963
	SE_2	0.0865	0.0666	0.0991	0.0694	0.1492	0.1034	0.1146	0.1080
	ECP	0.8469	0.8379	0.9287	0.9155	0.9312	0.9202	0.8988	0.9031

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 2.7: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are POI (5); $\rho = 0.2$, $\kappa = 1$, $\beta_X = \beta_Z = \log(1.25)$, $Z \sim N(0,1)$ and $\boldsymbol{X|Z \sim N(0,1)}$ such that $\boldsymbol{\rho_{XZ} = 0}$.*

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\gamma = 0.5$		$\gamma = 0.8$		$\gamma = 0.5$		$\gamma = 0.8$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.1138	-0.1138	-0.0439	-0.0436	0.0021	0.0023	0.0050	0.0050
	SE_1	0.0387	0.0388	0.0492	0.0492	0.0550	0.0550	0.0551	0.0552
	SE_2	0.0370	0.0373	0.0506	0.0502	0.0557	0.0560	0.0570	0.0570
	ECP	0.1463	0.1496	0.8417	0.8454	0.9619	0.9570	0.9439	0.9464
Likelihood									
Known	Bias	-0.0005	-0.0009	0.0027	0.0023	0.0058	0.0060	0.0065	0.0065
	SE_1	0.0809	0.0810	0.0626	0.0625	0.0560	0.0560	0.0555	0.0555
	SE_2	0.0772	0.0772	0.0648	0.0648	0.0566	0.0568	0.0568	0.0569
	ECP	0.9487	0.9531	0.9372	0.9370	0.9615	0.9616	0.9498	0.9496
Sample I	Bias	0.0074	0.0074	0.0039	0.0041	0.0058	0.0059	0.0066	0.0069
	SE_1	0.0842	0.0843	0.0631	0.0632	0.0569	0.0569	0.0557	0.0557
	SE_2	0.0918	0.0919	0.0676	0.0680	0.0627	0.0629	0.0587	0.0589
	ECP	0.9339	0.9318	0.9411	0.9370	0.9403	0.9382	0.9432	0.9475
Sample II	Bias	-0.0002	-0.0006	0.0025	0.0024	0.0065	0.0065	0.0062	0.0062
	SE_1	0.0813	0.0814	0.0626	0.0627	0.0562	0.0563	0.0555	0.0556
	SE_2	0.0789	0.0786	0.0652	0.0654	0.0591	0.0592	0.0572	0.0572
	ECP	0.9574	0.9616	0.9391	0.9391	0.9488	0.9531	0.9412	0.9412
SIMEX									
Quadratic	Bias	-0.0633	-0.0622	-0.0089	-0.0086	-0.0041	-0.0009	0.0013	0.0016
	SE_1	0.0711	0.0537	0.0627	0.0587	0.0781	0.0551	0.0604	0.0551
	SE_2	0.0595	0.0579	0.0626	0.0619	0.0648	0.0556	0.0561	0.0563
	ECP	0.7684	0.7651	0.9256	0.9248	0.9614	0.9585	0.9483	0.9520
Cubic	Bias	-0.0516	-0.0458	-0.0143	-0.0096	-0.0136	-0.0095	-0.0107	-0.0075
	SE_1	0.1035	0.0603	0.0799	0.0601	0.0822	0.0544	0.0817	0.0542
	SE_2	0.0929	0.0685	0.0772	0.0656	0.0711	0.0543	0.0704	0.0539
	ECP	0.8548	0.8420	0.9243	0.9290	0.9673	0.9522	0.9429	0.9457

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 2.8: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are POI* (5)*;* $\rho = 0.2$, $\boldsymbol{\kappa = 0.5}$, $\beta_X = \beta_Z = \log(1.25)$, $Z \sim N(0,1)$ *and* $\boldsymbol{X|Z \sim N(1.33Z, 1)}$ *such that* $\boldsymbol{\rho_{XZ} = 0.8}$.

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\gamma = 0.5$		$\gamma = 0.8$		$\gamma = 0.5$		$\gamma = 0.8$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.1137	-0.1130	-0.0412	-0.0399	0.1481	0.1499	0.0612	0.0628
	SE_1	0.0392	0.0392	0.0498	0.0499	0.0773	0.0774	0.0861	0.0862
	SE_2	0.0397	0.0399	0.0483	0.0488	0.0762	0.0765	0.0898	0.0906
	ECP	0.1751	0.1888	0.8853	0.8902	0.5292	0.5241	0.8833	0.8841
Likelihood									
Known	Bias	0.0019	0.0052	0.0071	0.0092	0.0030	0.0018	0.0002	0.0006
	SE_1	0.0825	0.0832	0.0637	0.0639	0.1179	0.1184	0.0990	0.0991
	SE_2	0.0849	0.0867	0.0630	0.0638	0.1185	0.1204	0.1038	0.1048
	ECP	0.9439	0.9413	0.9550	0.9525	0.9416	0.9390	0.9459	0.9389
Sample I	Bias	0.0126	0.0171	0.0103	0.0125	-0.0119	-0.0145	-0.0024	-0.0022
	SE_1	0.0871	0.0881	0.0647	0.0649	0.1236	0.1245	0.0996	0.0999
	SE_2	0.0985	0.1018	0.0656	0.0663	0.1378	0.1411	0.1071	0.1081
	ECP	0.9249	0.9155	0.9527	0.9548	0.9343	0.9296	0.9369	0.9276
Sample II	Bias	0.0042	0.0077	0.0080	0.0100	-0.0003	-0.0019	-0.0006	-0.0000
	SE_1	0.0836	0.0844	0.0640	0.0641	0.1193	0.1199	0.0991	0.0992
	SE_2	0.0870	0.0891	0.0638	0.0645	0.1217	0.1238	0.1056	0.1064
	ECP	0.9482	0.9390	0.9548	0.9548	0.9482	0.9460	0.9389	0.9389
SIMEX									
Quadratic	Bias	-0.0603	-0.0603	-0.0053	-0.0033	0.0706	0.0731	0.0052	0.0066
	SE_1	0.0743	0.0546	0.0706	0.0600	0.0981	0.0902	0.1026	0.0951
	SE_2	0.0611	0.0605	0.0594	0.6000	0.0959	0.0951	0.0986	0.0995
	ECP	0.7967	0.7805	0.9602	0.9546	0.8665	0.8600	0.9563	0.9381
Cubic	Bias	-0.0500	-0.0460	-0.0064	-0.0046	0.0311	0.0337	-0.0198	-0.0130
	SE_1	0.0879	0.0618	0.0932	0.0612	0.1101	0.0962	0.1332	0.0956
	SE_2	0.0859	0.0728	0.0703	0.0625	0.1083	0.1070	0.1142	0.1002
	ECP	0.8180	0.8004	0.9494	0.9443	0.9106	0.9065	0.9478	0.9402

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 2.9: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are POI* (5)*; $\rho = 0.2$, $\kappa = 0.5$, $\beta_X = \beta_Z = \log(1.25)$, $Z \sim N(0,1)$ and $\boldsymbol{X|Z \sim N(0,1)}$ such that $\boldsymbol{\rho_{XZ} = 0}$.*

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\gamma = 0.5$		$\gamma = 0.8$		$\gamma = 0.5$		$\gamma = 0.8$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.1125	-0.1123	-0.0454	-0.0446	-0.0009	-0.0001	0.0014	0.0023
	SE_1	0.0382	0.0381	0.0486	0.0487	0.0543	0.0543	0.0544	0.0544
	SE_2	0.0382	0.0382	0.0493	0.0494	0.0523	0.0526	0.0521	0.0525
	ECP	0.1697	0.1707	0.8477	0.8482	0.9576	0.9593	0.9519	0.9474
Likelihood									
Known	Bias	0.0039	0.0078	0.0001	0.0019	0.0041	0.0055	0.0043	0.0055
	SE_1	0.0801	0.0808	0.0619	0.0621	0.0554	0.0555	0.0547	0.0548
	SE_2	0.0803	0.0817	0.0623	0.0630	0.0534	0.0537	0.0512	0.0516
	ECP	0.9523	0.9545	0.9538	0.9581	0.9523	0.9545	0.9582	0.9558
Sample I	Bias	0.0148	0.0192	0.0032	0.0051	0.0051	0.0064	0.0040	0.0053
	SE_1	0.0843	0.0853	0.0629	0.0630	0.0563	0.0564	0.0549	0.0550
	SE_2	0.0974	0.1001	0.0656	0.0663	0.0594	0.0598	0.0532	0.0537
	ECP	0.9371	0.9264	0.9536	0.9470	0.9328	0.9329	0.9448	0.9426
Sample II	Bias	0.0063	0.0100	0.0002	0.0020	0.0053	0.0063	0.0040	0.0052
	SE_1	0.0812	0.0820	0.0620	0.0622	0.0556	0.0557	0.0548	0.0548
	SE_2	0.0822	0.0836	0.0630	0.0636	0.0540	0.0546	0.0508	0.0512
	ECP	0.9674	0.9610	0.9536	0.9536	0.9500	0.9524	0.9581	0.9492
SIMEX									
Quadratic	Bias	-0.0609	-0.0596	-0.0105	-0.0079	-0.0028	-0.0032	-0.0017	-0.0010
	SE_1	0.0740	0.0526	0.0656	0.0586	0.0704	0.0543	0.0636	0.0543
	SE_2	0.0593	0.0578	0.0612	0.0608	0.0539	0.0523	0.0518	0.0518
	ECP	0.7578	0.7469	0.9384	0.9427	0.9567	0.9568	0.9550	0.9509
Cubic	Bias	-0.0466	-0.0443	-0.0141	-0.0112	-0.0139	-0.0114	-0.0130	-0.0105
	SE_1	0.0947	0.0591	0.0697	0.0589	0.0899	0.0538	0.0677	0.0534
	SE_2	0.0827	0.0692	0.0710	0.0659	0.0860	0.0509	0.0561	0.0497
	ECP	0.8361	0.8333	0.9180	0.9182	0.9733	0.9712	0.9487	0.9489

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 2.10: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are POI* (5)*;* $\rho = 0.2$, $\boldsymbol{\kappa = 1}$, $\boldsymbol{\beta_X = \log(2)}$, $\beta_Z = \log(1.25)$, $Z \sim N(0,1)$ *and* $\boldsymbol{X|Z \sim N(1.33Z, 1)}$ *such that* $\boldsymbol{\rho_{XZ} = 0.8}$.

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\gamma = 0.5$		$\gamma = 0.8$		$\gamma = 0.5$		$\gamma = 0.8$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.3784	-0.3772	-0.1634	-0.1637	0.3954	0.3956	0.1703	0.1709
	SE_1	0.0445	0.0445	0.0594	0.0594	0.0865	0.0865	0.0951	0.0952
	SE_2	0.0457	0.0460	0.0557	0.0551	0.0877	0.0876	0.0910	0.0911
	ECP	0.0000	0.0000	0.2224	0.2090	0.0060	0.0064	0.5691	0.5736
Likelihood									
Known	Bias	0.0162	0.0130	0.0039	0.0026	0.0010	0.0032	0.0035	0.0041
	SE_1	0.1225	0.1200	0.0863	0.0855	0.1363	0.1355	0.1126	0.1123
	SE_2	0.1262	0.1232	0.0798	0.0796	0.1400	0.1385	0.1050	0.1044
	ECP	0.9421	0.9462	0.9525	0.9548	0.9464	0.9462	0.9676	0.9742
Sample I	Bias	0.0511	0.0419	0.0147	0.0134	-0.0230	-0.0166	-0.0063	-0.0057
	SE_1	0.1333	0.1276	0.0879	0.0871	0.1434	0.1410	0.1138	0.1137
	SE_2	0.1925	0.1752	0.0978	0.0972	0.2075	0.1978	0.1317	0.1311
	ECP	0.8777	0.8817	0.9312	0.9355	0.8369	0.8430	0.9269	0.9247
Sample II	Bias	0.0232	0.0185	0.0071	0.0059	-0.0075	-0.0043	-0.0015	-0.0010
	SE_1	0.1245	0.1214	0.0868	0.0861	0.1383	0.1371	0.1131	0.1130
	SE_2	0.1401	0.1334	0.0849	0.0845	0.1545	0.1519	0.1098	0.1096
	ECP	0.9206	0.9290	0.9419	0.9298	0.9206	0.9183	0.9505	0.9527
SIMEX									
Quadratic	Bias	-0.2164	-0.2165	-0.0403	-0.0430	0.2092	0.2101	0.0264	0.0273
	SE_1	0.0689	0.0643	0.0808	0.0759	0.1020	0.0983	0.1093	0.1059
	SE_2	0.0760	0.0771	0.0804	0.0795	0.1074	0.1081	0.1068	0.1077
	ECP	0.1377	0.1373	0.8934	0.8985	0.4216	0.4423	0.9398	0.9417
Cubic	Bias	-0.1568	-0.1588	-0.0403	-0.0432	0.1117	0.1125	-0.0110	-0.0103
	SE_1	0.0884	0.0755	0.0887	0.0810	0.1163	0.1056	0.1143	0.1093
	SE_2	0.1078	0.1017	0.0928	0.0922	0.1286	0.1241	0.1157	0.1171
	ECP	0.4636	0.4423	0.8765	0.8528	0.7798	0.7582	0.9416	0.9394

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 2.11: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are POI* (5)*;* $\rho = 0.2$, $\kappa = 1$, $\beta_X = \log(2)$, $\beta_Z = \log(1.25)$, $Z \sim N(0,1)$ *and* $\boldsymbol{X|Z \sim N(0,1)}$ *such that* $\boldsymbol{\rho_{XZ} = 0}$.

Method		Mismeasured covariate (β_X)				Error-free covariate (β_Z)			
		$\gamma = 0.5$		$\gamma = 0.8$		$\gamma = 0.5$		$\gamma = 0.8$	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	-0.3803	-0.3805	-0.1594	-0.1586	-0.0203	-0.0198	-0.0071	-0.0069
	SE_1	0.0414	0.0415	0.0551	0.0553	0.0562	0.0562	0.0567	0.0567
	SE_2	0.0434	0.0435	0.0574	0.0576	0.0570	0.0575	0.0546	0.0543
	ECP	0.0000	0.0000	0.1868	0.2008	0.9218	0.9208	0.9604	0.9626
Likelihood									
Known	Bias	0.0088	0.0101	0.0101	0.0097	0.0056	0.0062	0.0034	0.0033
	SE_1	0.1149	0.1157	0.0803	0.0806	0.0643	0.0644	0.0601	0.0602
	SE_2	0.1228	0.1238	0.0822	0.0824	0.0644	0.0643	0.0574	0.0573
	ECP	0.9386	0.9362	0.9384	0.9360	0.9534	0.9553	0.9642	0.9640
Sample I	Bias	0.0481	0.0503	0.0193	0.0189	0.0019	0.0024	0.0056	0.0060
	SE_1	0.1261	0.1269	0.0817	0.0819	0.0669	0.0670	0.0605	0.0605
	SE_2	0.1973	0.1991	0.1053	0.1054	0.1069	0.1072	0.0764	0.0758
	ECP	0.8771	0.8617	0.8924	0.9000	0.8199	0.8128	0.8765	0.8800
Sample II	Bias	0.0155	0.0161	0.0079	0.0075	0.0090	0.0095	0.0033	0.0033
	SE_1	0.1167	0.1177	0.0800	0.0802	0.0648	0.0649	0.0602	0.0602
	SE_2	0.1270	0.1285	0.0863	0.0862	0.0700	0.0703	0.0603	0.0604
	ECP	0.9363	0.9426	0.9361	0.9400	0.9257	0.9577	0.9521	0.9520
SIMEX									
Quadratic	Bias	-0.2233	-0.2214	-0.0361	-0.0354	-0.0140	-0.0133	-0.0022	-0.0025
	SE_1	0.0721	0.0593	0.0795	0.0706	0.0661	0.0584	0.0704	0.0588
	SE_2	0.0740	0.0724	0.0801	0.0796	0.0608	0.0614	0.0582	0.0567
	ECP	0.1107	0.0840	0.8651	0.8690	0.9363	0.9308	0.9691	0.9583
Cubic	Bias	-0.1742	-0.1680	-0.0408	-0.0385	-0.0166	-0.0181	-0.0113	-0.0109
	SE_1	0.1058	0.0696	0.1005	0.0758	0.0928	0.0595	0.0792	0.0590
	SE_2	0.1307	0.0951	0.1104	0.0898	0.1149	0.0630	0.0633	0.0566
	ECP	0.3931	0.3585	0.8511	0.8413	0.9293	0.9203	0.9574	0.9504

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

(Table 2.11), the naive β_Z estimator demonstrates slight bias and empirical coverage probability of 0.92 even though $\rho_{XZ} = 0$.

- The value of κ does not appear to have much of an impact on the results. The empirical biases and coverage probabilities appear to be similar for the three different values of κ investigated.

- Again, as in the binary situation, results based on Weibull and piecewise constant baseline hazards models are similar suggesting that the piecewise model may be a robust model to adopt in practice.

- The results in Tables 2.6 to 2.11 are based on an average of five assessments per patient. A small number of simulations based on an average of 20 assessments were also run for severe measurement error ($\gamma = 0.5$). These results were consistent with those summarized here.

- The performance of all approaches tends to deteriorate somewhat as the magnitude of the true underlying effect of X increases. Even in relative terms, the naive and SIMEX empirical biases tend to be larger in magnitude when $\beta_X = \log(2)$ versus $\beta_X = \log(1.25)$ (compare Tables 2.8 and 2.9 to 2.10 and 2.11). Since there does not seem to be much of a difference between the naive and SIMEX standard errors for the two values of β_X, it is not surprising that the empirical coverage probabilities are less for the larger (in magnitude) β_X as well. The correct maximum likelihood approach is still successful in reducing bias and results in empirical coverage probabilities which are closer to the nominal value of 0.95, however, the empirical standard errors appear to be substantially larger when $\beta_X = \log(2)$ than when $\beta_X = \log(1.25)$ and hence the empirical coverage probabilities are a little smaller. This is especially the case for the small validation study.

- The two standard errors, SE_1 (average model-based standard error) and SE_2 (empirical standard error), generally tended to be close with the largest difference resulting from the correct maximum likelihood approach with a small validation study. Therefore the observed variability in the estimates was larger than the expected variability

under the models. Since this does not appear to be as much of an issue for the correct maximum likelihood approach based on a large validation study, the difference in observed and expected variability is most likely a reflection of the variability in the estimates of the measurement error and conditional covariate distribution parameters which are treated as known in the likelihood function.

- For the SIMEX approach, although five extrapolation functions were fit along with an optimal choice of extrapolant and an average extrapolant, the cubic and quadratic functions appeared to perform the best overall. The cubic extrapolant was selected most often when choosing the optimal function based on adjusted R^2 for all estimators and the rational linear and quadratic models were selected next. The rational linear model appeared to result in the lowest biases and empirical coverage probabilities closest to 0.95, especially for the β_X estimator; however, convergence problems were often encountered when fitting this model. Therefore, results based on this extrapolant were not summarized here. In practice, however, when faced with measurement error in a continuous covariate it is an important candidate model to consider in the extrapolation step.

- The estimated variability of the SIMEX procedure estimates is much smaller than for the correct likelihood procedures. This may be due in part to the fact that the correct likelihood procedure requires estimation of more parameters (e.g. the distribution of X given Z). Alternatively, it could be that the SIMEX variance estimates are underestimated because the variance approximation of Stefanski & Cook (1996) assumes known measurement error variance and extrapolant. However, given the observed difference between the variability of estimators under the two methods, if one were able to obtain accurate variance estimates for SIMEX, then narrower confidence intervals might be obtained.

Based on the simulation results for misclassified binary covariates and continuous covariates, it is clear that the presence of mismeasured covariates can have a substantial impact on inference. Therefore, steps must be taken to incorporate measurement uncertainty in the analysis. This is especially so if it is suspected that there is severe misclassification or measurement error present. SIMEX and correct likelihood approaches can be used to

address such a problem. Each approach has strengths and weaknesses. Based on the extrapolation functions investigated here (i.e. linear, quadratic, cubic, exponential and rational linear), the SIMEX approach seems to work reasonably well when there is minor misclassification or measurement error present. However, it only provides a partial bias correction in the presence of large measurement error or misclassification probabilities. This is likely due to the fact that we do not know the true extrapolation function. The SIMEX variance approximation used in the simulation studies assumes that the measurement error (or misclassification probabilities) are known so the standard errors presented with the simulation results are underestimated. A two-stage bootstrap procedure could be used to estimate standard errors if it is believed that there is substantial variability associated with the mismeasurement distribution parameter estimators.

The correct likelihood approach appears to perform well for different levels of mismeasurement and for both small and large validation studies. However, the empirical coverage probabilities were observed to be smaller for estimators based on the small validation study than those based on the large validation study. This is probably due to the fact that the sampling variability in the mismeasurement and covariate distribution parameter estimators is ignored because they are estimated based on external supplementary data and used in the likelihood function. Estimators of these parameters would be expected to be more variable for a small validation sample than for a large validation sample. Therefore, if their sampling variability was incorporated, we would expect larger standard errors associated with the estimators of interest based on a small validation sample compared to a large validation sample. Bootstrapping could be used to incorporate this additional variability. Estimates for the mismeasurement and covariate distribution parameters could be obtained from a bootstrap sample drawn from the external validation study. Then the correct likelihood function based on a bootstrap sample from the primary dataset could be maximized. After repeating this a large number of times, the variability in the estimators of interest could be estimated by their respective sample variances. If the supplementary data were included in the primary dataset, the mismeasurement and covariate distributions could be modeled and estimated along with the other parameters in the likelihood function.

In terms of computation, SIMEX involves repeated analyses using existing software, while the likelihood approach has to be programmed based on the problem at hand and maximized using general optimization software. The SIMEX approach tends to take longer to run, but the likelihood approach requires the development of problem-specific code. The piecewise constant models resulted in biases and empirical coverage probabilities that closely resembled those under a Weibull model. This is not surprising because PCBH models are considered to be robust. In practice, we do not know the underlying model, so PCBH models are an attractive choice because they require fewer assumptions regarding the distributional form of the failure time distribution. In the next section, the naive maximum likelihood, correct maximum likelihood and the SIMEX approaches will be applied and compared based on data arising from the motivating study which was described in Section 2.2. Both Weibull models and PCBH models will be fit to these data.

2.6 Application: Psoriatic Arthritis Data

Based on reliability study results, the presence of measurement error and misclassification has been confirmed in factors that are commonly included in investigations on the progression of PsA. A multi-center reliability study was conducted by Gladman et al. (2004) investigating variation between physician's assessments performed on PsA patients. Ten PsA patients were selected to represent a broad range of joint damage, joint inflammation and spinal involvement. As well, ten rheumatologists who are members of the Spondyloarthritis Research Consortium of Canada thoroughly assessed each of the ten patients. A combination of continuous and categorical variables involving evaluation of peripheral joint disease, spinal involvement and enthesitis were included in the investigation. After examination of the reliability coefficients, it was concluded that the variables associated with the evaluation of peripheral joint disease demonstrated moderate to substantial reliability. However, those involving evaluation of spinal involvement and enthesitis did not perform as well (Gladman et al. 2004).

The objective in the application presented here is to incorporate the information available from this reliability study into an analysis similar to that of Gladman et al. (1995).

An obvious indication of the progression of PsA is the development of damaged joints. Therefore, we will consider a two-state model as in Figure 2.2 with State 1 defined as no damaged joints and State 2, one or more damaged joints. In addition, we will assume that the number of damaged joints determined via clinical assessment is a perfectly measured variable.

This analysis was based on data extracted from the PsA clinic database as of early 2005. For the purposes of this analysis, we will restrict attention to the 378 patients who entered the study in State 1 (i.e. with no damaged joints); the transition times for these 378 patients are either interval-censored or right-censored. Table 2.12 summarizes the demographics for this group of patients (Table 3.10 summarizes the demographics of the entire group of patients which will be included in the application in Chapter 3). The covariates labeled as Z are considered perfectly measured for the purposes of this analysis and those labeled as W are those which are prone to error. The purpose of these analyses is to demonstrate the effects of mismeasured covariates in practice. Information regarding the distribution of the W variables is available from the reliability study. One of the models which will be investigated involves one binary variable subject to misclassification along with several baseline (i.e. fixed) variables which are assumed to be precisely measured. A second model will include a continuous covariate subject to error which will be fit along with the precisely measured variables. No interactions will be considered at this time. In both cases SIMEX involved repeated estimation using naive maximum likelihood ($B = 150$, here) for different multiples of mismeasurement given by $\boldsymbol{\nu} = \{0, 0.5, 1, 1.5, 2\}$. Candidate extrapolation functions that were considered for both the parameter estimates and the variance estimates included linear, quadratic, exponential and rational linear (or nonlinear) functions. Error sums of squares and adjusted R^2 ($R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$) were considered to determine the extrapolation function that provided the best fit. Likelihood ratio tests were used for the naive and correct likelihood approach to determine the final models. To determine the final model based on SIMEX, variables were omitted if their coefficients did not appear to be significantly different from zero based on individual t-tests.

Table 2.12: *Patient demographics and covariates at clinic entry (patients entering in State 1).*

		n = 378
Gender (Z_G)	Women	163
	Men	215
Age at PsA Diagnosis (Z_{AP})	Average	**35.3**
	Range	(10-79)
PsA Duration (years) (Z_{DP})	Average	**5.3**
	Range	(0-47.3)
Number of Effused Joints (Z_E)	Average	**2.6**
	Range	(0-33)
Presence of Dactylitis (W_D)	Yes	129
	No	249
Back Measurements		
Upper Back (W_U)	Average	**2.2**
	Range	(0-4.5)
Middle Back (W_M)	Average	**3.3**
	Range	(0.4-6)
Lower Back (W_L)	Average	**4.0**
	Range	(0-8.5)

2.6.1 Misclassification in a Binary Covariate

Based on the reliability study, the observed value of X_D, the presence of dactylitis variable is prone to misclassification. Gladman et al. (2004) define dactylitis as the diffuse swelling of an entire digit. Dactylitis was coded as 1 if this swelling was observed in at least one digit. The misclassification probabilities along with the prevalence of dactylitis (i.e. $\pi = P(X = 1)$) were estimated using the reliability data based on the likelihood function given in (2.20). The resulting estimates were $\hat{\pi} = 0.61$, $\hat{\pi}_{01} = 0.31$ and $\hat{\pi}_{10} = 0.12$. The misclassification probabilities will be required for both the correct maximum likelihood and the SIMEX approaches. The prevalence estimate, $\hat{\pi}$, will also be used in the correct likelihood approach. For this analysis, W_D, along with several other variables assumed to be perfectly measured: gender (Z_G), age at PsA onset (Z_{AP}), duration of PsA at clinic entry (Z_{DP}), and the number of effused joints at clinic entry (Z_E) will be fit using both Weibull and piecewise constant baseline hazard (PCBH) regression models for comparison purposes. These variables were chosen because they are relevant factors in the study of PsA. The cut-points used in estimation of the PCBH model, a_1, a_2 and a_3, were calculated as the 25^{th}, 50^{th} and 75^{th} percentiles of all the observation times, respectively (6, 11 and 17 years).

Table 2.13: *Estimates obtained by fitting naive and correct Weibull regression and piecewise constant models and applying the SIMEX procedure to the PsA clinic data with a misclassified binary covariate (X_D).*

		Naive			**Correct**			**SIMEX**		
		Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
Weibull	Age at PsA Diagnosis ($\beta_{Z_{AP}}$)	0.0120	0.0071	0.0911	0.0123	0.0073	0.0894	0.0097	0.0511	0.8492
	PsA duration ($\beta_{Z_{DP}}$)	-0.0946	0.0131	<0.0001	-0.0970	0.0136	<0.0001	-0.1060	0.0492	0.0316
	Gender (β_{Z_G})	0.0728	0.1524	0.6331	0.0593	0.1572	0.7062	0.0351	0.1610	0.8275
	Effused joint count (β_{Z_E})	0.0520	0.0162	0.0015	0.0510	0.0169	0.0026	0.0458	0.0188	0.0152
	Presence of dactylitis (β_{X_D})	0.1973	0.1544	0.2019	0.5065	0.3522	0.1510	0.4328	0.3208	0.1781
PCBH	Age at PsA Diagnosis ($\beta_{Z_{AP}}$)	0.0113	0.0070	0.1063	0.0113	0.0071	0.1097	0.0123	0.0070	0.0783
	PsA duration ($\beta_{Z_{DP}}$)	-0.0916	0.0131	<0.0001	-0.0935	0.0135	<0.0001	-0.0904	0.0130	<0.0001
	Gender (β_{Z_G})	0.0819	0.1523	0.5910	0.0791	0.1547	0.6094	0.0465	0.1556	0.7651
	Effused joint count (β_{Z_E})	0.0479	0.0162	0.0032	0.0478	0.0165	0.0039	0.0398	0.0153	0.0097
	Presence of dactylitis (β_{X_D})	0.2099	0.1545	0.1749	0.4273	0.3172	0.1786	0.3866	0.2275	0.0902

First consider the full model, the results of which are summarized in Table 2.13. The estimates of the four "Z" variables appear to be quite similar for the three approaches. For the most part, the estimated standard errors tend to agree across the three methods with the exception of the SIMEX standard error estimates corresponding to $\beta_{Z_{AP}}$ and $\beta_{Z_{DP}}$ based on a Weibull regression model; these are much larger than those based on the maximum likelihood approaches, most likely due to the approximate nature of the SIMEX approach (i.e. we do not know the exact form of the extrapolant). For the misclassified variable, the correct likelihood and SIMEX estimates and standard errors are larger in magnitude than the naive ones. Therefore, the presence of misclassification in X_D appears to induce the attenuation phenomenon in the naive estimators that has been observed in other covariate measurement error problems. The Weibull and PCBH regression models give more or less similar estimates and standard errors. Although the estimates and standard errors based on the SIMEX approach differ slightly from the correct maximum likelihood approach, the directions and the significance of the effects appear to be consistent for the two approaches. Since the effect of X_D on the time to damage of at least one joint does not appear to be significant, the final models based on the naive and correct likelihood and the SIMEX approaches are the same and are summarized in Table 2.14.

Table 2.14: *Final model estimates obtained by fitting naive and correct Weibull regression and piecewise constant models and applying the SIMEX procedure to the PsA clinic data without a misclassified binary covariate (X_D).*

		Estimate	SE	P-value
Weibull	PsA duration ($\beta_{Z_{DP}}$)	-0.1016	0.0126	<0.0001
	Effused joint count (β_{Z_E})	0.0590	0.0156	0.0002
PCBH	PsA duration ($\beta_{Z_{DP}}$)	-0.0980	0.0126	<0.0001
	Effused joint count (β_{Z_E})	0.0548	0.0156	0.0005

There is no need to conduct an analysis incorporating covariate measurement error when the covariate subject to error is not included in the model. Based on these results, it appears that the duration of Psoriatic Arthritis (Z_{DP}) and the number of swollen joints at clinic entry (Z_E) are associated with the time to development of at least one dam-

aged joint. The variable Z_{DP} appears to exhibit a protective effect. Therefore, for each additional year of PsA duration at clinic entry, the relative risk of developing damaged joints is RR=0.9034 [95% CI (0.8814,0.9260)] based on a Weibull model and RR=0.9066 [95% CI (0.8845,0.9293)] under a piecewise constant model. For each additional swollen joint observed at clinic entry, the relative risk of joint damage is RR=1.0608 [95% CI (1.0288,1.0937)] and RR=1.0563 [95% CI (1.0245,1.0891)] based on Weibull and PCBH models, respectively. Note that the relative risk estimates and associated confidence intervals are very close for the Weibull and PCBH models. Since PCBH models are considered to be robust, this suggests that the Weibull model seems to be appropriate for these data.

2.6.2 Measurement Error in a Continuous Covariate

Consider the back measurement variables in Table 2.12. These measurements are based on the Smythe test (Gladman et al. 2004). With the patient in full flexion (i.e. bent forward as far as possible), a line is drawn on the patient's lower back at the level of the dimples of Venus. Three additional lines are drawn 10 cm apart. The back measurements are then recorded as the differences (in cm) between 10 cm (at full flexion) and the length of the three segments created by the four lines when the patient stands upright. Based on the reliability data, Gladman et al. (2004) report 95% confidence intervals for the intraclass correlation coefficient (ICC) corresponding to these variables as $(-0.01, 0.38)$, $(0.06, 0.53)$ and $(-0.01, 0.37)$, respectively. Based on the confidence intervals, these measurements appear to exhibit only moderate to poor reliability. However, since these measurements gauge patient mobility there is most likely substantial variability even in repeated measurements on the same patient by the same physician. To incorporate measurement error in an analysis, we first need to use the reliability data to estimate the measurement error distribution. No "true" values of the back variables are measured. What are available, however, are repeated measurements on 10 patients by 10 physicians. These patients were selected from the PsA clinic. However, there was no identifier contained in the reliability data to link them with the primary data from the PsA database. Therefore, this supplementary data was treated as an external dataset. As discussed in Section 2.4.3, these repeated measurements can be used to obtain information about the measurement error variance. Assuming the classical error model in (1.29) is appropriate, with $i = 1, 2, ..., 10$

Table 2.15: *Measurement error and covariate distributions.*

		X, $\boldsymbol{Z}$ Independent			X, $\boldsymbol{Z}$ Dependent		
Covariate	$\hat{\sigma}^2_U$	$\hat{\mu}_X$	$\hat{\sigma}^2_X$	$\hat{\gamma}_X$	$\hat{\mu}_{X\|Z}$	$\hat{\sigma}^2_{X\|Z}$	$\hat{\gamma}_{X\|Z}$
Upper Back	0.6199	2.1977	0.3141	0.3363	$2.9390 - 0.0153Z_{AP} - 0.0264Z_{DP}$	0.2566	0.2328
Middle Back	0.5882	3.1225	0.4930	0.4560	$4.0558 - 0.0201Z_{AP} - 0.0297Z_{DP}$	0.4087	0.4100
Lower Back	0.6153	3.8556	0.4058	0.3374	$4.8565 - 0.0207Z_{AP} - 0.0359Z_{DP}$	0.2981	0.3264

and $j = 1, 2, ..., 10$, we can use the following random effects model to estimate σ^2_U for each variable:

$$W_{ij} = \mu_x + \alpha_i + e_{ij}, \text{ where } e_{ij} \sim N\left(0, \sigma^2_U\right) \text{ and } \alpha_i \sim N\left(\mu_i, \sigma^2\right). \tag{2.25}$$

Here X_i is represented by two components; an unknown fixed effect, μ_x, and a random effect associated with patient i, α_i. In this analysis, the patient effect means are assumed to differ between patients, however the corresponding variances are assumed to be constant across patients. This model was fit to the reliability data using PROC MIXED in SAS. The resulting measurement error variance estimates, $\hat{\sigma}^2_U$, are displayed in Table 2.15.

For the correct likelihood approach, we also must assume a distributional form for the conditional distribution of X, the true covariate, given $\boldsymbol{Z}$, the precisely measured covariates. As is often assumed in practice, we assume that $X|\boldsymbol{Z}$ follows a normal distribution, $X|\boldsymbol{Z} \sim N(\mu_{X|Z}, \sigma^2_{X|Z})$. The back measurements were selected for consideration here in part due to the symmetric, bell-shaped pattern exhibited in their histograms, suggesting that normality is probably not an unreasonable assumption for these variables. Moreover, there is an increasing interest in rheumatology on the impact of back involvement on disease course. When validation data are available, where X is measured along with $\boldsymbol{Z}$ on a group of patients possibly included in the study, estimation of the parameters of this distribution is straightforward. However, in this situation, X is not actually measured for any of the patients and the reliability data do not include measurements on $\boldsymbol{Z}$ in addition to those on W. However, the classical error model (1.29) can be used to determine expressions for $\mu_{X|Z}$ and $\sigma^2_{X|Z}$ in terms of quantities that can be estimated using the primary and reliability data.

Suppose, first that the distribution of X does not depend of $\boldsymbol{Z}$. Then, by taking the expectation with respect to X of both sides of $E(W|X) = X$, it follows that $\mu_X = \mu_W$. Also, from (1.29), $VAR(W|X) = \sigma_U^2$ so $VAR(W) = E[VAR(W|X)] + VAR[E(W|X)] = \sigma_U^2 + \sigma_X^2$. Therefore, an estimate of μ_X is given by $\hat{\mu}_W$ and σ_X^2 could be estimated by the difference, $\hat{\sigma}_W^2 - \hat{\sigma}_U^2$. The primary data could be used to estimate μ_W and σ_W^2, whereas only the reliability data would be of use to estimate σ_U^2. Alternatively, we could use only the reliability data to estimate the $X|\boldsymbol{Z}$ distribution parameters based on the random effects model introduced above to estimate the measurement error variance. The only source of variability in $\mu_X + \alpha_i$ is in the random effect α_i. Therefore an estimate for σ_X^2 would be $\hat{\sigma}^2$ in this random effects model.

Instead, if the distribution of X depends on $\boldsymbol{Z}$, since measurements of $\boldsymbol{Z}$ are not included in the reliability study, a random effects model such as this one cannot be used to estimate these quantities. Assume that (1.29) holds and the distribution of U does not depend on $\boldsymbol{Z}$. Then if $W = X + U = \beta_0 + \boldsymbol{\beta}_Z'\boldsymbol{Z} + \epsilon$, where $\epsilon \sim N\left(0, \sigma_{W|Z}^2\right)$, it follows that $\hat{\mu}_{X|Z} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_Z'\boldsymbol{Z}$ and $\hat{\sigma}_{X|Z}^2 = \hat{\sigma}_{W|Z}^2 - \hat{\sigma}_U^2$. Table 2.15 summarizes the parameter estimates associated with the back variable distributions. All Z variables in Table 2.12 were fit in a linear regression, but only the effects of Z_{AP} and Z_{DP} appeared to be significantly different from 0. These represent age at PsA onset and PsA duration at clinic entry, respectively. For this analysis, X will be permitted to depend on Z is this way. One error-prone variable, the middle back variable (W_M), along with several other variables assumed to be perfectly measured: gender (Z_G), age at PsA onset (Z_{AP}), duration of PsA at clinic entry (Z_{DP}), and the number of effused joints at clinic entry (Z_E) will be fit using both Weibull and piecewise constant baseline hazard regression models. These variables were selected to be investigated because they have been identified as relevant factors in the study of PsA.

Table 2.16 summarizes the results based on fitting the full model. The error-prone variable, X_M appears to be significant based on the three approaches, although the estimate and standard error appear to be underestimated in the naive maximum likelihood approach. As in the binary case, the presence of measurement error appears to induce attenuation. The SIMEX standard error estimate for the PCBH regression model is smaller

than the naive estimate. However, this is likely a result of the true extrapolation function being unknown. The estimates associated with the other variables and their corresponding standard error estimates also appear to be smaller in magnitude for the naive maximum likelihood approach compared to the other two methods. As was observed in the binary case, the Weibull and PCBH regression models tend to agree. The final model results are summarized in Table 2.17. Figures 2.15 to 2.17 illustrate the SIMEX approach based on Weibull and PCBH regression models. The results based on the likelihood and SIMEX approaches and the two models tend to be more or less consistent, suggesting that the magnitude of the estimates and standard errors appear to be underestimated by the naive approach.

In addition to the two variables that were observed to be associated with time to damage of at least one joint in Section 2.6.1, the error-prone variable, X_M also appears to be associated with the outcome of interest here. This suggests that the more middle back mobility a patient has, the lower the risk of developing at least one damaged joint. Another way of interpreting this is that back mobility is protective for the development of damaged joints. Based on a Weibull model, the correct likelihood approach results in an estimate of the relative risk (of joint damage with a 1 cm increase in middle back mobility) of RR=0.1351 [95% CI (0.0208,0.8755)], while the SIMEX procedure gives RR=0.1202 [95% CI (0.0680,0.2122)]. Similarly, based on the piecewise constant model, RR=0.3216 [95% CI (0.1313,0.7874)] and RR=0.0693 [95% CI (0.0474,0.1013)] for the likelihood and SIMEX approaches, respectively. Although these relative risk estimates differ, they suggest there is a reduction of joint damage risk with increased back mobility. This reduction appears to be underestimated by the naive approach. Note that the SIMEX estimate for the effect of duration of PsA at clinic entry is not significant under the piecewise constant model and is therefore not included in Table 2.17. In this application, treating the back measurement as precisely measured appears to underestimate the magnitude of its effect (and overestimate the corresponding relative risk) as well as those corresponding to the correctly measured variables. Therefore, if it is of interest to learn about the true underlying effects, it is critical that an analysis that incorporates measurement error be conducted.

Table 2.16: *Estimates obtained by fitting naive and correct Weibull regression and piecewise constant models and applying the SIMEX procedure to the PsA clinic data with an error-prone continuous covariate (X_M).*

		Naive			**Correct**			**SIMEX**		
		Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
Weibull	Age at PsA Diagnosis ($\beta_{Z_{AP}}$)	-0.0195	0.0153	0.2036	-0.0404	0.0265	0.1285	-0.0301	0.0219	0.1710
	PsA duration ($\beta_{Z_{DP}}$)	-0.1284	0.0291	<0.0001	-0.2361	0.0991	0.0176	-0.1557	0.0420	0.0002
	Gender (β_{Z_G})	-0.5490	0.3611	0.1293	-0.8928	0.6409	0.1644	-0.4141	0.4506	0.3587
	Effused joint count (β_{Z_E})	0.2303	0.0722	0.0016	0.3923	0.1832	0.0329	0.2586	0.1089	0.0181
	Middle back ROM (β_{X_M})	-0.7263	0.2228	0.0012	-2.8828	1.4324	0.0449	-1.2308	0.5025	0.0148
PCBH	Age at PsA Diagnosis ($\beta_{Z_{AP}}$)	-0.0176	0.0150	0.2409	-0.0204	0.0181	0.2617	-0.0162	0.0684	0.8132
	PsA duration ($\beta_{Z_{DP}}$)	-0.1234	0.0284	<0.0001	-0.1376	0.0396	0.0005	-0.1665	0.1028	0.1167
	Gender (β_{Z_G})	-0.5680	0.3609	0.1162	-0.6742	0.4214	0.1105	-0.5412	0.6557	0.4098
	Effused joint count (β_{Z_E})	0.2278	0.0708	0.0014	0.2427	0.0933	0.0096	0.2402	0.0813	0.0033
	Middle back ROM (β_{X_M})	-0.6259	0.2007	0.0020	-1.2850	0.5124	0.0126	-0.9127	0.1674	<0.0001

Table 2.17: *Final model estimates obtained by fitting naive and correct Weibull regression and piecewise constant models and applying the SIMEX procedure to the PsA clinic data with an error-prone continuous covariate (X_M).*

		Naive			**Correct**			**SIMEX**		
		Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
Weibull	PsA duration ($\beta_{Z_{DP}}$)	-0.1064	0.0267	<0.0001	-0.1511	0.0576	0.0091	-0.1522	0.0270	<0.0001
	Effused joint count (β_{Z_E})	0.1743	0.0645	0.0072	0.2359	0.1195	0.0491	0.2815	0.1002	0.0052
	Middle back ROM (β_{X_M})	-0.6859	0.2119	0.0013	-2.0020	0.9536	0.0365	-2.1190	0.2901	<0.0001
PCBH	PsA duration ($\beta_{Z_{DP}}$)	-0.1002	0.0253	<0.0001	-0.1085	0.0318	0.0007			
	Effused joint count (β_{Z_E})	0.1762	0.0643	0.0064	0.1820	0.0798	0.0231	0.1016	0.0391	0.0108
	Middle back ROM (β_{X_M})	-0.5875	0.1931	0.0025	-1.1345	0.4569	0.0135	-2.6696	0.1939	<0.0001

Figure 2.15: *Final model estimates of parameters obtained by applying the SIMEX procedure to the PsA clinic data based on a Weibull model with an error-prone continuous covariate (X_M).*

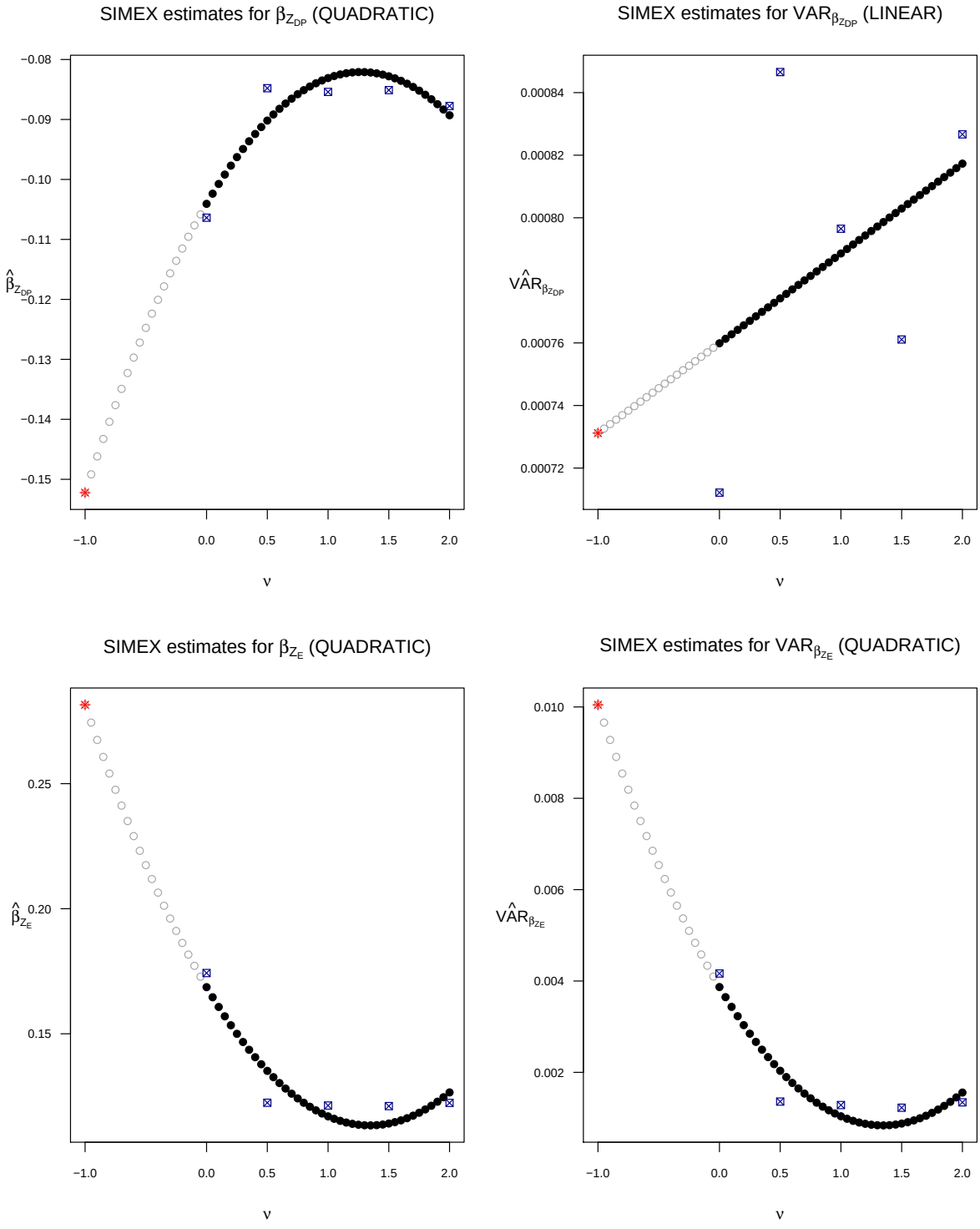


Figure 2.16: *Final model estimates of parameters obtained by applying the SIMEX procedure to the PsA clinic data based on a Weibull model with an error-prone continuous covariate (X_M).*

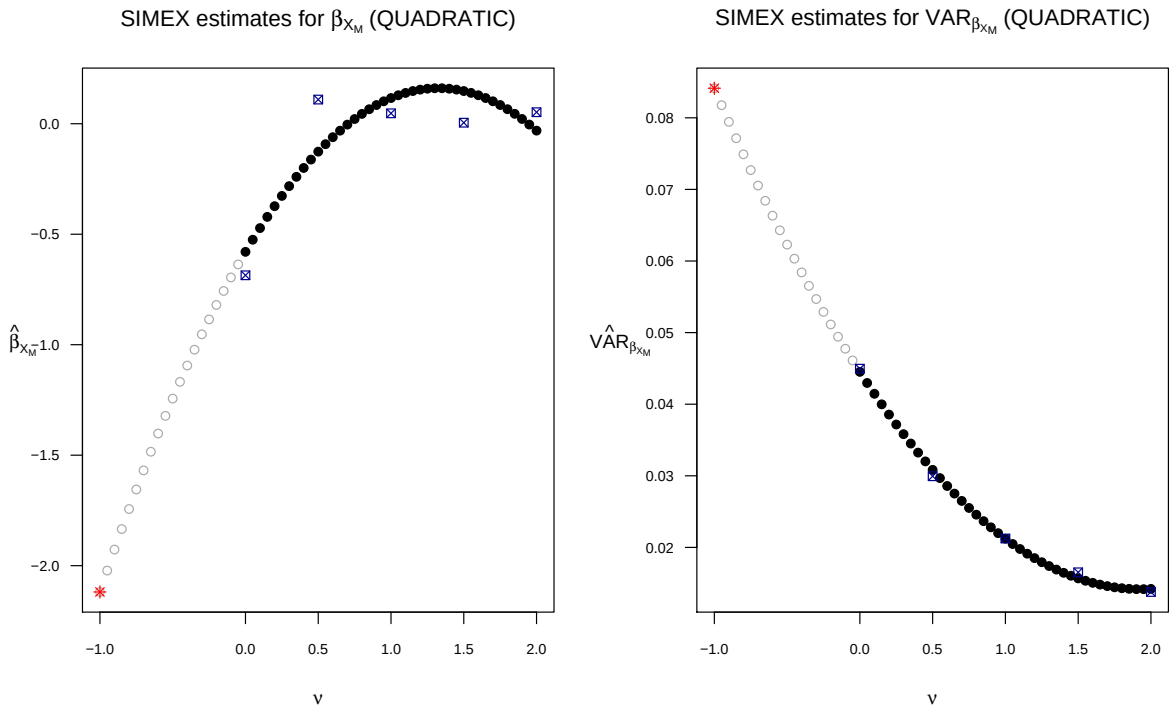
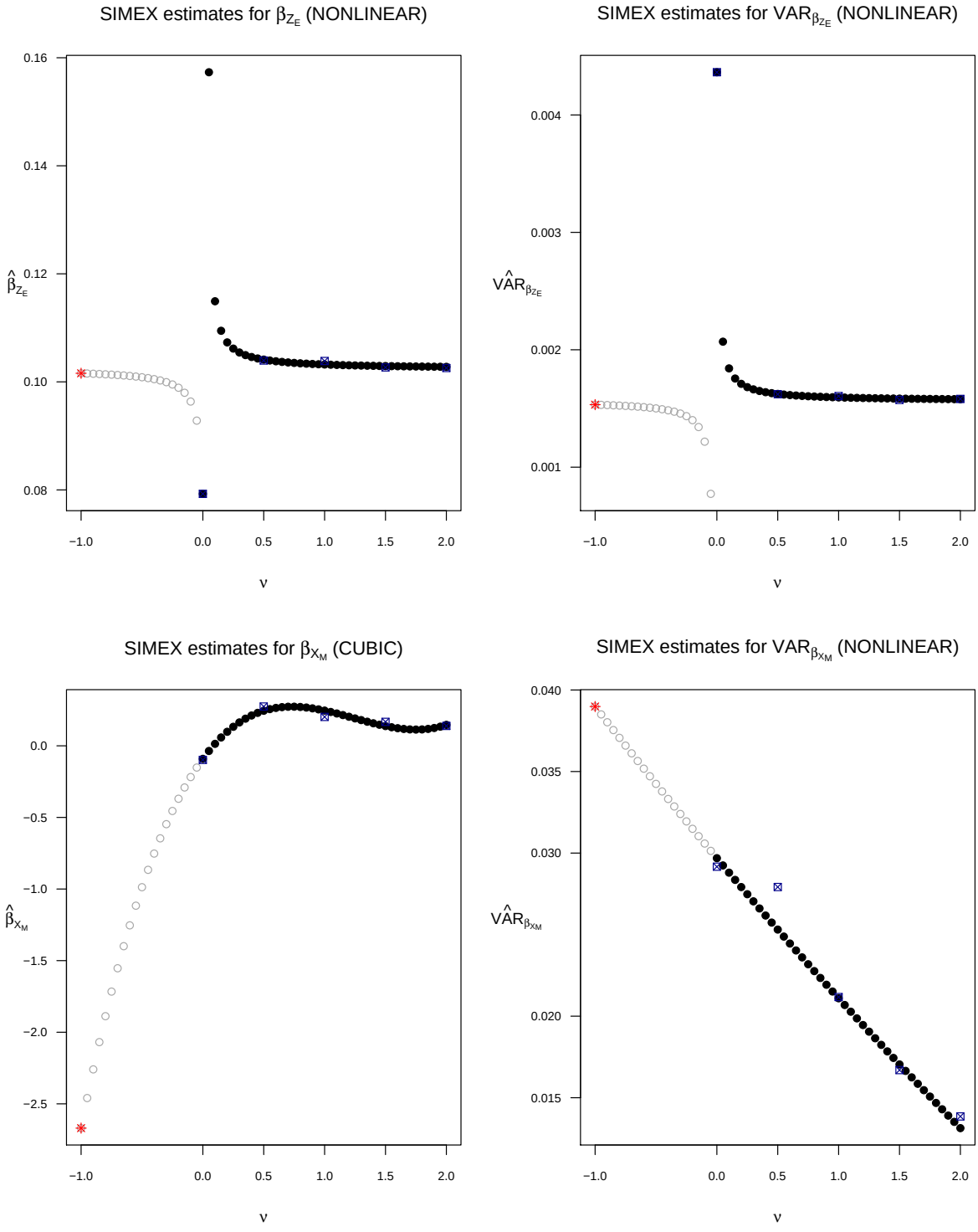


Figure 2.17: *Final model estimates of parameters obtained by applying the SIMEX procedure to the PsA clinic data based on a piecewise constant model with an error-prone continuous covariate (X_M).*



2.6.3 Discussion

Based on the application presented above, parameter estimates and standard errors appear to differ between the correct maximum likelihood approach and the SIMEX approach. The covariate effects have the same signs, but the maximum likelihood estimates often appear to be larger (in magnitude) than the SIMEX estimators. In addition, the likelihood standard errors often seem to be larger than those based on SIMEX. These differences could be due to the fact that the SIMEX estimators are only approximately consistent since we do not know the exact form of the extrapolation function. Also, the variance approximation procedure used is valid for small, known measurement error (Carroll et al. 2006). The measurement error and misclassification were estimated from a reliability sample for this application and the mismeasurement does not appear to be minor based on these estimates. Therefore, we would expect that the SIMEX standard errors are underestimated and that SIMEX would provide only a partial correction for mismeasurement.

Another reason why there may be a difference between results based on the maximum likelihood approach and SIMEX is that, unlike the likelihood approach, SIMEX does not require any assumptions regarding the underlying distribution of $X|\boldsymbol{Z}$. If the maximum likelihood estimators are affected by misspecification of this distribution, then we would expect the SIMEX and likelihood estimates to differ when the $X|\boldsymbol{Z}$ distribution is misspecified. Interestingly enough, Huang et al. 2006 describe a procedure that is similar to SIMEX to examine the sensitivity of assumptions in structural measurement error models (e.g. likelihood approach). They argue that if the assumed distribution of $X|\boldsymbol{Z}$ is not appropriate in that it introduces asymptotic bias in the estimators of the parameters of interest, then as σ_U^2 increases, the resulting bias will increase in magnitude. This can be investigated empirically by assuming different measurement error variances and, as in SIMEX, taking the average of B maximum likelihood estimates at each level of σ_U^2 and plotting the average of these estimates against σ_U^2. A nonconstant relationship would suggest that the assumed model for $X|\boldsymbol{Z}$ is not robust.

Ideally, supplementary data in the form of internal validation data would be available for a large number of subjects. Ten patients assessed by ten physicians is not a very

large dataset. However, this study was not originally designed for use in an errors-in-variables analysis, but rather was designed to provide information on the extent to which different rheumatologists could agree on the measurements of different signs or symptoms of patients. With a high degree of agreement on particular measures there is rationale for considering these as the basis for outcomes in multi-center trials. Even though it was not conducted for this purpose, data arising from this study are useful in providing information regarding the error distribution as was demonstrated above. To study this claim, Table 2.18 summarizes the results of a small simulation study which compares the performance of the correct likelihood and SIMEX methods based on a small external reliability study generated similarly to that which was available in this application. The SAS code from the simulations discussed earlier in this chapter was used, but the parameter configurations were chosen to be close to the values that were observed in the PsA application.

These results demonstrate much poorer performance than those based on simulations using larger supplementary datasets. The biases are larger here and the empirical coverage probabilities are considerably farther away from the nominal values of 0.95. Although estimators based on a correct likelihood approach do appear to result in smaller biases and larger empirical coverage probabilities, there still appears to be some bias present and the empirical coverage probabilities are less than 0.95. Based on these results, the SIMEX approach only provides a partial correction for the bias. This is consistent with the simulation results. Estimators based on a correct likelihood approach and SIMEX would likely demonstrate better performance for larger reliability studies. As in the validation data generated for the simulations, the larger the reliability sample, the less sampling variability will be present in the estimators for the mismeasurement and covariate distribution parameters. The PsA reliability study was not designed with this purpose in mind. In planning similar studies in the future, it would be best to strive to obtain a large internal validation sample. If this is not feasible, a large reliability dataset would also provide valuable information regarding the mismeasurement and covariate distributions required to account for measurement error or misclassification in covariates.

In this chapter, it has been demonstrated that mismeasured covariates induce bias

Table 2.18: *Empirical performance of estimators of the regression parameters associated with X and Z based on parameter values close to those observed in PsA application; Number of assessments are POI* (10)*;* $\rho = 0.06$, $\kappa = 1.75$, $\beta_X = -2$, $\beta_Z = -1.5$, *reliability sample of 10 independent observations on 10 subjects Binary X:* $P(Z = 1) = 0.5$, $P(X = 1|Z) = 0.6$, $\pi_{00} = 0.9$ *and* $\pi_{11} = 0.7$ *Continuous X:* $Z \sim N(0, 1.5)$, $X|Z \sim N(-0.03Z, 0.41)$, $\gamma = 0.41$.

Method		Binary X				Continuous X			
		β_X		β_Z		β_X		β_Z	
		Weibull	PCBH	Weibull	PCBH	Weibull	PCBH	Weibull	PCBH
Naive	Bias	1.2206	1.2339	0.3585	0.3844	1.3870	1.3986	0.3957	0.4157
	SE_1	0.1184	0.1187	0.1209	0.1219	0.0812	0.0808	0.0820	0.0807
	SE_2	0.1331	0.1197	0.1455	0.1294	0.0880	0.0875	0.0895	0.0834
	ECP	0.0000	0.0000	0.1901	0.1604	0.0000	0.0000	0.0183	0.0028
Likelihood	Bias	-0.1811	0.0453	-0.0701	0.0562	-0.2206	0.3272	-0.0553	0.1749
	SE_1	0.2895	0.3258	0.1820	0.1714	0.4918	0.2884	0.2123	0.1302
	SE_2	0.4814	0.4883	0.2187	0.1798	1.1177	0.6831	0.3401	0.1370
	ECP	0.8489	0.8912	0.9151	0.9163	0.7862	0.5449	0.8839	0.6597
SIMEX		Exponential				Cubic			
	Bias	-1.0184	-1.2156	0.3426	0.3763	-0.5977	0.8628	-0.4158	0.3285
	SE_1	0.7257	0.5871	0.1363	0.1207	0.2755	0.1096	2.6196	0.1111
	SE_2	1.5740	1.6592	0.1358	0.1319	0.2972	0.2260	4.0708	0.1241
	ECP	0.5387	0.4494	0.2744	0.1582	0.3672	0.0299	0.9938	0.1851

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)

in estimators which treat them as precisely measured when modeling interval-censored lifetime data. The performance of likelihood and SIMEX approaches accounting for the mismeasurement have been examined and applied to data from a study on PsA. In the next chapter, this work will be extended to a progressive three-state model.

Chapter 3

Interval-censored Three-state Data with Mismeasured Covariates

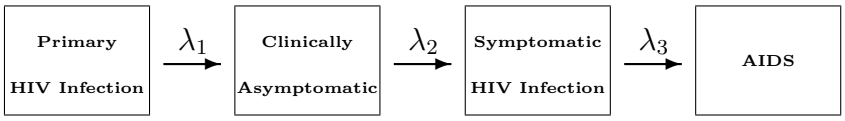
3.1 Overview

Clinical trials of progressive diseases (e.g., HIV-AIDS) are often conducted to estimate rates of transitions between disease states and the effects of covariates on these transition rates. Consider the multi-state model in Figure 3.1 representing the progression of HIV-AIDS (Toronto General Hospital University Health Network 2005). Information on covariates such as CD4 cell count or viral load may be collected in a study of this disease, but measurements on both are known to be error-prone. It has been well established that naive regression analyses based on measured values can lead to seriously biased estimators and misleading standard errors in generalized linear models and survival models with right-censored data. In Chapter 2 we considered the impact of covariate mismeasurement on interval-censored lifetime data. To date there does not appear to be much research that specifically addresses mismeasured covariates in the context of interval-censored multi-state models.

The purpose of this chapter is to explore the effects of covariate mismeasurement on the estimation of regression parameters and to propose and evaluate methods to account for this mismeasurement problem. This methodology will be applied to the motivating

study described in Chapter 2.

Figure 3.1: *Progressive model for HIV-AIDS involving three transition intensities.*



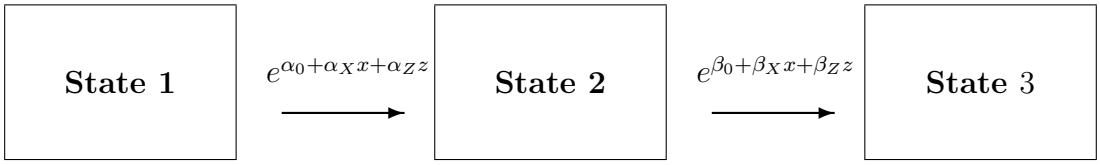
3.2 Impact of Ignoring Error in Covariates

The following notation will be used throughout this discussion. Let

- $i = 1, 2, ..., n$ index the subjects under observation,
- $j = 1, 2, ..., m_i$ index the observation times for subject i,
- $k = 1, 2$ index the different transitions,
- $\boldsymbol{y_i} = (y_i(u_{i1}), y_i(u_{i2}), y_i(u_{i3}), ..., y_i(u_{im_i}))'$ represent the observed states at the m_i observation times, $u_{i1}, u_{i2}, ..., u_{im_i}$ for subject i,
- $\boldsymbol{x_i}$ be a $(p_x \times 1)$ covariate vector which is imperfectly measured,
- $\boldsymbol{w_i}$ be the mismeasured version of $\boldsymbol{x_i}$, and
- $\boldsymbol{z_i}$ be a perfectly measured $(p_z \times 1)$ covariate vector.

For simplicity, we will consider the effects of mismeasured covariates on estimation for a progressive three-state model given in Figure 3.2 rather than the four-state model of Gladman et al. (1995). However, extensions to models with more than three states are straightforward. Also, we will restrict our attention to one-dimensional X, W, and Z here.

Figure 3.2: *A time homogeneous three-state progressive model with multiplicative covariate effects.*



If modeling is based on the structure of Figure 3.2, a naive maximum likelihood function would be as follows:

$$\mathcal{L}\left(\boldsymbol{\theta}^*\right) = \prod_{i=1}^{n}\prod_{j=1}^{m_i} P_{y_i(u_{i,j-1}),y_i(u_{ij})}(u_{ij} - u_{i,j-1}|w_i, z_i; \boldsymbol{\theta}^*), \tag{3.1}$$

where the transition probabilities are given by (1.24). Specifically, for this three state model, the probabilities are as follows:

$$\begin{aligned}
P_{1,1}(t;\boldsymbol{\theta}^*) &= \exp\left[-\exp\left(\alpha_0^* + \alpha_X^* w_i + \alpha_Z^* z_i\right) t\right], \\
P_{1,2}(t;\boldsymbol{\theta}^*) &= \frac{\exp\left(\alpha_0^* + \alpha_X^* w_i + \alpha_Z^* z_i\right)}{\exp\left(\beta_0^* + \beta_X^* w_i + \beta_Z^* z_i\right) - \exp\left(\alpha_0^* + \alpha_X^* w_i + \alpha_Z^* z_i\right)} \cdot \\
&\quad \left\{\exp\left[-\exp\left(\alpha_0^* + \alpha_X^* w_i + \alpha_Z^* z_i\right) t\right] - \exp\left[-\exp\left(\beta_0^* + \beta_X^* w_i + \beta_Z^* z_i\right) t\right]\right\}, \\
P_{1,3}(t;\boldsymbol{\theta}^*) &= 1 - P_{1,1}(t;\boldsymbol{\theta}^*) - P_{1,2}(t;\boldsymbol{\theta}^*), \\
P_{2,2}(t;\boldsymbol{\theta}^*) &= \exp\left[-\exp\left(\beta_0^* + \beta_X^* w_i + \beta_Z^* z_i\right) t\right], \\
P_{2,3}(t;\boldsymbol{\theta}^*) &= 1 - \exp\left[-\exp\left(\beta_0^* + \beta_X^* w_i + \beta_Z^* z_i\right) t\right], \\
P_{3,3}(t;\boldsymbol{\theta}^*) &= 1.
\end{aligned}$$

A "$*$" is attached to the parameters in this model to emphasize that they differ from the true model parameters in Figure 3.2 when the true covariate x_i is replaced by the measured version, w_i. We note that in this formulation, we assume the assessment scheme is noninformative as outlined in Grüger et al. (1991) and the form of the transition intensities is specified correctly with the exception of the mismeasured covariates.

Maximization of (3.1) will result in estimates for $\boldsymbol{\theta}^*$, a vector of parameters that are possibly different from the parameters of interest, $\boldsymbol{\theta}$. Since the estimators for $\boldsymbol{\theta}^*$ are based

on mismeasured covariates, they will potentially be biased for $\boldsymbol{\theta}$. Asymptotic biases provide insight regarding the impact of mismeasured covariates. Since the asymptotic biases are complicated functions of $\boldsymbol{\theta}$ and the covariate and measurement error distributions it is difficult to derive closed-form expressions. We can investigate the relationship graphically as in Chapter 2 for different parameter configurations (White 1982). Note that the multistate formulation means that some unique measurement error problems can arise when dealing with interval-censored transition times. For example, if $\beta_X = 0$ in the model for Figure 3.2, then with right-censored data we would not expect any bias in the estimator for β_Z because the likelihood can be factorized. With interval-censored transition times this factorization is not possible and the measurement error in X can even impact parameter estimation in transition rates where X does not appear in the linear predictor. For this reason, covariate measurement error or misclassification can have a wide ranging impact in these more involved models.

In creating the following plots, all subjects are assumed to begin in state 1 at $t = u_{i0} = 0$ (i.e. $y_i(u_{i0}) = 1$) and to be assessed at five equally spaced assessment times. The study duration, τ, was selected such that $P_{13}(\tau|X, Z; \boldsymbol{\theta})$ was at least 0.6 for all combinations of (X, Z) (binary covariates) or such that $P_{13}(\tau|X, Z; \boldsymbol{\theta}) = 0.6$ at $\boldsymbol{\mu}' = (\mu_X, \mu_Z)' = (0, 0)'$ (continuous (X, Z)). With $\boldsymbol{Y}$ representing the states occupied at the assessment times, the data $(\boldsymbol{Y}, X, W, Z)'$ are assumed to be i.i.d. across patients. The equations to be solved to determine the limiting values of the naive maximum likelihood estimators were determined based on (2.6) and (3.1) as follows. The regularity conditions outlined in White (1982), including that the derivative and the expectation operators can be interchanged, lead to the equation

$$\frac{\partial E_{Y,W,X,Z|Y_0}\left[l_{naive}\left(\boldsymbol{\theta}^*\right)\right]}{\partial \boldsymbol{\theta}^*} = \mathbf{0} \tag{3.2}$$

following from (2.6). The value of $\boldsymbol{\theta}^*$ satisfying (3.2) can be equivalently obtained by determining the value of $\boldsymbol{\theta}^*$ that maximizes $E_{Y,W,X,Z|Y_0}\left[l_{naive}\left(\boldsymbol{\theta}^*\right)\right]$, provided some conditions are satisfied. As in Section 2.3 let $\mathcal{P}$ represent the set of all possible values of $\boldsymbol{Y}$ (i.e. all possible state paths) and $\boldsymbol{v}$ be a six dimensional vector. Then, the function to be maximized with respect to θ^* can be written as follows:

$$
\begin{aligned}
& E_{Y,W,X,Z|Y_0}\left[l_{naive}\left(\boldsymbol{\theta^*}\right)\right] \\
= \; & E_{Y,W,X,Z|Y_0}\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}\log\left[P_{v_{j-1},v_j}(u_{ij}-u_{i,j-1}|\boldsymbol{W_i},\boldsymbol{Z_i};\boldsymbol{\theta}^*)\right]^{I(\boldsymbol{Y_i}=\boldsymbol{v})}\right\} \\
= \; & E_{Y,W,X,Z|Y_0}\left\{\sum_{i=1}^{n} I\left(\boldsymbol{Y_i}=\boldsymbol{v}\right)\left\{\sum_{j=1}^{5}\log\left[P_{v_{j-1},v_j}(\tau/5|\boldsymbol{W_i},\boldsymbol{Z_i};\boldsymbol{\theta}^*)\right]\right\}\right\} \qquad (3.3) \\
= \; & nE_{W,X,Z}\left\{E_{Y|W,X,Z,Y_0}\left\{I\left(\boldsymbol{Y}=\boldsymbol{v}\right)\left[\sum_{j=1}^{5}\log\left[P_{v_{j-1},v_j}(\tau/5|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*)\right]\right]\right\}\right\} \\
= \; & nE_{W,X,Z}\left\{\sum_{\boldsymbol{v}\in\mathcal{P}}\left[P\left(\boldsymbol{Y}=\boldsymbol{v}|\boldsymbol{X},\boldsymbol{Z},Y_0;\boldsymbol{\theta}^*\right)\sum_{j=1}^{5}\log\left[P_{v_{j-1},v_j}(\tau/5|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*)\right]\right]\right\} \\
= \; & nE_{W,X,Z}\left\{\sum_{\boldsymbol{v}\in\mathcal{P}}\left[\prod_{j=1}^{5}P_{v_{j-1},v_j}(\tau/5|\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\theta})\sum_{j=1}^{5}\log\left(P_{v_{j-1},v_j}(\tau/5|\boldsymbol{W},\boldsymbol{Z};\boldsymbol{\theta}^*)\right)\right]\right\}.
\end{aligned}
$$

The sixth line follows from the fact that we are considering nondifferential mismeasurement here so $f_{Y|Y_0,W,X,Z}\left(\cdot\right) = f_{Y|Y_0,X,Z}\left(\cdot\right)$. Also, in the above formulation, it is assumed that the distribution of (W, X, Z) does not depend on Y_0. The expectation $E_{W,X,Z}$, and therefore, the form of the objective function depend on whether we are dealing with binary or continuous covariates. This expectation is simply a sum for the binary covariates, but is an integral that has no closed form in general for the continuous covariates. In creating the plots that will follow, these functions were maximized using PROC NLP and PROC NLMIXED in SAS for binary and continuous covariates, respectively.

3.2.1 Binary Covariates

Here we let X, Z and W be binary covariates so, (3.3) becomes

$$
\begin{aligned}
& \sum_{x=0}^{1}\sum_{w=0}^{1}\sum_{z=0}^{1}P\left(X=x,W=w,Z=z\right)\sum_{\boldsymbol{v}\in\mathcal{P}}\left\{\prod_{j=1}^{5}P_{v_{j-1},v_j}(\tau/5|X=x,Z=z;\boldsymbol{\theta})\right. \\
& \left.\cdot\sum_{j=1}^{5}\log\left[P_{v_{j-1},v_j}(\tau/5|W=w,Z=z;\boldsymbol{\theta}^*)\right]\right\},
\end{aligned}
$$

where $P(X = x, W = w, Z = z) = P(Z = z)\, P(X = x|Z = z)\, P(W = w|X = x, Z = z)$, $P(Z = 1) = 0.5$, $P(X = x|Z = z) = e^{\log(2)z}/\left(1 + e^{\log(2)z}\right)$ and $P(W = w|X = x, Z = z) = P(W = w|X = x)$ is defined by the misclassification probabilities, $\pi_{01} = P(W = 0|X = 1)$ and $\pi_{10} = P(W = 1|X = 0)$. Maximization of this function with respect to

$$\boldsymbol{\theta}^* = (\alpha_0^*, \alpha_X^*, \alpha_Z^*, \beta_0^*, \beta_X^*, \beta_Z^*)'$$

will give the limiting values of the naive estimators.

Figures 3.3 to 3.6 illustrate the asymptotic bias based on several parameter configurations that may be encountered in practice. As one would expect, it is clear from these plots and the simulation results that follow that the magnitude of the bias increases as the misclassification increases in severity. However, from these plots it appears that π_{00} (or equivalently, π_{10}) appears to have less of an impact on bias than π_{11} (or π_{01}). For instance, the asymptotic bias summarized in the plots is larger in magnitude when $\pi_{00} = 1$ and $\pi_{11} = 0.7$ than when $\pi_{00} = 0.7$ and $\pi_{11} = 1$.

The difference between the asymptotic bias summarized in Figures 3.3 and 3.4 appears to be negligible based on these scales for this parameter configuration. This suggests that the relative magnitude of β_0 to α_0 does not have a substantial impact on the asymptotic bias in the naive estimators. Based on Figures 3.3 and 3.5, it appears that the underlying true values of α_X and β_X appear to affect the extent of the bias observed in the six naive estimators; the bias is larger in magnitude for $\alpha_X = \beta_X = \log(2)$ than for $\alpha_X = \beta_X = \log(1.25)$. This relationship is examined further in Figure 3.6 where the asymptotic biases in the naive estimators are plotted against $\alpha_X = \beta_X$. Based on this plot, bias in the naive estimators for α_0 and β_0 appears to be more severe when $\alpha_X = \beta_X < 0$ than when $\alpha_X = \beta_X > 0$ and increases as the magnitude of the true underlying effects increase. Not surprisingly, the magnitude of biases in the naive estimators for α_X and β_X, the effects associated with the misclassified covariate, increase as the magnitude of $\alpha_X = \beta_X$ increases. When $\alpha_X = \beta_X < 0$, the bias in these estimators is positive; whereas, when $\alpha_X = \beta_X > 0$, the bias is negative, suggesting that, as has been observed in other mismeasured covariate problems, regression estimates are attenuated. Based on Figure 3.6, this attenuation appears to become larger as the magnitude of $\alpha_X = \beta_X$ increases. Figure

3.6 also suggests the asymptotic bias associated with naive estimators for effects on the error-free covariates is much smaller in magnitude than the bias in the naive estimators associated with the error-prone covariate effects for this particular configuration. However, there does appear to be a larger asymptotic bias present when $\alpha_X = \beta_X < 0$ than when $\alpha_X = \beta_X > 0$. It may also be of interest to investigate the effect on bias when only one quantity, either α_X or β_X, is varied. However, in practice often the simpler model assuming common covariate effects across transitions is adopted.

Since we are considering interval-censored data, one question that arises is whether a bias is introduced for estimation of the coefficient of Z in the second transition if the misclassified covariate only affects the first transition under the true and assumed model. Figures 3.7 and 3.8 illustrate the asymptotic bias of the estimators based on a naive model of the same form as the true underlying model (but with W used in place of X). These plots suggest that bias is introduced in the estimator of the effect of Z on the second transition when X does not have an effect on this transition under the true model. The magnitude of the bias depends on the true effect of X on the first transition and appears to be less than that we would observe if X had an effect on the second transition.

Another interesting question involves the impact of m, the number of assessments, on the asymptotic bias. Figures 3.9 and 3.10 compare the asymptotic bias for the naive estimators for a specific parameter configuration based on two and six equally spaced assessments. Based on these plots, there does not appear to be much of a difference in the asymptotic bias between the two assessment schemes. This is slightly counter-intuitive since the more frequent the assessments, the closer the data are to right-censored data where the factorization would suggest negligible bias would result. This could be explored further by increasing the number of inspections and assessing whether the bias decreases in the parameters associated with the second transition intensity.

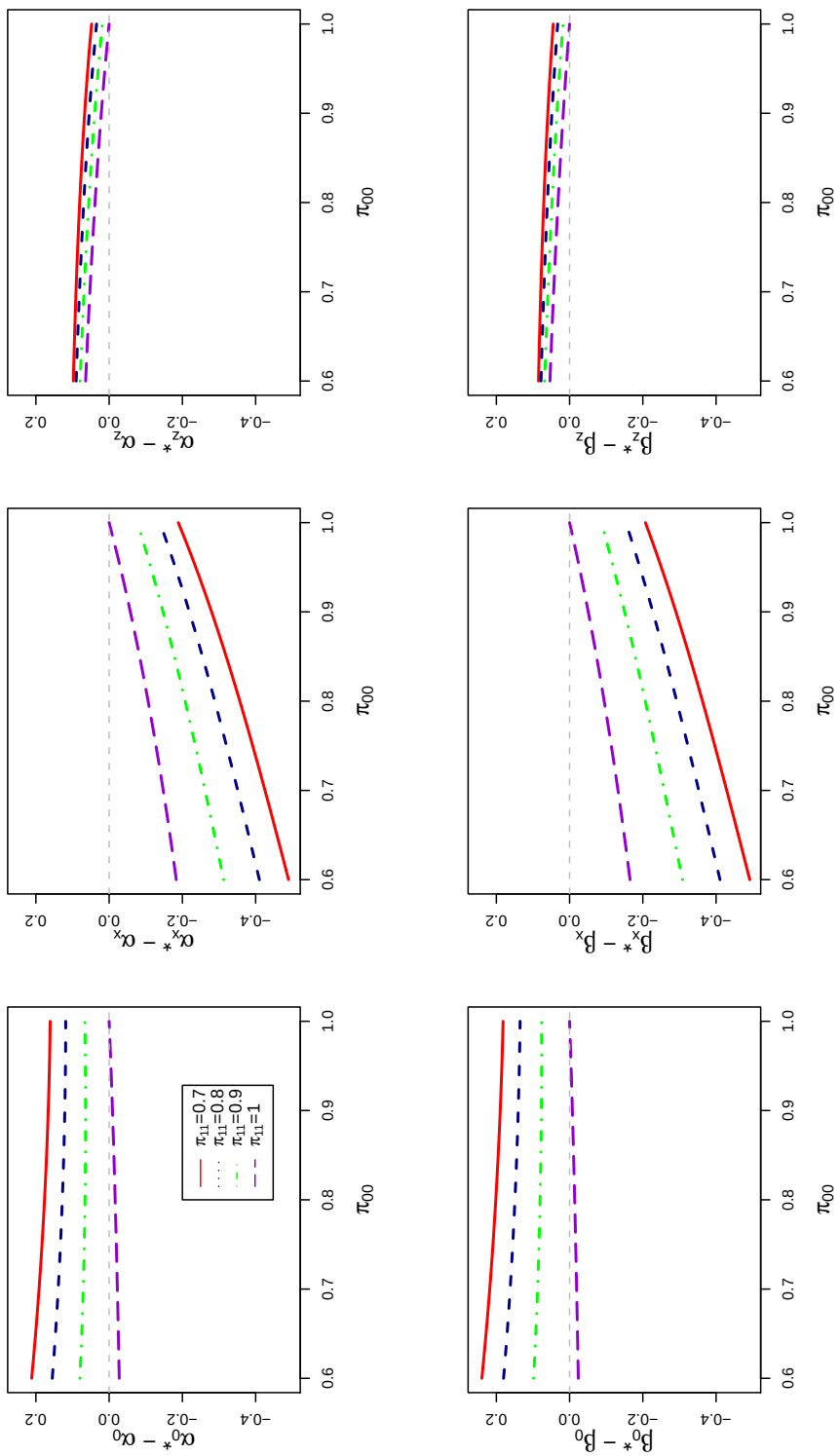
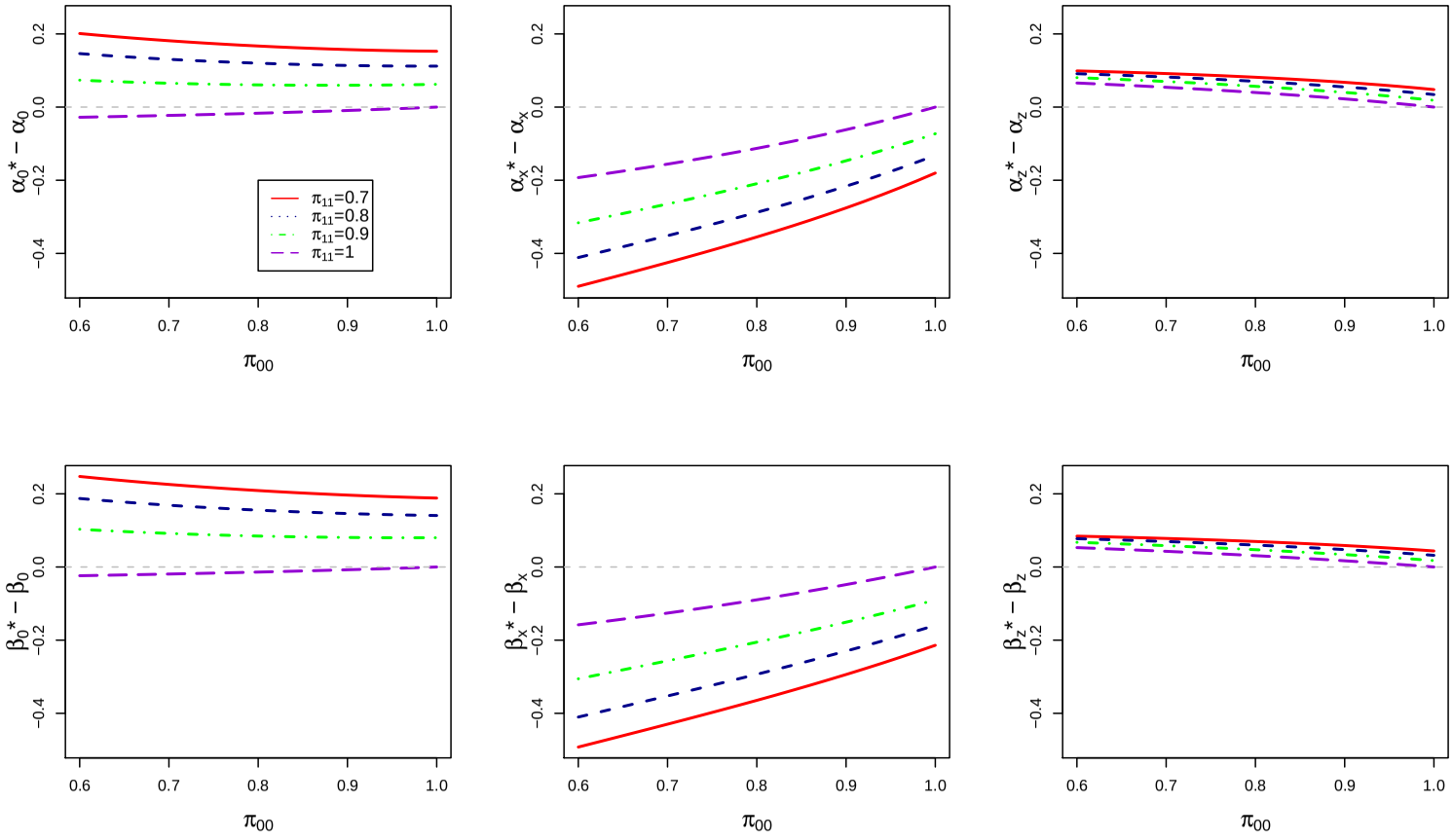


Figure 3.3: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional transition intensities model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$*; maximum right censoring rate at* τ *is 40%;* $P(Z = 1) = 0.5$ *and* $\text{logit}[P(X = 1|Z = z)] = \log(2)\, z$.

Figure 3.4: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional transition intensities model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\boldsymbol{\beta_0 = \log(0.22)}$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$*; maximum right censoring rate at* τ *is 40%;* $P(Z = 1) = 0.5$ *and* $\text{logit}\left[P(X = 1|Z = z)\right] = \log(2)\, z$.



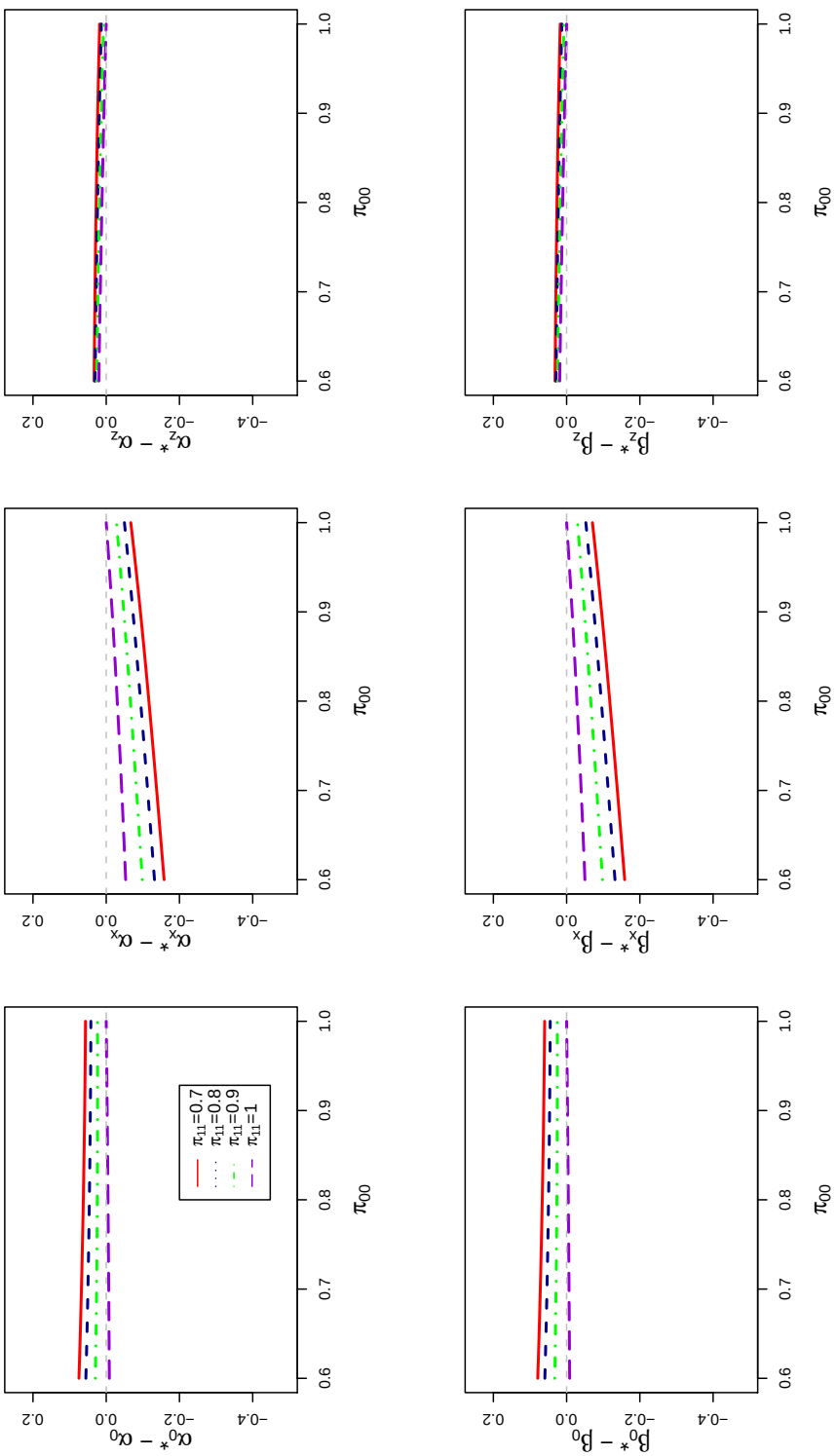
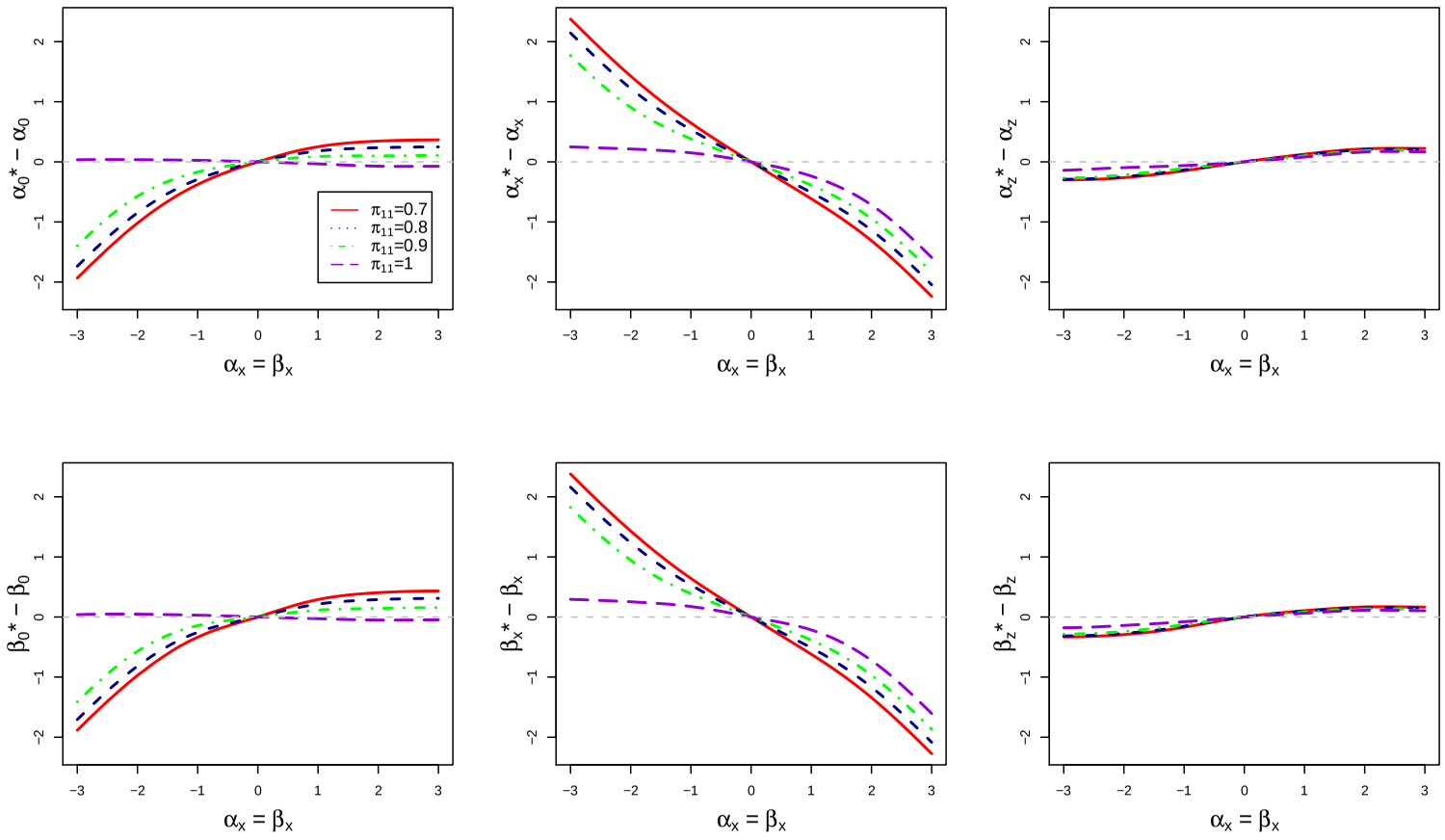


Figure 3.5: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional transition intensities model with a misclassified binary covariate; $m = 5$ equally spaced assessments; $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\boldsymbol{\alpha_X} = \boldsymbol{\beta_X} = \boldsymbol{\alpha_Z} = \boldsymbol{\beta_Z} = \mathbf{log\,(1.25)}$; maximum right censoring rate at τ is 40%; $P(Z = 1) = 0.5$ and $\text{logit}[P(X = 1|Z = z)] = \log(2)\, z$.*

Figure 3.6: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional transition intensities model with a misclassified binary covariate;* $m = 5$ *equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(1.25)$; *maximum right censoring rate at* τ *is 40%;* $P(Z = 1) = 0.5$, $\pi_{00} = 0.7$ *and* $\text{logit}[P(X = 1|Z = z)] = \log(2)\, z$.



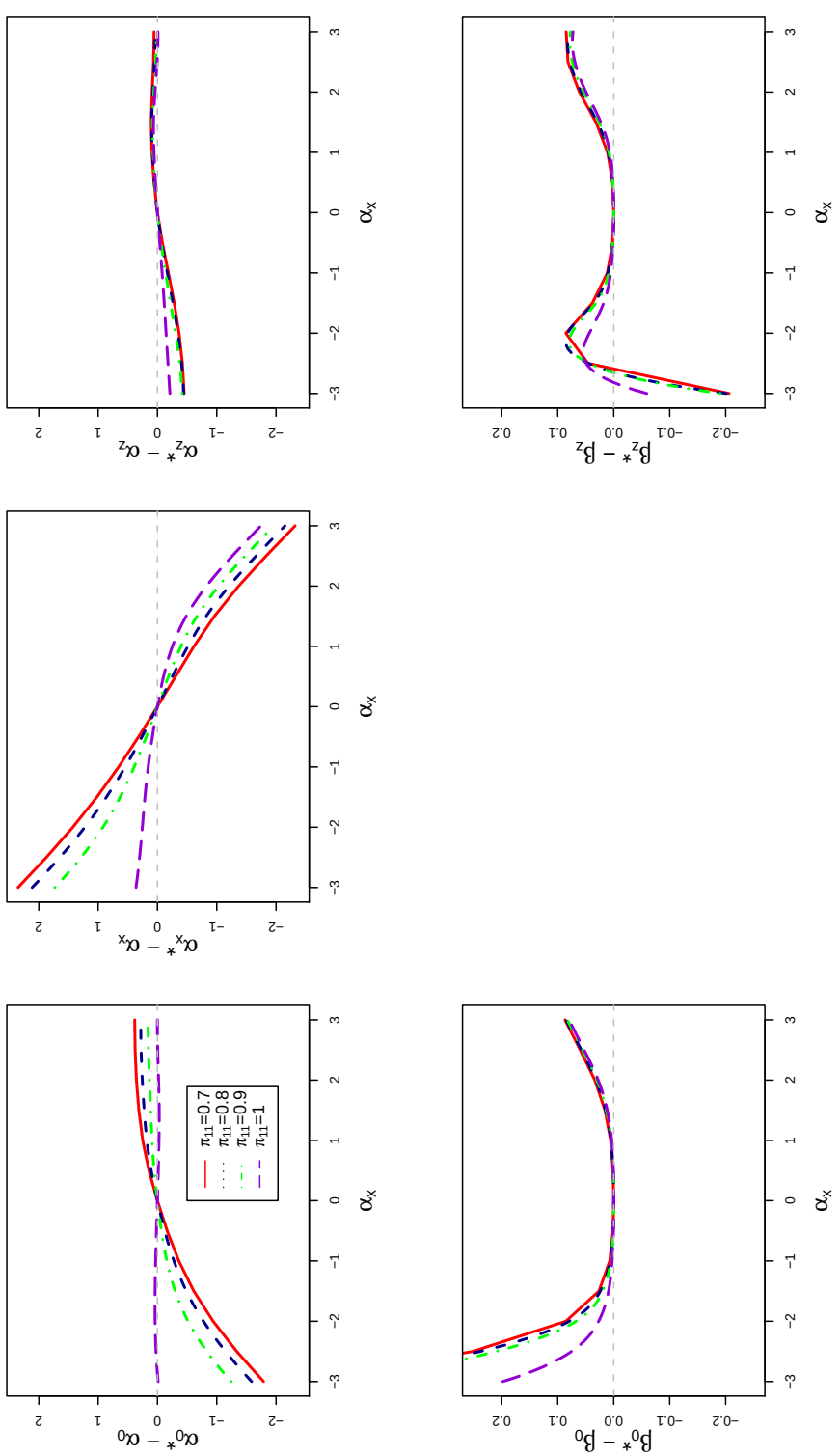
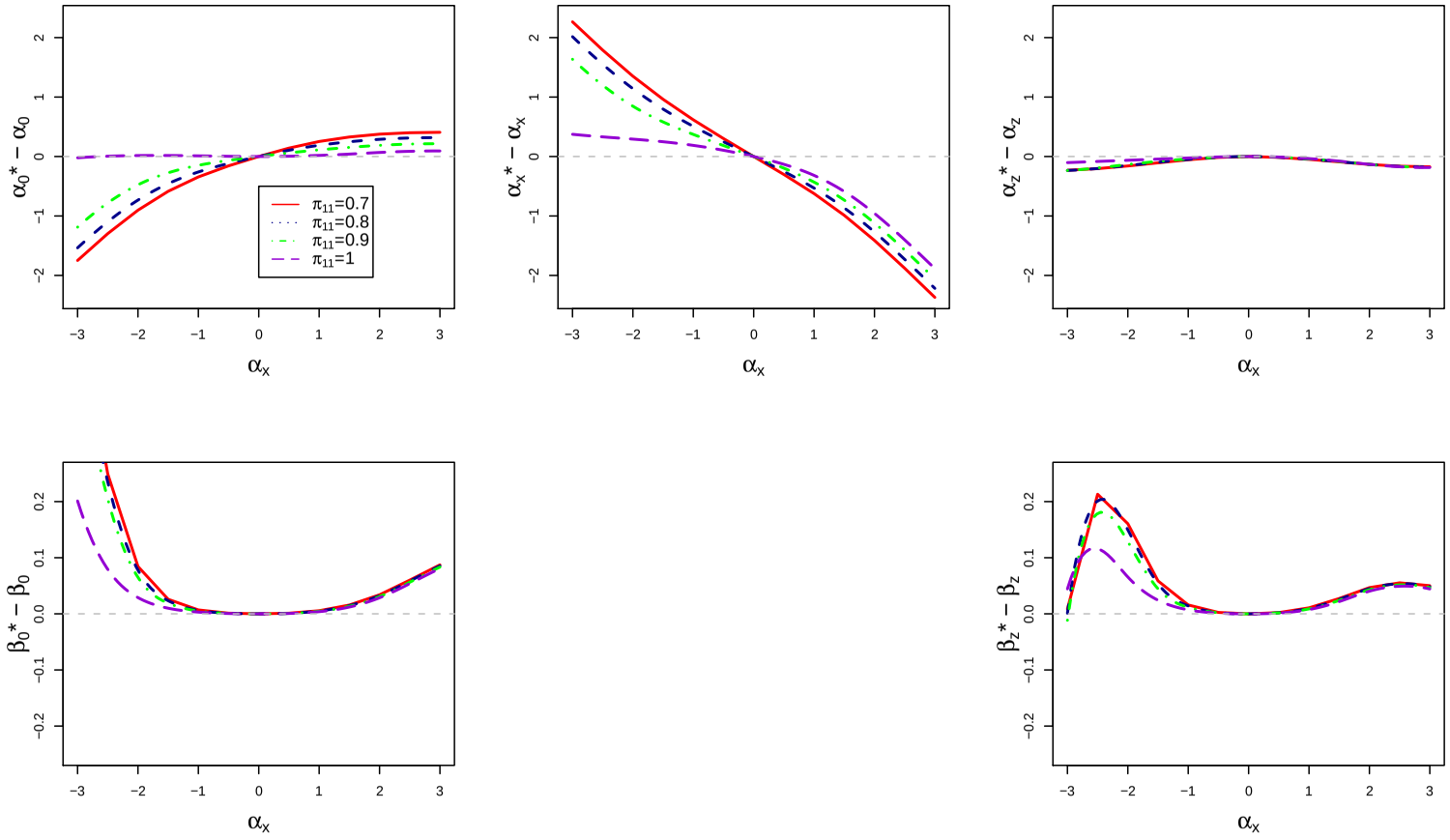


Figure 3.7: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional transition intensities model with a misclassified binary covariate affecting only the first transition;* $m = 5$ *equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(2)$; *maximum right censoring rate at* τ *is 40%;* $P(Z = 1) = 0.5$, $\pi_{00} = 0.7$ *and* $\text{logit}[P(X = 1|Z = z)] = \log(2)\, z$.

Figure 3.8: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional transition intensities model with a misclassified binary covariate affecting only the first transition;* $m = 5$ *equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(2)$*; maximum right censoring rate at* τ *is 40%;* $\boldsymbol{P(Z = 1) = P(X = 1|Z = z) = 0.5}$ *and* $\pi_{00} = 0.7$.



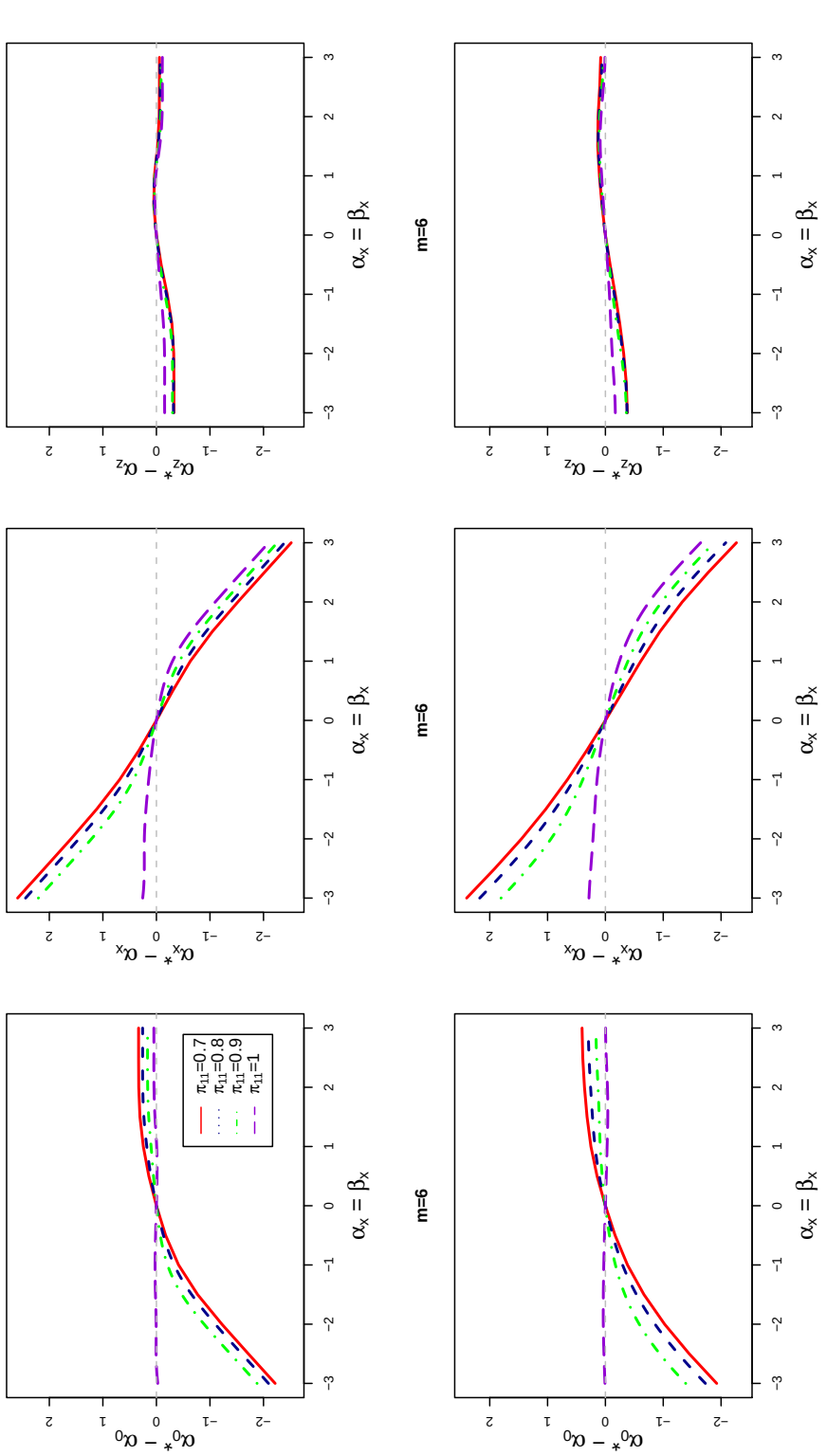
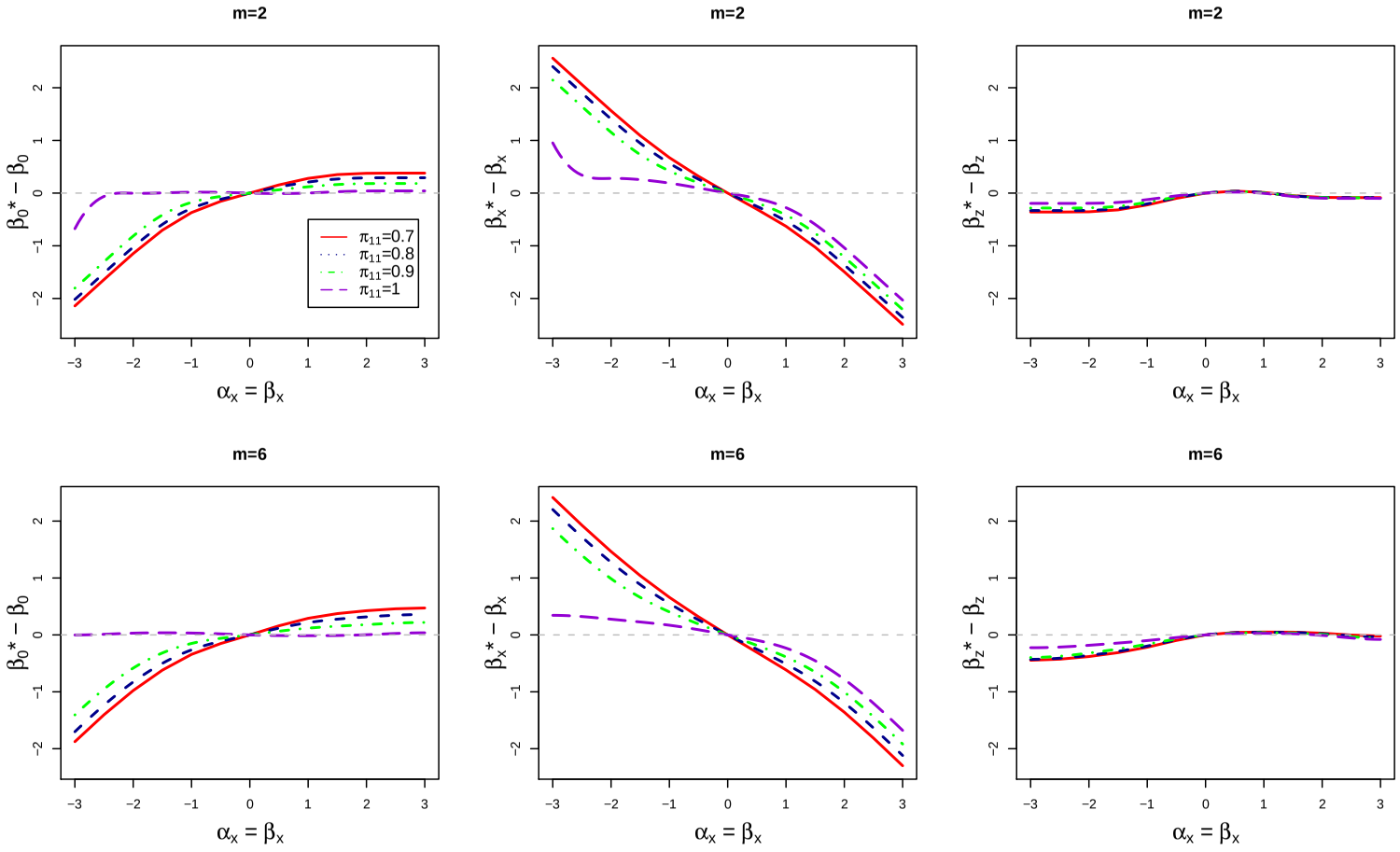


Figure 3.9: *Comparison of the asymptotic bias of naive maximum likelihood estimators for a proportional transition intensities model with a misclassified binary covariate for m=2 and 6 equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(2)$*; maximum right censoring rate at* τ *is 40%;* $P(Z=1) = 0.5$, $\pi_{00} = 0.7$ *and* $\text{logit}[P(X=1|Z=z)] = \log(2)\, z$.

Figure 3.10: *Comparison of the asymptotic bias of naive maximum likelihood estimators for a proportional transition intensities model with a misclassified binary covariate for m=2 and 6 equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(2)$*; maximum right censoring rate at* τ *is 40%;* $\pi_{00} = 0.7$ *and* $\boldsymbol{P(Z=1) = P(X=1|Z=z) = 0.5}$.



3.2.2 Continuous Covariates

The function maximized with respect to $\boldsymbol{\theta}^* = (\alpha_0^*, \alpha_X^*, \alpha_Z^*, \beta_0^*, \beta_X^*, \beta_Z^*)'$ when X, W and Z are continuous is

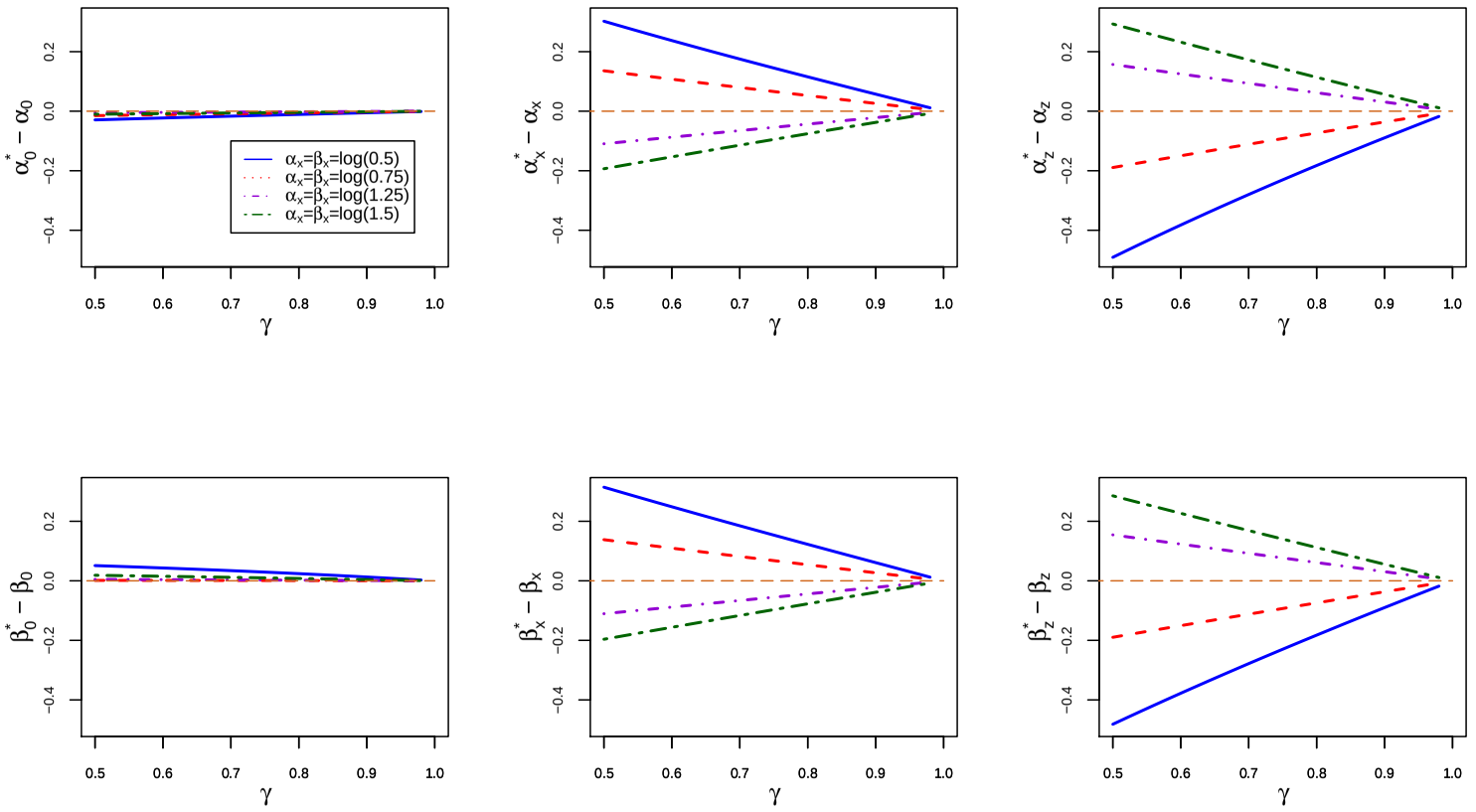
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,W,Z}\left(x,w,z\right)\left\{\sum_{\boldsymbol{v}\in\mathcal{P}}\left[\prod_{j=1}^{5} P_{v_{j-1},v_j}(\tau/5|X=x,Z=z;\boldsymbol{\theta})\right.\right.$$
$$\left.\left.\cdot\sum_{j=1}^{5}\log\left[P_{v_{j-1},v_j}(\tau/5|W=w,Z=z;\boldsymbol{\theta}^*)\right]\right]\right\}\mathrm{d}x\mathrm{d}w\mathrm{d}z.$$

Here $f_{X,W,Z}\left(x,w,z\right)$ denotes the probability density function of a trivariate normal distribution specified by the conditional distributions:

- $Z \sim N\left(0,\sigma_Z^2\right)$,
- $X|Z \sim N\left(\xi_Z Z, \sigma_{X|Z}^2\right)$, and
- $W|X,Z \sim N\left(\mu_{W|X,Z}, \sigma_U^2\right)$.

The specific parameter values used for each plot are outlined in the titles of the figures. The parameter ξ_Z was fixed such that ρ_{XZ} was 0 or 0.8 when $\sigma_{X|Z}^2 = \sigma_Z^2 = 1$ based on the expression $\rho_{XZ} = \xi_Z\sigma_Z/\sqrt{\sigma_{X|Z}^2 + \xi^2\sigma_Z^2}$. Adaptive Gaussian quadrature, as implemented in PROC NLMIXED (SAS), was used to approximate the integrals for each parameter configuration.

Figure 3.11: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards transition intensities model with a mismeasured continuous covariate;* $m = 5$ *equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(1.25)$*; right censoring rate at* τ *is 40% when evaluated at the means of* X *and* Z*;* $Z \sim N(0,1)$, $X|Z \sim N(1.33Z, 1)$ *(*$\rho_{XZ} = 0.8$*) and* $W = X + U$*, where* $U \sim N\left(0, \sigma_U^2\right)$ *(*$\sigma_U^2 = (1-\gamma)/\gamma$*).*



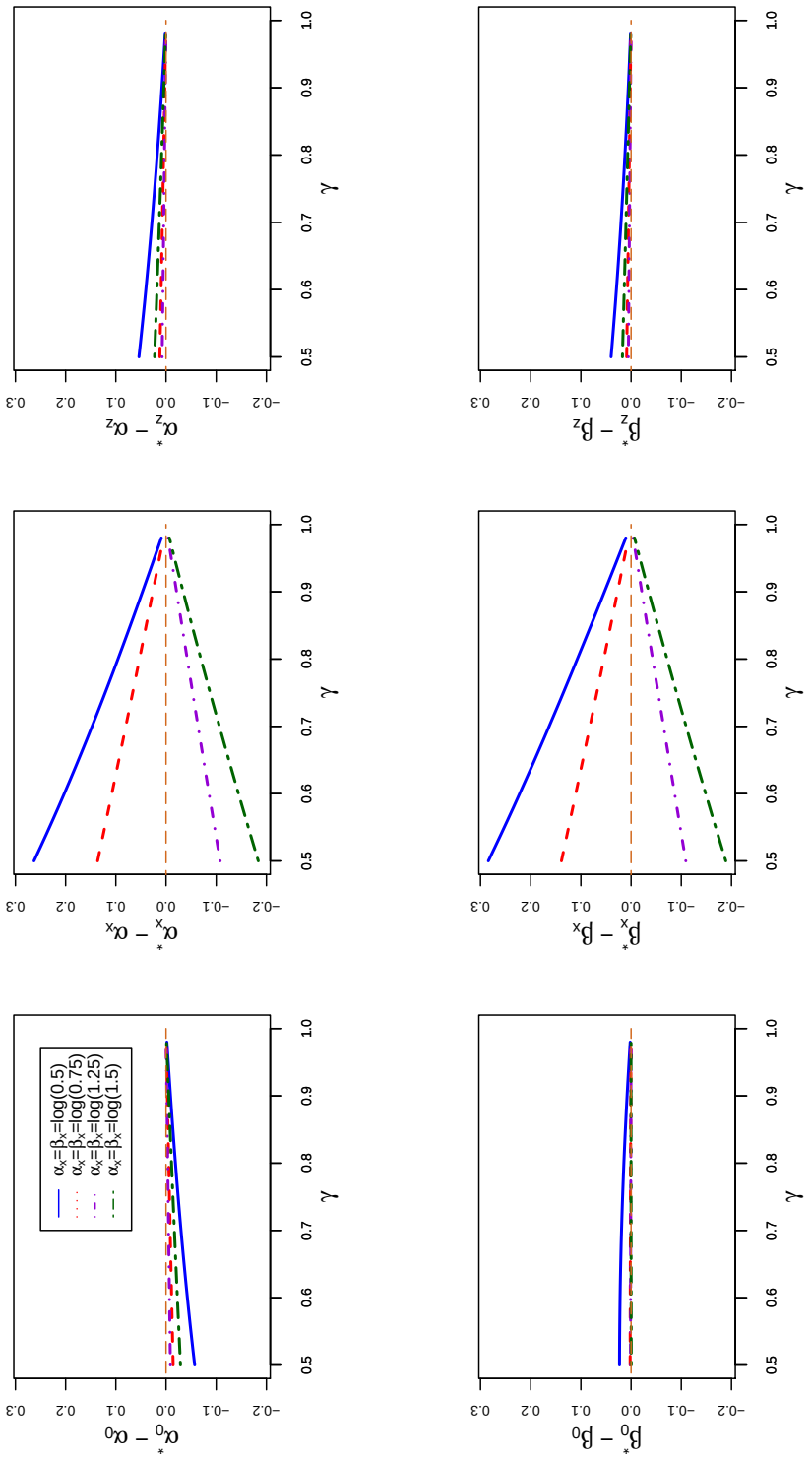
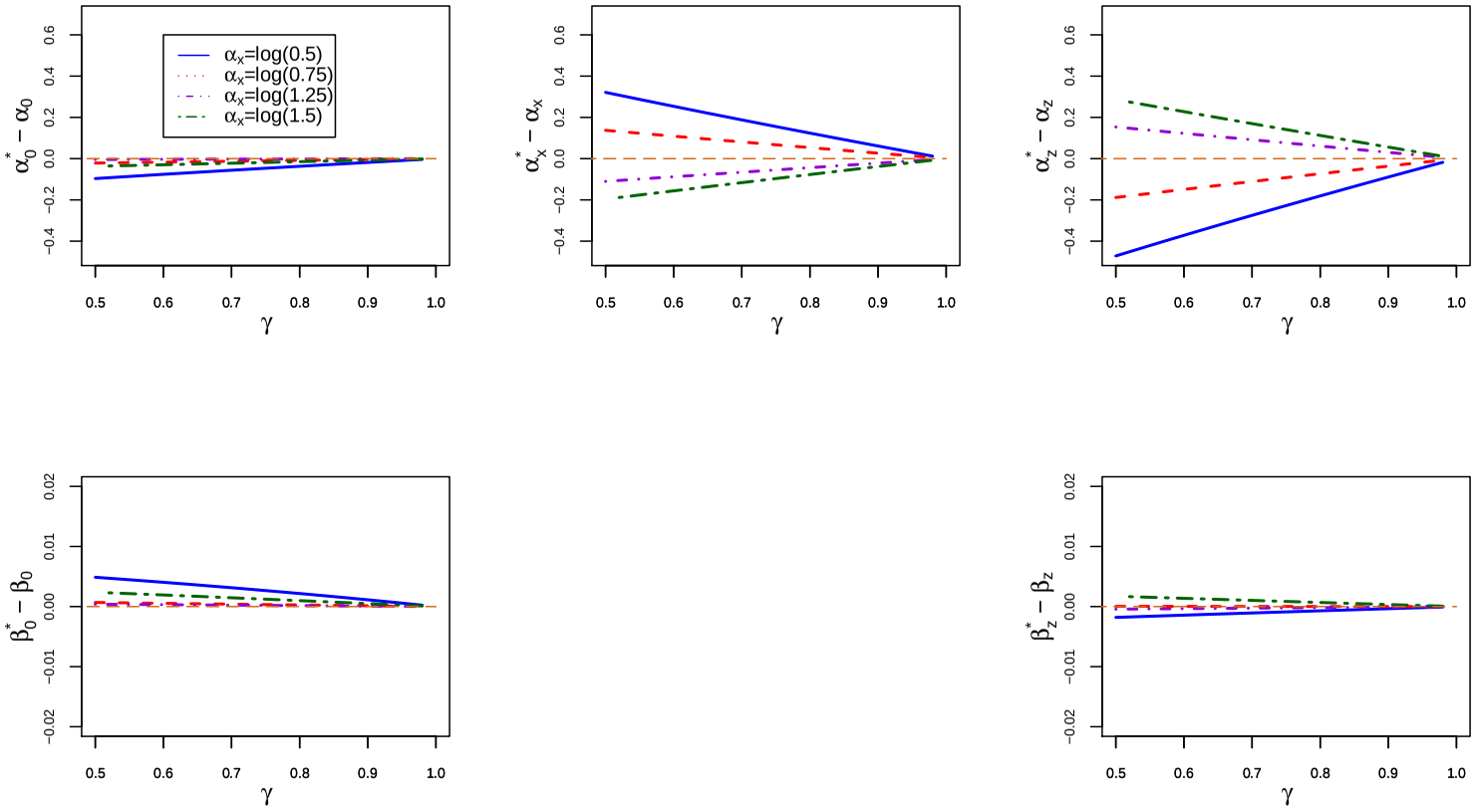


Figure 3.12: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards transition intensities model with a mismeasured continuous covariate; $m = 5$ equally spaced assessments; $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(1.25)$; right censoring rate at τ is 40% when evaluated at the means of X and Z; $Z \sim N(0,1)$, $\boldsymbol{X}|\boldsymbol{Z} \sim \boldsymbol{N}(\mathbf{0},\mathbf{1})$ $(\boldsymbol{\rho_{XZ}} = \mathbf{0})$ and $W = X + U$, where $U \sim N(0, \sigma_U^2)$ $(\sigma_U^2 = (1-\gamma)/\gamma)$.*

Figure 3.13: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards transition intensities model with a mismeasured continuous covariate affecting only the first transition;* $m = 5$ *equally spaced assessments;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(1.25)$*;right censoring rate at* τ *is 40% when evaluated at the means of* X *and* Z*;* $Z \sim N(0,1)$, $X|Z \sim N(1.33Z, 1)$ *(*$\rho_{XZ} = 0.8$*) and* $W = X + U$*, where* $U \sim N\left(0, \sigma_U^2\right)$ *(*$\sigma_U^2 = (1-\gamma)/\gamma$*).*



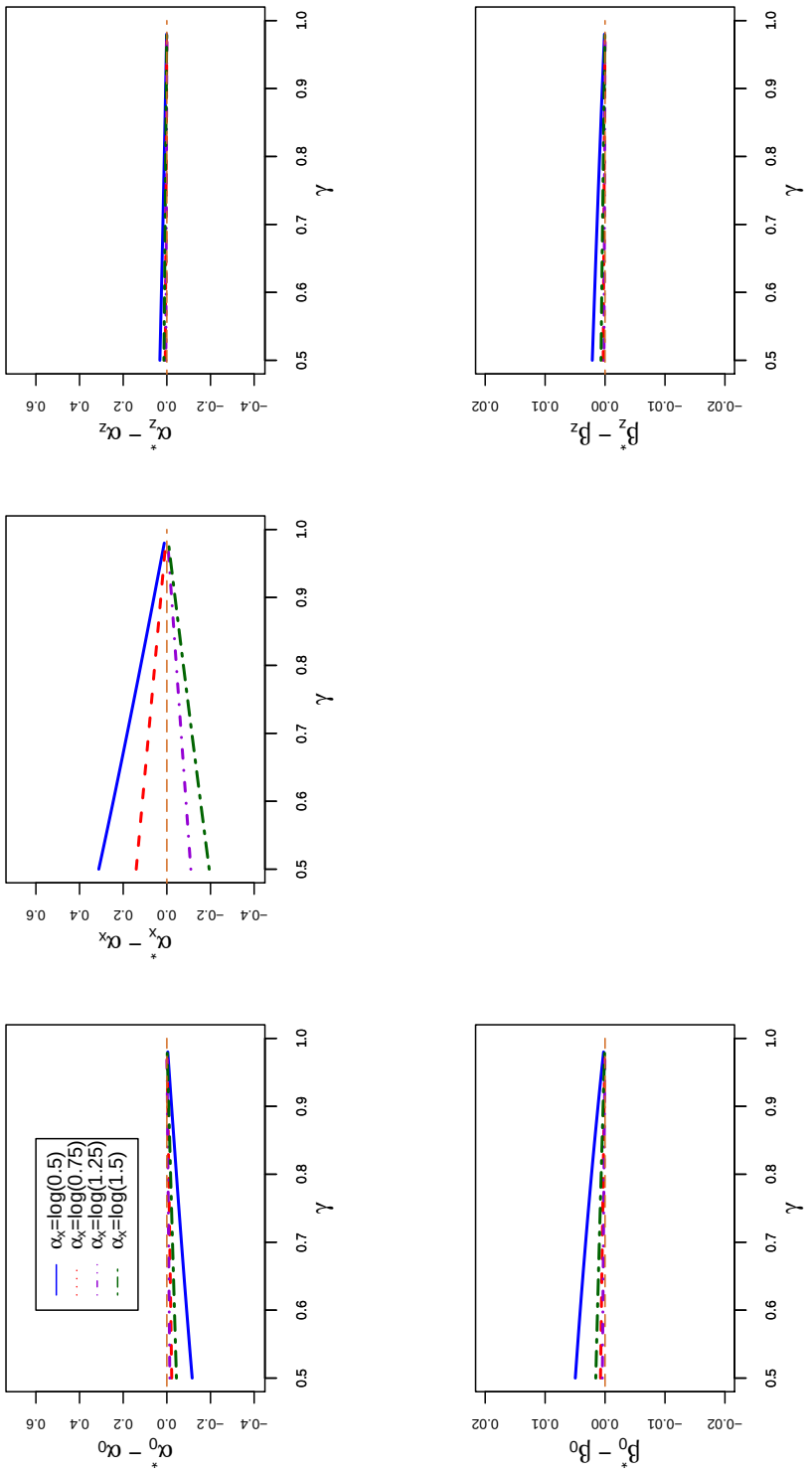


Figure 3.14: *Plot of the asymptotic bias of naive maximum likelihood estimators for a proportional hazards transition intensities model with a mismeasured continuous covariate affecting only the first transition; $m = 5$ equally spaced assessments; $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_Z = \beta_Z = \log(1.25)$;right censoring rate at τ is 40% when evaluated at the means of X and Z; $Z \sim N(0,1)$, $\boldsymbol{X}|\boldsymbol{Z} \sim \boldsymbol{N(0,1)}$ $(\boldsymbol{\rho_{XZ} = 0})$ and $W = X + U$, where $U \sim N\left(0, \sigma_U^2\right)$ $(\sigma_U^2 = (1-\gamma)/\gamma)$.*

As expected, the plots given in Figures 3.11 to 3.14 demonstrate that the asymptotic biases in the naive estimators increase in magnitude as the reliability ratio, γ, decreases (or equivalently, as the measurement error increases). There appears to be substantial asymptotic biases in the naive estimators associated with X and Z, but a lesser degree of bias for the baseline intensity estimators. As was observed in the binary covariate case, it appears that the magnitude of the asymptotic bias depends on the magnitude of $\alpha_X = \beta_X$. The biases look to be smaller when $\alpha_X = \beta_X = \log(1.25) \approx 0.2231$ than they are when $\alpha_X = \beta_X = \log(0.5) \approx -0.6931$. When the error-free covariate, Z, is uncorrelated with X, the asymptotic biases in the estimators of effects on Z seem to be relatively small. However, there still appears to be some bias present in the estimators corresponding to the Z effects which increases in magnitude as γ decreases and as the magnitude of the true values of α_X and β_X increase. When X and Z are highly correlated, there appears to be considerable bias in the estimators associated with Z in addition to those associated with X. Based on the parameter configurations explored, α_X and β_X seem to be underestimated (in absolute value) by the naive maximum likelihood approach; whereas, the magnitude of the Z effect parameters seem to be overestimated sometimes and underestimated sometimes. Figures 3.13 and 3.14 address the question of whether estimation of the effect of Z on the second transition is affected when X has an effect on the first, but not the second transition for a particular configuration. Based on these plots, there does appear to be asymptotic bias in the naive estimators of both the intercept and the effect of Z on the second transition, although it appears to be quite small. This bias appears to be present regardless of whether $\rho_{XZ} = 0$ or $\rho_{XZ} = 0.8$.

3.3 Correcting for Mismeasured Covariates

The notation and methodology for a progressive multi-state model with panel data in the absence of mismeasured covariates were outlined in Chapter 1. When model (1.13) is appropriate, an illustration of the observation and state path for an arbitrary individual is given in Figure 1.7 and the complete data likelihood is given in (1.26). Again, we assume a multiplicative model for the transition intensities model to relate the intensities to the

true covariates of interest:

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp\left\{\boldsymbol{\beta}'_{xk}\boldsymbol{x_i} + \boldsymbol{\beta}'_{zk}\boldsymbol{z_i}\right\}, \tag{3.4}$$

where $\lambda_{ik}(t)$ is the transition intensity associated with the $k \to k+1$ transition for subject i at time t and the baseline intensity, $\lambda_{0k}(t)$, is piecewise constant as in (1.14). As before we assume the assessment scheme is noninformative as outlined in Grüger et al. (1991). The necessary notation and the description of the SIMEX procedure and maximum likelihood approaches are given in Section 2.4. The maximum likelihood and SIMEX approaches follow similar steps here. However, since we are now considering three states the likelihood function is more complicated. The $f_{Y|X,Z}$ term, which is needed to proceed with the SIMEX procedure and appears in the correct likelihood function, is now given by (3.1). The SIMEX approach will involve repeated maximization of (3.1) based on simulated data. Suppose we know the parameters of the mismeasurement and conditional distribution of X given Z and let the transition probabilities be given by (1.24). As mentioned previously, for a three-state progressive model these would be given as follows:

$$\begin{aligned}
P_{1,1}(t;\boldsymbol{\theta}) &= \exp\left[-\exp\left(\alpha_0 + \alpha_X x + \alpha_Z z_i\right) t\right], \\
P_{1,2}(t;\boldsymbol{\theta}) &= \frac{\exp\left(\alpha_0 + \alpha_X x + \alpha_Z z_i\right)}{\exp\left(\beta_0 + \beta_X x + \beta_Z z_i\right) - \exp\left(\alpha_0 + \alpha_X x + \alpha_Z z_i\right)} \\
&\quad \cdot \left\{\exp\left[-\exp\left(\alpha_0 + \alpha_X x + \alpha_Z z_i\right) t\right] - \exp\left[-\exp\left(\beta_0 + \beta_X x + \beta_Z z_i\right) t\right]\right\}, \\
P_{1,3}(t;\boldsymbol{\theta}) &= 1 - P_{1,1}(t;\boldsymbol{\theta}) - P_{1,2}(t;\boldsymbol{\theta}), \\
P_{2,2}(t;\boldsymbol{\theta}) &= \exp\left[-\exp\left(\beta_0 + \beta_X x + \beta_Z z_i\right) t\right], \\
P_{2,3}(t;\boldsymbol{\theta}) &= 1 - \exp\left[-\exp\left(\beta_0 + \beta_X x + \beta_Z z_i\right) t\right], \\
P_{3,3}(t;\boldsymbol{\theta}) &= 1.
\end{aligned}$$

Then, the correct likelihood function involving a misclassified binary covariate obtained by conditioning on Z is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\sum_{x=0}^{1} f_{Y|X,Z}\left(\boldsymbol{y_i}|x,z_i\right) f_{W|X,Z}\left(w_i|x,z_i\right) f_{X|Z}\left(x|z_i\right) \\
&= \prod_{i=1}^{n}\sum_{x=0}^{1}\left\{\prod_{j=1}^{m_i} P_{y_i(u_{i,j-1}),y_i(u_{ij})}(u_{ij}-u_{i,j-1}|x,z_i;\boldsymbol{\theta})\right. \\
&\qquad \left.\left[\pi_{10}^{w_i}(1-\pi_{10})^{(1-w_i)}\right]^{(1-x)}\left[\pi_{01}^{(1-w_i)}(1-\pi_{01})^{w_i}\right]^{x}\frac{\left[\mathrm{e}^{\phi_0+\boldsymbol{\phi}_Z'\boldsymbol{z}_i}\right]^{x}}{1+\mathrm{e}^{\phi_0+\boldsymbol{\phi}_Z'\boldsymbol{z}_i}}\right\},
\end{aligned}$$

and the likelihood function based on a continuous covariate is

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{\theta}\right) &= \prod_{i=1}^{n}\int_{-\infty}^{\infty} f_{Y|X,Z}\left(\boldsymbol{y_i}|x,z_i;\boldsymbol{\theta}\right) f_{W|X,Z}\left(\boldsymbol{w_i}|x,z_i;\boldsymbol{\theta}_{W|X,Z}\right) f_{X|Z}\left(x|z_i;\boldsymbol{\theta}_{X|Z}\right) dx \\
&= \prod_{i=1}^{n}\int_{-\infty}^{\infty}\left\{\prod_{j=1}^{m_i} P_{y_i(u_{i,j-1}),y_i(u_{ij})}(u_{ij}-u_{i,j-1}|x,z_i;\boldsymbol{\theta})\right. \\
&\qquad \left.\frac{1}{\sqrt{2\pi}\sigma_U}\mathrm{e}^{\frac{-(w_i-x)^2}{2\sigma_U^2}}\frac{1}{\sqrt{2\pi}\sigma_{X|Z}}\mathrm{e}^{\frac{-\left(x-\mu_{X|Z}\right)^2}{2\sigma_{X|Z}^2}}\right\} dx.
\end{aligned}$$

In practice the measurement error parameters and the parameters of the distribution of X given Z must be estimated. This can be achieved with data from a so-called validation or reliability sample. The resulting estimates can then be used in the appropriate likelihood expressions and maximization with respect to the remaining parameters can be carried out. The empirical performance of these approaches will be compared to a naive maximum likelihood approach in the next section.

3.4 Simulation Studies

The objective of these simulations was to compare the performance of the naive and correct estimation approaches in the presence of measurement error and misclassification. Three-state progressive models with time homogeneous transition intensities, $\lambda_1(x,z,\boldsymbol{\alpha})$ and $\lambda_2(x,z,\boldsymbol{\beta})$ were investigated. Values for the baseline transition intensities and the

covariate effects were selected to represent a range of situations that may be encountered in practice. Consider the model given in Figure 3.2 as representing the "true" model. The variable W is the mismeasured version of X and will be used to fit models along with Z, a perfectly measured covariate. As before, parameters associated with $\lambda_1(x, z, \boldsymbol{\alpha})$ are denoted by $\boldsymbol{\alpha} = (\alpha_0, \alpha_X, \alpha_Z)'$, and those associated with the second transition are denoted by $\boldsymbol{\beta} = (\beta_0, \beta_X, \beta_Z)'$. In practice, interest often lies in covariate effects on transitions rather than the baseline transition intensities. Therefore, attention will be primarily directed at the estimators of the regression coefficients for X and Z in this section. All simulations were conducted in SAS using PROC NLP and PROC IML and the plots were generated in R.

3.4.1 Binary Covariates

DATA GENERATION

Data have been generated based on the true models and the joint distribution of (X, W) as follows:

- Number of datasets: $N = 500$,
- Number of subjects per dataset: $n = 500$,
- Years of follow-up: τ was selected such that $P_{1,3}(\tau)$ was at least 0.6 based on all possible values of (X, W),
- Average number of assessments was: $\mu = 5$,
- Baseline intensities: $e^{\alpha_0} = 0.1, 0.2$ and e^{β_0} was set such that $\frac{e^{\beta_0}}{e^{\alpha_0}} = 1.02, 2$ (Note that 1.02 was chosen to set α_0 and β_0 to be close. If 1 had been used, the expression for the transition probabilities in Section 3.3 would have involved division by 0 when X=Z=0.),
- Covariate effects: $e^{\alpha_X} = e^{\beta_X} = 1.25, 2$ and $e^{\alpha_Z} = e^{\beta_Z} = 1.25$, and
- SIMEX parameters: $M = 5$ with $\{\nu_1, \nu_2, \nu_3, \nu_4, \nu_5\} = \{0, 0.5, 1, 1.5, 2\}$ and $B = 100$.

The transition times for each individual were simulated independently as $T_1 \sim EXP(\lambda_1(x, z, \boldsymbol{\alpha}))$ and $T_2 \sim EXP(\lambda_2(x, z, \boldsymbol{\beta}))$, respectively. The time of the first transition was denoted as t_1 and the second was $t_1 + t_2$. The number of follow-up times were generated as $m_i \sim POI(\mu)$. The assessment times, $u_{ij}, j = 1, 2, ..., m_i$ were then generated from m_i independent $UNIF(0, \tau)$ random variables. The i^{th} subject's contribution to the dataset was obtained by recording the state occupied at each of the m_i assessment times.

With binary covariates, misclassification is characterized by misclassification probabilities, $\pi_{01} = 1 - \pi_{11}$ and $\pi_{10} = 1 - \pi_{00}$, or equivalently, by $\pi_{00} = P(W = 0|X = 0)$ and $\pi_{11} = P(W = 1|X = 1)$. Covariate values were generated by the following steps:

- $Z \sim BIN(1, p_Z)$, with $p_Z = 0.5$.
- $X|Z \sim BIN(1, \text{expit}(\xi_0 + \xi_Z Z))$, where $\text{expit}(x) = e^x/(1 + e^x)$ and with $\xi_0 = -\log(3), 0$ and $\xi_Z = -\log(2), \log(2)$, representing negative and positive effects of Z on X.
- $\pi_{11} = P(W = 1|X = 1) = 0.7, 1$ (sensitivity), and
- $\pi_{00} = P(W = 0|X = 0) = 0.7, 0.9, 1$ (specificity). These values were selected to represent minor to moderate misclassification. These configurations also allow us to investigate the situation when only false negatives are possible ($\pi_{11} = 1$ and $\pi_{00} < 1$) or only false positive are possible ($\pi_{00} = 1$ and $\pi_{01} < 1$). As implied by the above expressions, we assume here that the misclassification probabilities do not depend on Z.

Validation samples (one of size 50 and one of size 200) were randomly selected to estimate the misclassification probabilities and for the corrected maximum likelihood approach, the conditional distribution of X given Z. Analyses were based on models consistent in structure to those from which the data were generated (i.e. a proportional transition intensities model was assumed) so there was no misspecification other than the mismeasured covariates to complicate the situation.

ESTIMATION

Estimates of π_{01} and π_{10} were obtained by fitting a logistic regression of W on X in the validation sample:

$$\hat{\pi}_{01} = \frac{1}{1 + \mathrm{e}^{\hat{\phi}_0 + \hat{\phi}_X}} \text{ and } \hat{\pi}_{10} = \frac{\mathrm{e}^{\hat{\phi}_0}}{1 + \mathrm{e}^{\hat{\phi}_0}}.$$

As is customary in measurement error models, the misclassification probabilities were treated as if they were "known" (or at least that there was negligible variation in the corresponding estimators) so $\hat{\pi}_{01}$ and $\hat{\pi}_{10}$ were used to generate the misclassification in the SIMEX approach and in the likelihood function for the maximum likelihood approach. A logistic regression of X on Z provided estimates of ξ_0 and ξ_Z for $P(X = 1|z)$ for use in the likelihood function:

$$P(X = 1|z) = \hat{p}_{X|Z} = \frac{\mathrm{e}^{\hat{\xi}_0 + \hat{\xi}_Z Z}}{1 + \mathrm{e}^{\hat{\xi}_0 + \hat{\xi}_Z Z}}.$$

The SIMEX approach involved repeated simulations and estimation based on the naive likelihood function. The original misclassification was increased by factors given by ν_m, $m = 2, 3, 4, 5$. For each level of induced misclassification, $B = 100$ revised $\boldsymbol{w}_b$'s were generated and each time,

$$\hat{\boldsymbol{\theta}}_b(\nu_m) = \left(\hat{\alpha}_{0b}(\nu_m), \hat{\alpha}_{Xb}(\nu_m), \hat{\alpha}_{Zb}(\nu_m), \hat{\beta}_{0b}(\nu_m), \hat{\beta}_{Xb}(\nu_m), \hat{\beta}_{Zb}(\nu_m)\right)'$$

was obtained by maximizing the following likelihood function:

$$\mathcal{L}_{naive}(\boldsymbol{\theta}(\nu_m)) = \prod_{i=1}^{500} \prod_{j=1}^{m_i} P_{y_i(u_{i,j-1}), y_i(u_{ij})}(u_{ij} - u_{i,j-1}|w_i, z_i; \boldsymbol{\theta}(\nu_m)) \tag{3.5}$$

At each ν_m, $\hat{\boldsymbol{\theta}}(\nu_m)$ was obtained by taking the average of the $B = 100$ naive maximum likelihood estimates. $\hat{\boldsymbol{\theta}}(\nu_1)$ is simply the original naive maximum likelihood estimate. An extrapolation model was then fit to these five values and the SIMEX estimates were obtained by extrapolating back to the case where $\nu = -1$ as described in Section 2.4.1. The simple variance approximation approach described in Stefanski & Cook (1995) for continuous measurement error was used in Küchenhoff et al. (2005) for misclassification and so it was applied here. Variance estimates for the SIMEX estimators were obtained

by first fitting a model to the differences, $\tau^2(\nu_m) - s^2(\nu_m)$, $m = 1, 2, ..., 5$, where $\tau^2(\nu_m)$ is the average of the $B = 100$ model-based variance estimates at each ν_m (based on the inverse of the naive information matrix here) and $s^2(\nu_m)$ is the sample variance of the $B = 100$ parameter estimates at ν_m. The SIMEX variance estimate was then obtained by extrapolating this relationship back to $\nu = -1$. Quadratic ($\theta(\nu) = a + b\nu + c\nu^2$) and exponential ($\theta(\nu) = ae^{b\nu}$) extrapolation functions were considered and fit using least squares in SAS (using PROC REG and PROC NLIN, respectively), as in Küchenhoff et al. (2005).

For simplicity, the same extrapolation function was used to obtain the SIMEX parameter and variance estimates. In practice, extrapolation function selection would not necessarily be automated in this way and model building techniques could be used along with diagnostic checks based on residuals to assess the adequacy of the models. Moreover the parameter and variance estimates need not have the same extrapolant. It is difficult to imitate this in simulation studies, but two alternative approaches were considered in the simulation based on quadratic and exponential functions. First, the optimal model of the two based on adjusted R^2 (i.e. $R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$) was selected to estimate the parameters and the variances. Second, since both extrapolation functions appeared to perform reasonably well, the average of the estimates arising from the two functions were considered.

The maximum likelihood approach accommodating misclassification was based on the following likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{x=0}^{1} \prod_{j=1}^{m_i} P_{y_i(u_{i,j-1}),y_i(u_{ij})}(u_{ij} - u_{i,j-1}|x, z_i; \boldsymbol{\theta}) \left[\frac{\left(e^{\hat{\phi}_0 + \hat{\phi}_X x}\right)^{w_i}}{1 + e^{\hat{\phi}_0 + \hat{\phi}_X x}} \right] \left[\frac{\left(e^{\hat{\xi}_0 + \hat{\xi}_Z z_i}\right)^{x}}{1 + e^{\hat{\xi}_0 + \hat{\xi}_Z z_i}} \right]. \quad (3.6)$$

This function was maximized with respect to $\boldsymbol{\theta} = (\alpha_0, \alpha_X, \alpha_Z, \beta_0, \beta_X, \beta_Z)'$. For both SIMEX and the maximum likelihood approaches, as in the simulations for the two-state problem, the objective functions were maximized based on a quasi-Newton algorithm using PROC NLP in SAS which was described in Chapter 2. For each set of parameter estimates, approximate 95% confidence intervals using the model-based standard errors were constructed and compared to the true parameter values. Empirical coverage probabilities

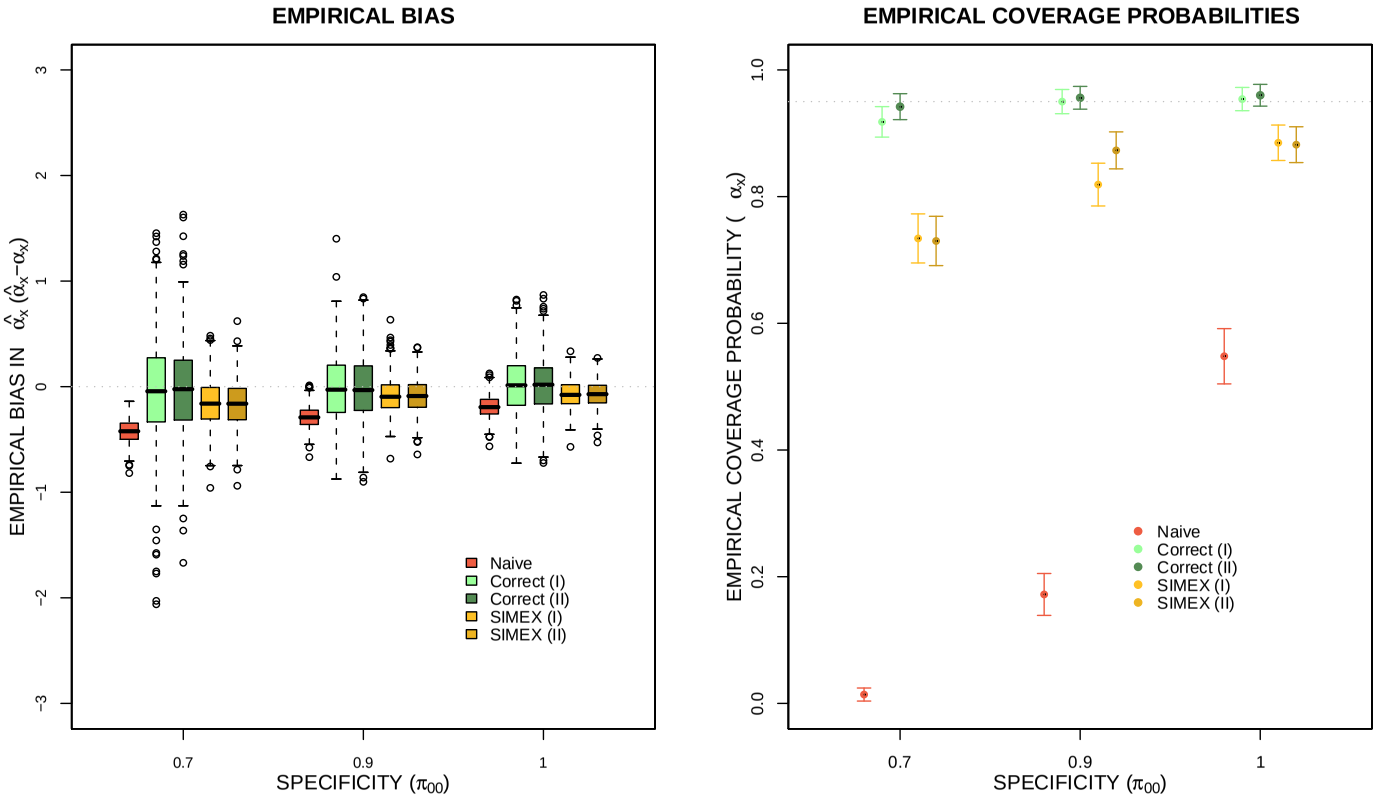
(ECPs) were then calculated as the proportion of the 500 95% confidence intervals containing the true value of the parameter of interest. Error bars were included in the plots by constructing approximate 95% confidence intervals for these proportions based on the observed ECPs. However, if the true confidence level associated with these intervals is 0.95, then we would expect the empirical coverage probabilities to be close to 0.95. Further we would expect the ECPs to fall between $0.95 \pm 1.96\sqrt{\frac{(0.95)(0.05)}{500}}$ or 0.9309 and 0.9691 approximately 95% of the time. A visual comparison can be made between the ECP intervals and the nominal coverage probability of 0.95 in the plots. Representative results from this simulation study are displayed in Figures 3.15 to 3.22.

DISCUSSION

As expected, the biases from the naive analyses were larger in magnitude for the estimators associated with the covariate subject to misclassification, X. In addition, the ECPs corresponding to these estimators were generally farther from the nominal value of 0.95 than those corresponding to Z, the error-free covariate. Both the correct likelihood approach and the SIMEX approach exhibit smaller biases and ECPs which are closer to the nominal level. Also, the biases tended to be smaller and the ECPs were better for estimation based on a large validation sample (size 200). This seems reasonable because the more validation data we have, the more information is available about the misclassification matrix and for the distribution of X given Z. The SIMEX approach, which was presented based on a quadratic extrapolation function for both parameter and variance estimates, is only a moderate improvement over the naive maximum likelihood approach. There is still some bias present and quite a few of the 95% confidence intervals for the true coverage probability lie below 0.95. Based on the parameter configurations explored in this study, the exponential extrapolation function performed better for estimation of the parameters associated with X (i.e. α_X and β_X), but not as well for the other parameters; the quadratic results are therefore displayed.

Consistent with the asymptotic bias plots presented in Figures 3.3 to 3.6, the impact of misclassification does not appear to be symmetric. The situation with only false negatives did not appear to induce the same magnitude of bias and decrease in ECP for the naive

Figure 3.15: *Empirical performance of estimators for the regression parameter associated with a misclassified binary covariate on the first transition; Number of assessments are POI* (5)*;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$*;* $P(Z = 1) = 0.5$ *and logit*$[P(X = 1|Z = z)] = \log(2)z$*;* $\pi_{11} = P(W = 1|X = 1) = 0.7$ *(empirical coverage probabilities are shown as* $\widehat{ECP} \pm 1.96\sqrt{\widehat{ECP}(1 - \widehat{ECP})/500}$*).*



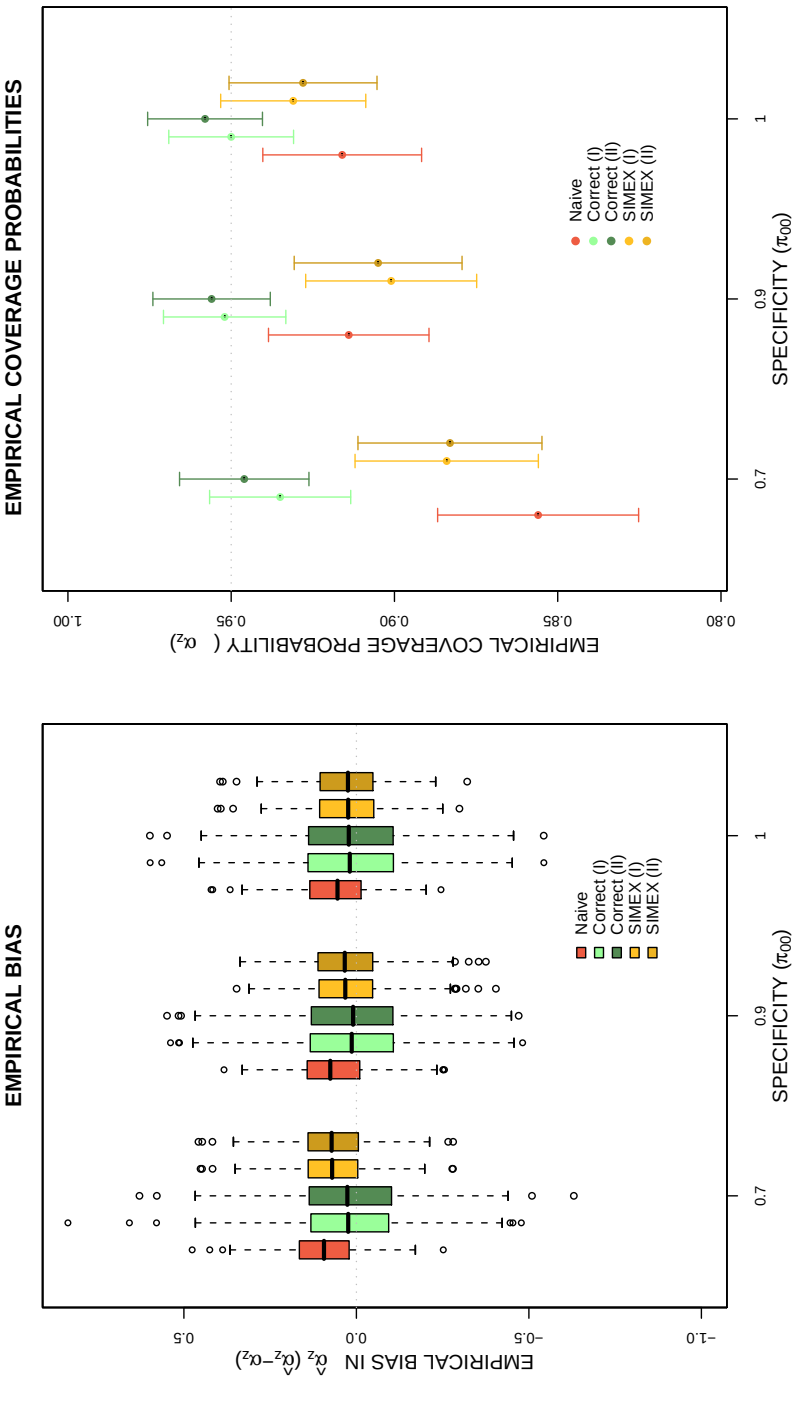
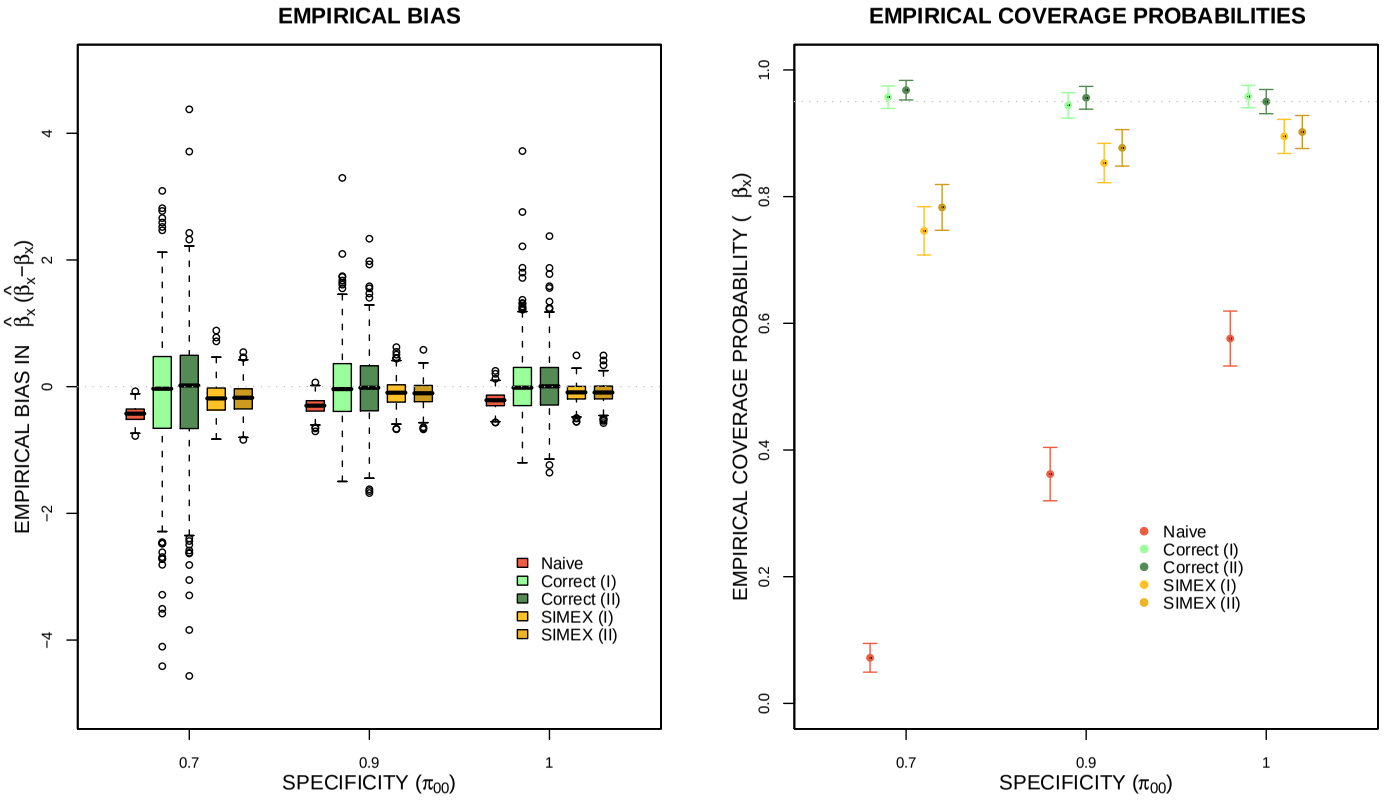


Figure 3.16: *Empirical performance of estimators for the regression parameter associated with a correctly classified binary covariate on the first transition; Number of assessments are* $POI(5)$*;* $\alpha_0 = \log(0.2)$*,* $\beta_0 = \log(0.4)$*,* $\alpha_X = \beta_X = \log(2)$*,* $\alpha_Z = \beta_Z = \log(1.25)$*;* $P(Z = 1) = 0.5$ *and* $logit[P(X = 1|Z = z)] = \log(2)z$*;* $\pi_{11} = P(W = 1|X = 1) = 0.7$ *(empirical coverage probabilities are shown as* $\widehat{ECP} \pm 1.96\sqrt{\widehat{ECP}(1 - \widehat{ECP})/500}$*).*

Figure 3.17: *Empirical performance of estimators for the regression parameter associated with a misclassified binary covariate on the second transition; Number of assessments are POI* (5)*;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$; $P(Z = 1) = 0.5$ *and* $logit[P(X = 1|Z = z)] = \log(2)z$; $\pi_{11} = P(W = 1|X = 1) = 0.7$ *(empirical coverage probabilities are shown as* $\widehat{ECP} \pm 1.96\sqrt{\widehat{ECP}(1 - \widehat{ECP})/500}$*).*



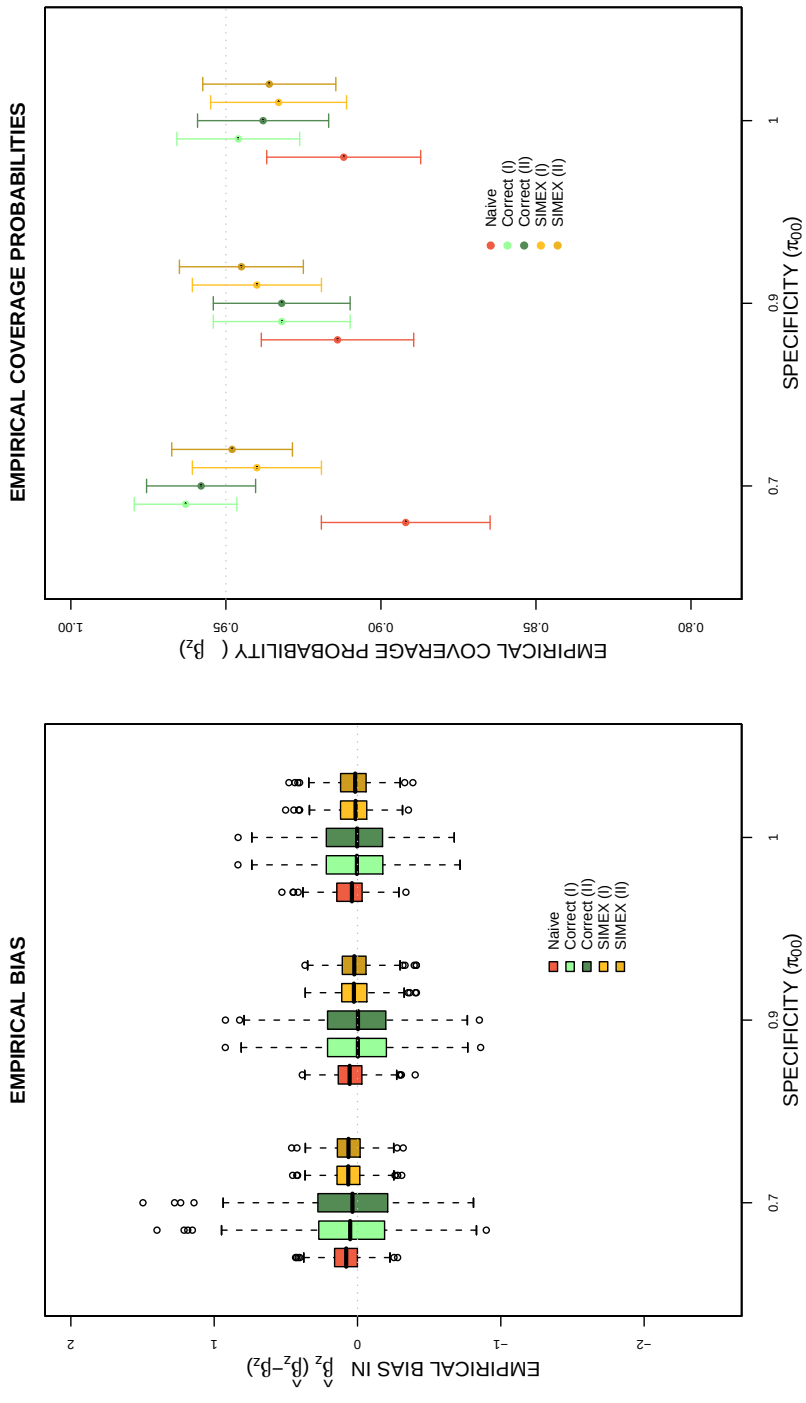
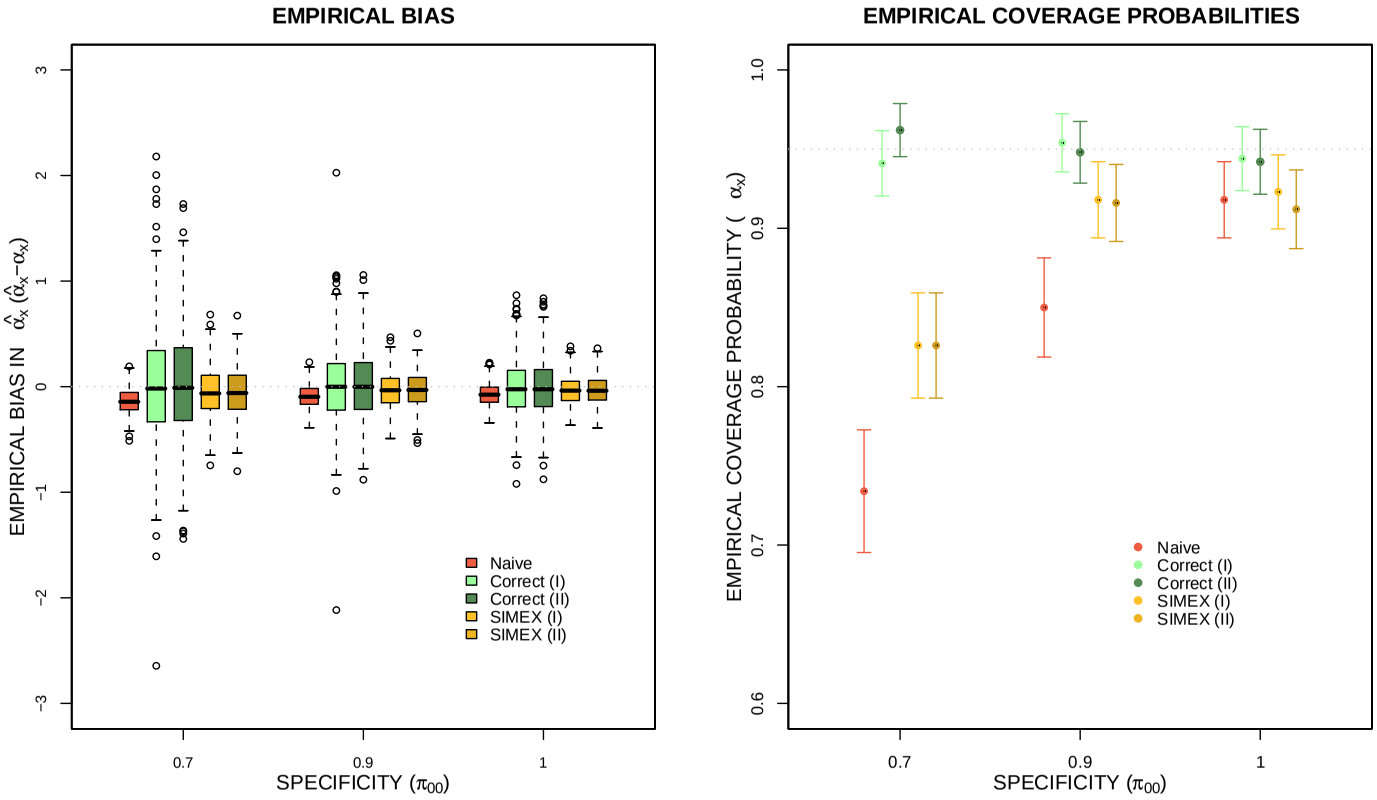


Figure 3.18: *Empirical performance of estimators for the regression parameter associated with a correctly classified binary covariate on the second transition; Number of assessments are POI* (5); $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$; $P(Z = 1) = 0.5$ *and* $logit[P(X = 1|Z = z)] = \log(2)z$; $\pi_{11} = P(W = 1|X = 1) = 0.7$ *(empirical coverage probabilities are shown as* $\widehat{ECP} \pm 1.96\sqrt{\widehat{ECP}(1 - \widehat{ECP})/500}$*).*

Figure 3.19: *Empirical performance of estimators for the regression parameter associated with a misclassified binary covariate on the first transition; Number of assessments are* $POI\,(5)$*;* $\alpha_0 = \log(0.2)$*,* $\beta_0 = \log(0.4)$*,* $\boldsymbol{\alpha_X = \beta_X = \alpha_Z = \beta_Z = \log(1.25)}$*;* $P(Z = 1) = 0.5$ *and* $logit[P(X = 1|Z = z)] = \log(2)z$*;* $\pi_{11} = P(W = 1|X = 1) = 0.7$ *(empirical coverage probabilities are shown as* $\widehat{ECP} \pm 1.96\sqrt{\widehat{ECP}(1 - \widehat{ECP})/500}$*).*



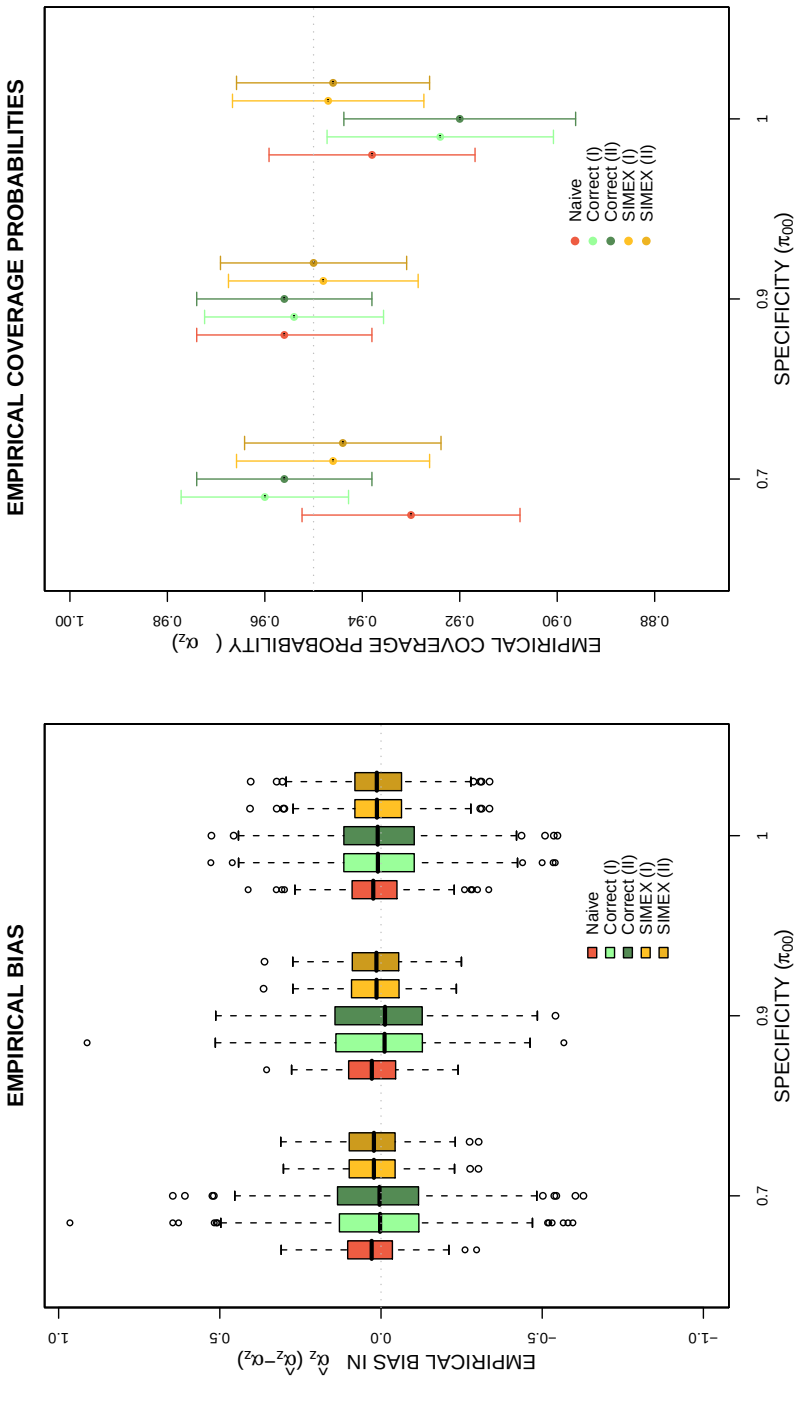
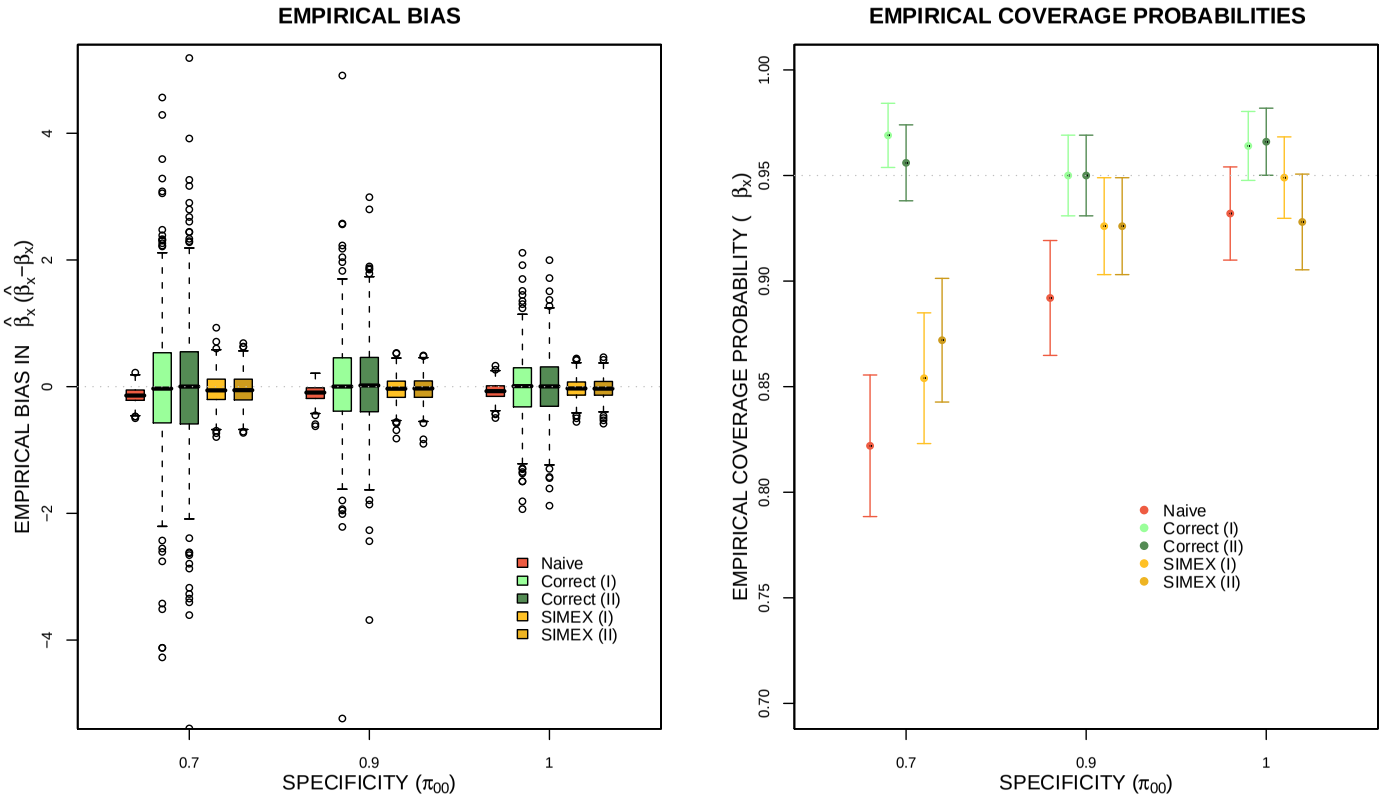


Figure 3.20: *Empirical performance of estimators for the regression parameter associated with a correctly classified binary covariate on the first transition; Number of assessments are* $POI(5)$; $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\boldsymbol{\alpha_X} = \boldsymbol{\beta_X} = \boldsymbol{\alpha_Z} = \boldsymbol{\beta_Z} = \mathbf{log\,(1.25)}$;$P(Z = 1) = 0.5$ *and* $logit[P(X = 1|Z = z)] = \log(2)z$; $\pi_{11} = P(W = 1|X = 1) = 0.7$ *(empirical coverage probabilities are shown as* $\widehat{ECP} \pm 1.96\sqrt{\widehat{ECP}(1 - \widehat{ECP})/500}$*).*

Figure 3.21: *Empirical performance of estimators for the regression parameter associated with a misclassified binary covariate on the second transition; Number of assessments are POI* (5)*;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\boldsymbol{\alpha_X = \beta_X = \alpha_Z = \beta_Z = \log(1.25)}$*;* $P(Z = 1) = 0.5$ *and* $logit[P(X = 1|Z = z)] = \log(2)z$*;* $\pi_{11} = P(W = 1|X = 1) = 0.7$ *(empirical coverage probabilities are shown as* $\widehat{ECP} \pm 1.96\sqrt{\widehat{ECP}(1 - \widehat{ECP})/500}$*).*



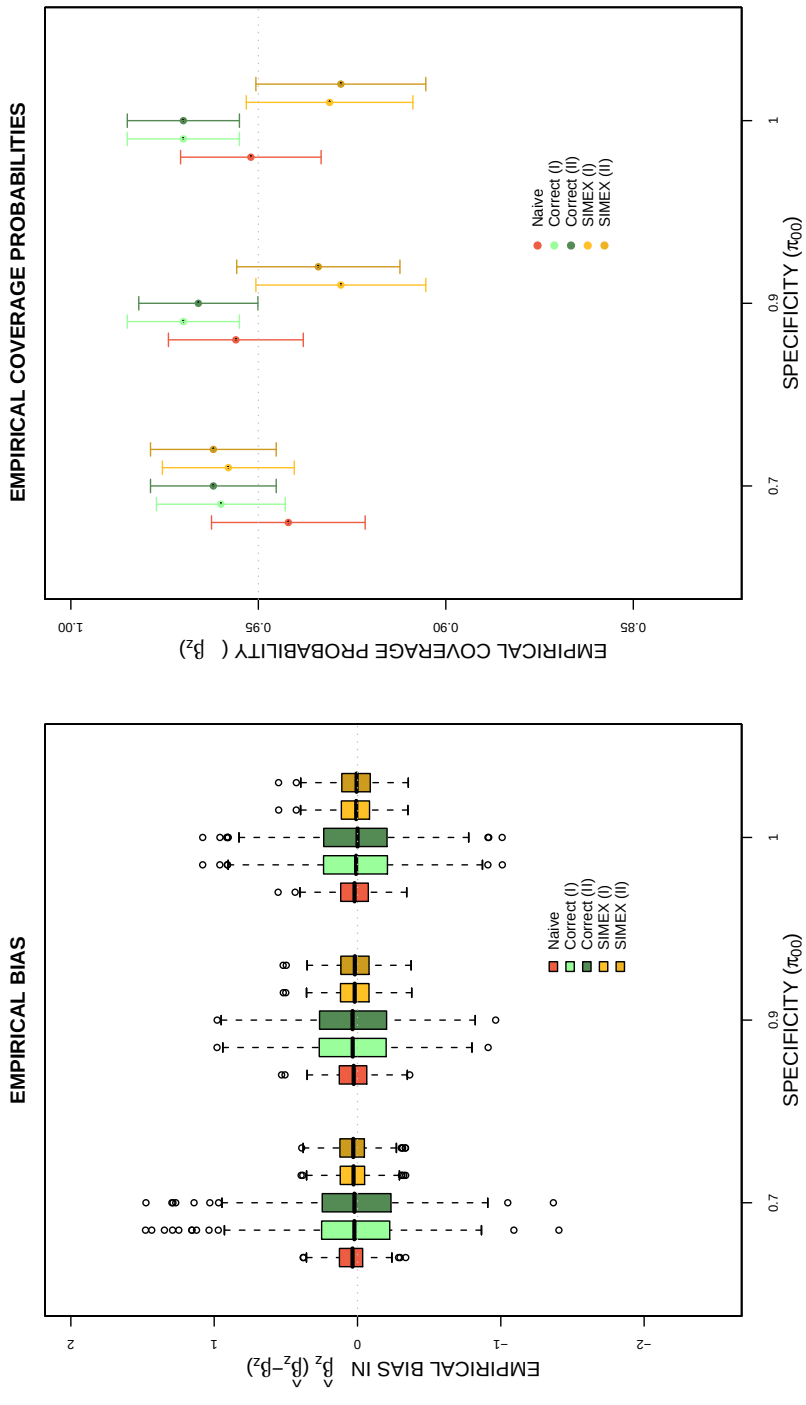


Figure 3.22: *Empirical performance of estimators for the regression parameter associated with a correctly classified binary covariate on the second transition; Number of assessments are* $POI(5)$*;* $\alpha_0 = \log(0.2)$*,* $\beta_0 = \log(0.4)$*,* $\boldsymbol{\alpha_X} = \boldsymbol{\beta_X} = \boldsymbol{\alpha_Z} = \boldsymbol{\beta_Z} = \mathbf{log\,(1.25)}$*;* $P(Z=1) = 0.5$ *and* $logit[P(X=1|Z=z)] = \log(2)z$*;* $\pi_{11} = P(W=1|X=1) = 0.7$ *(empirical coverage probabilities are shown as* $\widehat{ECP} \pm 1.96\sqrt{\widehat{ECP}(1-\widehat{ECP})/500}$*).*

Table 3.1: *Comparison of naive maximum likelihood results for* $\pi_{00} = 0.7$, $\pi_{11} = 1$ *and* $\pi_{00} = 1$, $\pi_{11} = 0.7$.

		$\pi_{00} = 0.7$, $\pi_{11} = 1$				$\pi_{00} = 1$, $\pi_{11} = 0.7$			
		X		Z		X		Z	
Configuration	Transition	Bias	ECP	Bias	ECP	Bias	ECP	Bias	ECP
1	$1 \rightarrow 2$	-0.1457	0.750	0.0500	0.906	-0.1866	0.548	0.0536	0.916
	$2 \rightarrow 3$	-0.1199	0.858	0.0460	0.892	-0.2243	0.576	0.0480	0.912
2	$1 \rightarrow 2$	-0.0405	0.938	0.0160	0.938	-0.0642	0.918	0.0166	0.938
	$2 \rightarrow 3$	-0.0483	0.942	0.0230	0.958	-0.0644	0.932	0.0208	0.952
3	$1 \rightarrow 2$	-0.0477	0.936	0.0149	0.946	-0.1350	0.750	0.0334	0.934
	$2 \rightarrow 3$	-0.0390	0.946	0.0290	0.940	-0.1375	0.818	0.0291	0.952

1 $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$
2 $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_X = \beta_X = \alpha_Z = \beta_Z = \log(1.25)$
3 $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.22)$, $\alpha_X = \beta_X = \alpha_Z = \beta_Z = \log(1.25)$

maximum likelihood estimation approach as the situation with only false positives. The setting with $\pi_{00} = 0.7$, $\pi_{11} = 1$ (false positives only) was consistently observed to result in lower estimated biases and ECPs closer to 0.95 than when $\pi_{00} = 1$, $\pi_{11} = 0.7$ (false negatives only); this was particularly true for the estimates of α_X and β_X. Table 3.1 illustrates this for several parameter configurations; two for which the full simulation results are given in Figures 3.15 to 3.22. Although not summarized in this table, the naive estimated standard errors associated with α_0, α_X, β_0 and β_X also tended to be slightly smaller when $\pi_{00} = 1$, $\pi_{11} = 0.7$ as compared to when $\pi_{00} = 0.7$, $\pi_{11} = 1$. The same pattern seemed to be apparent in the two-state simulations (Chapter 2) although the difference was not nearly as dramatic.

For the parameter configurations investigated in this simulation study, $SE_{naive} < SE_{SIMEX} < SE_{correct}$. This was also somewhat apparent in the two-state simulations in Chapter 2 but the difference did not appear to be as great, and whether the SIMEX or the correct standard errors were larger very much depended on the form of the assumed extrapolation function. The difference observed in the three-state model-based standard error estimates may have been partially due to the way the programming was carried for these simulation studies. For the maximum likelihood approach, although internal validation data were generated, the analysis was performed as if the estimates for the

misclassification probabilities and the $X|Z$ distribution were obtained from an external data source. When internal validation data are available, the following likelihood function should be maximized with respect to all parameters; $(\alpha_0, \alpha_X, \alpha_Z, \beta_0, \beta_X, \beta_Z, \phi_0, \phi_X, \xi_0, \xi_Z)$. Let $\Delta_i = 1$ when subject i is in the validation study.

$$\mathcal{L}_i(\boldsymbol{\theta}) = \begin{cases} \sum_{x=0}^{1} P_{y_i(u_{i,j-1}),y_i(u_{ij})}\left(u_{ij} - u_{i,j-1}|x, z_i; \boldsymbol{\theta}\right) \left[\frac{\left(\mathrm{e}^{\phi_0+\phi_X x}\right)^{w_i}}{1+\mathrm{e}^{\phi_0+\phi_X x}}\right] \left[\frac{\left(\mathrm{e}^{\xi_0+\xi_Z z_i}\right)^{x}}{1+\mathrm{e}^{\xi_0+\xi_Z z}}\right], & \Delta_i = 0 \\ P_{y_i(u_{i,j-1}),y_i(u_{ij})}\left(u_{ij} - u_{i,j-1}|x_i, z_i; \boldsymbol{\theta}\right) \left[\frac{\left(\mathrm{e}^{\phi_0+\phi_X x_i}\right)^{w_i}}{1+\mathrm{e}^{\phi_0+\phi_X x_i}}\right], & \Delta_i = 1 \end{cases} \tag{3.7}$$

To confirm that similar results would be observed if external validation data were available and to compare results based on (3.6) to (3.7) based on an internal validation sample, a small numerical study was conducted. Table 3.2 summarizes results based on the parameter configuration in Figures 3.15 to 3.18 with $\pi_{00} = \pi_{11} = 0.7$. "Correct 1" represents the correct likelihood approach based on treating an internal validation study as external, "Correct 2" represents correct maximum likelihood using an external validation study and "Correct 3", the correct likelihood based on an internal validation sample. The likelihood function (3.6) was maximized for "Correct 1" and "Correct 2"; whereas (3.7) was maximized to obtain results for "Correct 3". The reported bias is the difference between the average of the estimates from 500 samples and the true value and the SE is the average of the 500 model-based standard errors for each of the six parameters.

The results based on (3.6) appear to be pretty much consistent regardless of whether external or internal validation data are used to estimate the parameters associated with the error and covariate distributions. This is probably due to the fact that in both cases point estimates for these parameters are substituted into (3.6) which is then maximized with respect to $\boldsymbol{\theta}$. Although the standard errors based on the likelihood function in (3.7) appear to be smaller than those based on (3.6), the likelihood function used in the simulations, and this difference seems to be largest for the estimators of the effects associated with the misclassified covariate, they are still larger than the SIMEX standard errors. However, the SIMEX variance estimates were based on an approximate method that assumes the misclassification probabilities and extrapolants are known. Since the variability in the estimated misclassification rates was not taken into account, it may be the case that another variance estimation procedure, such as bootstrap variance estimation, would result

Table 3.2: *Comparison of correct maximum likelihood results based on external versus internal validation samples of size 200.*

	Naive			**SIMEX (Quadratic)**			**Correct 1**			**Correct 2**			**Correct 3**		
Parameter	Bias	SE	ECP	Bias	SE	ECP	Bias	SE	ECP	Bias	SE	ECP	Bias	SE	ECP
α_0	0.1958	0.0901	0.414	0.0730	0.1112	0.845	-0.0117	0.2875	0.956	-0.0099	0.2835	0.952	-0.0078	0.2104	0.952
α_X	-0.4280	0.1021	0.010	-0.1670	0.1618	0.730	-0.0206	0.4503	0.942	0.0002	0.4461	0.936	0.0231	0.2728	0.950
α_Z	0.0904	0.1017	0.848	0.0689	0.1027	0.883	0.0156	0.1846	0.946	0.0021	0.1849	0.952	-0.0138	0.1842	0.946
β_0	0.2284	0.1247	0.450	0.1077	0.1377	0.859	0.0313	0.7060	0.979	-0.0274	0.7498	0.980	-0.0511	0.4086	0.958
β_X	-0.4252	0.1247	0.074	-0.1838	0.1979	0.783	-0.0710	0.9324	0.968	-0.0084	0.9745	0.956	0.0247	0.5064	0.948
β_Z	0.0802	0.1242	0.898	0.0617	0.1259	0.968	0.0144	0.3245	0.958	0.0239	0.3254	0.932	0.0246	0.3094	0.950

in larger standard error estimates; this may also improve the SIMEX empirical coverage probabilities.

3.4.2 Continuous Covariates

DATA GENERATION

Data were generated as in Section 3.4.1. We assumed $Z \sim N(\mu_Z, \sigma_Z^2)$, where without loss of generality we set $\mu_Z = 0$ and considered $\sigma_Z^2 = 0.1$ and 1 to represent low and high variability in Z. We let $X|Z \sim N\left(\mu_{X|Z}, \sigma_{X|Z}^2\right)$, where $\mu_{X|Z} = \xi_0 + \xi_Z Z$, with $\xi_0 = 0$ and $\xi_Z = 0$ and 1.33 to represent a couple of plausible relationships between X and Z (i.e. when $\xi_Z = 0$, X and Z are independent and when $\xi_Z = 1.33$, $\rho_{XZ} = CORR(X, Z) = \xi_Z \sigma_Z / \sqrt{\sigma_{X|Z}^2 + \xi_Z^2 \sigma_Z^2} = 0.8$). The parameter $\sigma_{X|Z}^2$ was set to 0.1 and 1 to represent low and high variability in X given Z. Note that we are making the simplifying assumption that the distribution of X only depends on Z through its mean. We are considering the situation where $\sigma_{X|Z}^2$ does not depend on Z. The classical error model given by $\mu_{W|X,Z} = \zeta_0 + \zeta_X X$ was considered where $\zeta_0 = 0$ and $\zeta_X = 1$, and $\sigma_{W|X,Z}^2 = \sigma_U^2$ was selected to result in values of 0.5 and 0.8 for the *reliability ratio*, which is defined as $\gamma = \sigma_{X|Z}^2 / \left(\sigma_{X|Z}^2 + \sigma_u^2\right)$. These values of γ were selected to represent low to moderate reliability of W as a measure for X. Values for σ_U^2 are summarized in the following table based on the selected simulation values for $\sigma_{X|Z}^2$ and γ:

	σ_U^2	
	$\sigma_{X\|Z}^2 = 0.1$	$\sigma_{X\|Z}^2 = 1$
$\gamma = 0.5$	0.1	1
$\gamma = 0.8$	0.025	0.25

Values of the covariates were generated from a trivariate normal distribution given by (2.24). Again, two validation samples were randomly selected from the 500 subjects in each dataset to estimate the measurement error and conditional covariate distributions.

ESTIMATION

Based on the validation data, the measurement error distribution was modeled as $W = \zeta_0 + \zeta_X X + \zeta_Z Z + \epsilon$ and estimates for ζ_0, ζ_X and ζ_Z were obtained using least squares. The model $X = \xi_0 + \xi_Z Z + \epsilon$ was also fit using least squares to obtain $\hat{\xi_0}$ and $\hat{\xi_Z}$ to substitute into the likelihood function for the correct maximum likelihood approach. The models fit to the data had the same structure as the models used to generate the data so there was

no model misspecification other than the mismeasurement in X.

The SIMEX approach was implemented as in Section 3.4.1 and involved repeated maximization of the likelihood function in (3.5). PROC NLP in SAS was used here to obtain the maximum likelihood estimates. The correct maximum likelihood approach was based on the following likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{j=1}^{m_i} P_{y_i(u_{i,j-1}),y_i(u_{ij})}\left(u_{ij} - u_{i,j-1}|x, z_i; \boldsymbol{\theta}\right) f_{W|X}\left(w_i|x\right) f_{X|Z}\left(x|z_i\right) dx, \tag{3.8}$$

where $f_{W|X}\left(w|x\right) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_U} \mathrm{e}^{\frac{-(w-x)^2}{2\hat{\sigma}_U^2}}$ and $f_{X|Z}\left(x|z\right) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_{X|Z}} \mathrm{e}^{\frac{-\left(x-\hat{\mu}_{X|Z}\right)^2}{2\hat{\sigma}_{X|Z}^2}}$. This function was maximized with respect to $\boldsymbol{\theta} = (\alpha_0, \alpha_X, \alpha_Z, \beta_0, \beta_X, \beta_Z)'$. Gaussian quadrature was used to numerically approximate the integrals; an abscissas and weights based on 20 points were used. For both SIMEX and the maximum likelihood approaches, the objective functions were maximized based on a quasi-Newton algorithm using PROC NLP in SAS. Representative results from this simulation study are displayed in Tables 3.3 to 3.8.

In these tables, the term "known" indicates that the measurement error and the distribution of X given Z were known exactly (i.e. the true values are used). Essentially, these represent the best case scenarios in terms of bias and estimated standard errors based on the correct maximum likelihood function. "Sample I" refers to the results based on the small validation sample and "Sample II" refers to the results with a large validation sample. Linear ($\theta(\nu) = a + b\nu$), quadratic ($\theta(\nu) = a + b\nu + c\nu^2$), cubic ($\theta(\nu) = a + b\nu + c\nu^2 + d\nu^3$), exponential ($\theta(\nu) = a\mathrm{e}^{b\nu}$) and rational linear or nonlinear functions ($\theta(\nu) = a + \frac{b}{c+\nu}$) were fit to obtain the SIMEX parameter and variance estimates. The nonlinear extrapolation function looked to provide the best results in terms of estimated bias and empirical coverage probabilities for all parameters for the cases which convergence was reached. There were often convergence problems or negative extrapolated variance estimates. Therefore, the quadratic and cubic extrapolation function results are presented. There did not appear to be much of a difference in the SIMEX results based on measurement error variance estimates from a small versus large validation study, so the results are summarized based on a small validation study. The small validation study was chosen since it was thought it

Table 3.3: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are $POI\,(5)$; $\alpha_0 = \log(0.1)$, $\beta_0 = \log(0.2)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$; $Z \sim N(0,1)$ and $X|Z \sim N(1.33Z, 1)$ such that $\rho_{XZ} = 0.8$.*

		$\boldsymbol{\alpha_X}$		$\boldsymbol{\alpha_Z}$		$\boldsymbol{\beta_X}$		$\boldsymbol{\beta_Z}$	
Method		$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$
Naive	Bias	-0.3690	-0.1550	0.4174	0.1798	-0.3880	-0.1600	0.3910	0.1676
	SE_1	0.0439	0.0578	0.0856	0.0952	0.0607	0.0829	0.1187	0.1291
	SE_2	0.0453	0.0620	0.0944	0.1040	0.0625	0.0826	0.1304	0.1361
	ECP	0.0000	0.2325	0.0000	0.5190	0.0000	0.4709	0.0820	0.7415
Likelihood									
Known	Bias	0.0081	-0.0006	-0.0015	0.0033	-0.0043	0.0091	0.0153	0.0033
	SE_1	0.0997	0.0792	0.1309	0.1512	0.1411	0.1158	0.1712	0.1512
	SE_2	0.0965	0.0824	0.1303	0.1503	0.1367	0.1115	0.1709	0.1503
	ECP	0.9619	0.9460	0.9479	0.9540	0.9399	0.9620	0.9399	0.9540
Sample I	Bias	0.0085	0.0105	-0.0025	-0.0021	-0.0057	0.0184	0.0147	-0.0040
	SE_1	0.0997	0.0804	0.1314	0.1138	0.1410	0.1174	0.1719	0.1524
	SE_2	0.1262	0.1057	0.1751	0.1565	0.1590	0.1217	0.2077	0.1668
	ECP	0.8896	0.8840	0.8795	0.8700	0.8956	0.9440	0.8876	0.9340
Sample II	Bias	0.0090	0.0020	-0.0019	0.0067	-0.0033	0.0113	0.0153	0.0022
	SE_1	0.0999	0.0795	0.1311	0.1130	0.1413	0.1162	0.1715	0.1514
	SE_2	0.1025	0.0866	0.1363	0.1237	0.1420	0.1127	0.1738	0.1524
	ECP	0.9559	0.9200	0.9339	0.9140	0.9339	0.9620	0.9379	0.9500
SIMEX									
Quadratic	Bias	-0.2038	-0.0251	0.2318	0.0350	-0.2246	-0.0221	0.2298	0.0332
	SE_1	0.0621	0.0721	0.0985	0.1063	0.0869	0.1051	0.1326	0.1427
	SE_2	0.0749	0.0854	0.1161	0.1233	0.1039	0.1108	0.1537	0.1526
	ECP	0.1303	0.8968	0.3627	0.8947	0.3046	0.9332	0.5611	0.9190
Cubic	Bias	-0.1353	-0.0062	0.1554	0.0153	-0.1590	0.0029	0.1651	0.0093
	SE_1	0.0719	0.0759	0.1059	0.1092	0.1017	0.1107	0.1420	0.1468
	SE_2	0.0993	0.0945	0.1374	0.1298	0.1350	0.1268	0.1769	0.1662
	ECP	0.4790	0.8765	0.6333	0.8968	0.5731	0.9211	0.7395	0.9231

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 3.4: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are $POI\,(5)$; $\alpha_0 = \log(0.1)$, $\beta_0 = \log(0.2)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$; $Z \sim N(0,1)$ and $\boldsymbol{X|Z \sim N(0,1)}$ such that $\boldsymbol{\rho_{XZ} = 0}$.*

		$\boldsymbol{\alpha_X}$		$\boldsymbol{\alpha_Z}$		$\boldsymbol{\beta_X}$		$\boldsymbol{\beta_Z}$	
Method		$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$
Naive	Bias	-0.3726	-0.1555	-0.0141	-0.0035	-0.3921	-0.1671	-0.0291	-0.0141
	SE_1	0.0415	0.0544	0.0567	0.0571	0.0556	0.0757	0.0735	0.0748
	SE_2	0.0443	0.0567	0.0581	0.0571	0.0581	0.0777	0.0758	0.0742
	ECP	0.0000	0.1920	0.9299	0.9540	0.0000	0.3820	0.9359	0.9520
Likelihood									
Known	Bias	-0.0003	0.0025	0.0036	0.0040	0.0101	0.0082	0.0045	0.0005
	SE_1	0.0965	0.0752	0.0637	0.0602	0.1380	0.1082	0.0845	0.0797
	SE_2	0.0971	0.0769	0.0632	0.0589	0.1390	0.1071	0.0837	0.0775
	ECP	0.9460	0.9440	0.9600	0.9560	0.9640	0.9500	0.9620	0.9680
Sample I	Bias	0.0028	0.0072	0.0071	0.0045	0.0127	0.0128	0.0072	0.0012
	SE_1	0.0968	0.0758	0.0647	0.0604	0.1385	0.1090	0.0858	0.0801
	SE_2	0.1256	0.0923	0.0978	0.0717	0.1677	0.1192	0.1181	0.0870
	ECP	0.8540	0.8916	0.8260	0.8957	0.9100	0.9378	0.8700	0.9337
Sample II	Bias	0.0027	0.0025	0.0023	0.0060	0.0130	0.0082	0.0030	0.0030
	SE_1	0.0969	0.0753	0.0639	0.0603	0.1385	0.1082	0.0846	0.0799
	SE_2	0.1023	0.0814	0.0707	0.0616	0.1464	0.1086	0.0882	0.0813
	ECP	0.9400	0.9319	0.9320	0.9459	0.9400	0.9499	0.9420	0.9559
SIMEX									
Quadratic	Bias	-0.2079	-0.0272	-0.0069	0.0030	-0.2277	-0.0272	-0.0167	-0.0025
	SE_1	0.0585	0.0675	0.0581	0.0960	0.0795	0.0960	0.0759	0.0772
	SE_2	0.0725	0.0770	0.0628	0.1072	0.0982	0.1072	0.0805	0.0777
	ECP	0.1167	0.8831	0.9336	0.8891	0.2455	0.8891	0.9316	0.9617
Cubic	Bias	-0.1410	-0.0086	-0.0038	0.0048	-0.1567	-0.0024	-0.0108	-0.0001
	SE_1	0.0677	0.0710	0.0591	0.0589	0.0929	0.1018	0.0776	0.0780
	SE_2	0.0959	0.0866	0.0655	0.0609	0.1268	0.1210	0.0859	0.0790
	ECP	0.4507	0.8851	0.9256	0.9516	0.5352	0.8931	0.9235	0.9536

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 3.5: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are POI (5);* $\boldsymbol{\alpha_0 = \log(0.2)}$, $\boldsymbol{\beta_0 = \log(0.4)}$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$*;* $Z \sim N(0,1)$ *and* $X|Z \sim N(1.33Z, 1)$ *such that* $\rho_{XZ} = 0.8$.

		$\boldsymbol{\alpha_X}$		$\boldsymbol{\alpha_Z}$		$\boldsymbol{\beta_X}$		$\boldsymbol{\beta_Z}$	
Method		$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$
Naive	Bias	-0.3685	-0.1499	0.4152	0.1720	-0.3851	-0.1580	0.3963	0.1588
	SE_1	0.0440	0.0577	0.0852	0.0946	0.0607	0.0823	0.1187	0.1286
	SE_2	0.0455	0.0579	0.0932	0.0959	0.0636	0.0842	0.1268	0.1319
	ECP	0.0000	0.2540	0.0000	0.5460	0.0000	0.4800	0.0800	0.7740
Likelihood									
Known	Bias	0.0024	0.0053	0.0026	-0.0003	-0.0031	0.0095	0.0241	-0.0065
	SE_1	0.0989	0.0790	0.1299	0.1123	0.1408	0.1149	0.1703	0.1507
	SE_2	0.0952	0.0787	0.1302	0.1110	0.1416	0.1135	0.1665	0.1473
	ECP	0.9600	0.9438	0.9520	0.9478	0.9620	0.9478	0.9600	0.9538
Sample I	Bias	0.0173	0.0163	-0.0219	-0.0156	0.0101	0.0207	0.0026	-0.0204
	SE_1	0.1011	0.0804	0.1332	0.1138	0.1438	0.1170	0.1746	0.1529
	SE_2	0.1223	0.0936	0.1742	0.1370	0.1583	0.1234	0.1968	0.1672
	ECP	0.8898	0.9160	0.8637	0.9120	0.9399	0.9440	0.9259	0.9340
Sample II	Bias	0.0111	0.0066	-0.0105	-0.0028	0.0066	0.0108	0.0108	-0.0091
	SE_1	0.1002	0.0791	0.1315	0.1125	0.1427	0.1151	0.1724	0.1511
	SE_2	0.1011	0.0819	0.1379	0.1139	0.1486	0.1185	0.1714	0.1538
	ECP	0.9439	0.9337	0.9419	0.9538	0.9499	0.9378	0.9619	0.9478
SIMEX									
Quadratic	Bias	-0.1996	-0.0195	0.2269	0.0273	-0.2185	-0.0201	0.2323	0.0240
	SE_1	0.0623	0.0719	0.0980	0.1053	0.0870	0.1042	0.1325	0.1420
	SE_2	0.0778	0.0830	0.1180	0.1182	0.1051	0.1133	0.1504	0.1523
	ECP	0.1747	0.9056	0.3594	0.9217	0.3133	0.9116	0.5562	0.9317
Cubic	Bias	-0.1310	0.0008	0.1519	0.0042	-0.1481	0.0060	0.1634	-0.0015
	SE_1	0.0719	0.0765	0.1054	0.1093	0.1016	0.1100	0.1415	0.1464
	SE_2	0.1064	0.0955	0.1432	0.1283	0.1386	0.1281	0.1741	0.1645
	ECP	0.4819	0.8775	0.6325	0.9157	0.5964	0.9116	0.7329	0.9197

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 3.6: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are $POI(5)$; $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$; $Z \sim N(0,1)$ and $\boldsymbol{X|Z \sim N(0,1)}$ such that $\boldsymbol{\rho_{XZ} = 0}$.*

		$\boldsymbol{\alpha_X}$		$\boldsymbol{\alpha_Z}$		$\boldsymbol{\beta_X}$		$\boldsymbol{\beta_Z}$	
Method		$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$
Naive	Bias	-0.3711	-0.1603	-0.0132	-0.0068	-0.3995	-0.1693	-0.0411	-0.0166
	SE_1	0.0415	0.0542	0.0565	0.0569	0.0553	0.0851	0.0728	0.0743
	SE_2	0.0427	0.0545	0.0582	0.0580	0.0608	0.0750	0.0804	0.0717
	ECP	0.0000	0.1584	0.9220	0.9505	0.0000	0.3802	0.8960	0.9545
Likelihood									
Known	Bias	0.0032	-0.0050	0.0048	0.0002	-0.0075	0.0060	-0.0101	-0.0025
	SE_1	0.0965	0.0747	0.0635	0.0600	0.1361	0.1072	0.0834	0.0791
	SE_2	0.0926	0.0730	0.0637	0.0597	0.1352	0.1043	0.0854	0.0742
	ECP	0.9613	0.9444	0.9505	0.9603	0.9527	0.9603	0.9333	0.9722
Sample I	Bias	0.0137	-0.0050	0.0071	0.0002	-0.0021	0.0063	-0.0076	-0.0024
	SE_1	0.0975	0.0748	0.0644	0.0601	0.1378	0.1073	0.0846	0.0794
	SE_2	0.1288	0.0840	0.0905	0.0674	0.1602	0.1167	0.1099	0.0856
	ECP	0.8520	0.9105	0.8584	0.9205	0.9070	0.9463	0.8816	0.9264
Sample II	Bias	0.0104	-0.0031	0.0041	-0.0025	-0.0036	0.0083	-0.0081	-0.0051
	SE_1	0.0972	0.0750	0.0638	0.0600	0.1374	0.1077	0.0839	0.0792
	SE_2	0.1024	0.0746	0.0686	0.0623	0.1432	0.1079	0.0921	0.0756
	ECP	0.9514	0.9505	0.9323	0.9545	0.9493	0.9604	0.9112	0.9663
SIMEX									
Quadratic	Bias	-0.2033	-0.0344	-0.0146	0.0001	-0.2282	-0.0328	-0.0100	-0.0060
	SE_1	0.0584	0.0671	0.0581	0.0582	0.0793	0.0950	0.0757	0.0767
	SE_2	0.0748	0.0740	0.0645	0.0598	0.0976	0.1054	0.0852	0.0744
	ECP	0.1320	0.8628	0.9160	0.9543	0.2340	0.8966	0.9060	0.9642
Cubic	Bias	-0.1340	-0.0144	-0.0117	0.0013	-0.1578	-0.0111	-0.0041	-0.0048
	SE_1	0.0672	0.0696	0.0592	0.0586	0.0927	0.1008	0.0774	0.0774
	SE_2	0.1030	0.0835	0.0673	0.0617	0.1291	0.1202	0.0893	0.0772
	ECP	0.4540	0.8767	0.9120	0.9443	0.5120	0.9006	0.9000	0.9583

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 3.7: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are POI* (5)*;* $\alpha_0 = \log(0.2)$, $\boldsymbol{\beta_0 = \log(0.22)}$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$*;* $Z \sim N(0,1)$ *and* $X|Z \sim N(1.33Z, 1)$ *such that* $\rho_{XZ} = 0.8$.

		$\boldsymbol{\alpha_X}$		$\boldsymbol{\alpha_Z}$		$\boldsymbol{\beta_X}$		$\boldsymbol{\beta_Z}$	
Method		$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$
Naive	Bias	-0.3690	-0.1417	0.4165	0.1592	-0.3882	-0.1561	0.3863	0.1597
	SE_1	0.0429	0.0566	0.0834	0.0925	0.0572	0.0773	0.1110	0.1211
	SE_2	0.0449	0.0615	0.0884	0.0985	0.0632	0.0814	0.1137	0.1295
	ECP	0.0000	0.3220	0.0000	0.5800	0.0020	0.4420	0.0661	0.7160
Likelihood									
Known	Bias	0.0001	0.0154	0.0042	-0.0165	-0.0076	0.0178	0.0184	-0.0067
	SE_1	0.0959	0.0774	0.1267	0.1099	0.1335	0.1092	0.1603	0.1422
	SE_2	0.0934	0.0825	0.1269	0.1143	0.1352	0.1118	0.1574	0.1474
	ECP	0.9609	0.9362	0.9568	0.9300	0.9547	0.9547	0.9588	0.9403
Sample I	Bias	0.0106	0.0219	-0.0174	-0.0235	0.0019	0.0252	-0.0032	-0.0162
	SE_1	0.0972	0.0782	0.1295	0.1108	0.1356	0.1103	0.1643	0.1433
	SE_2	0.1363	0.1020	0.1895	0.1478	0.1605	0.1238	0.2019	0.1722
	ECP	0.8569	0.8765	0.8487	0.8745	0.8916	0.9218	0.8978	0.9053
Sample II	Bias	0.0039	0.0167	-0.0019	-0.0185	-0.0054	0.0205	0.0144	-0.0103
	SE_1	0.0963	0.0776	0.1273	0.1102	0.1341	0.1094	0.1611	0.1426
	SE_2	0.0986	0.0864	0.1385	0.1229	0.1399	0.1178	0.1646	0.1521
	ECP	0.9501	0.9136	0.9335	0.9280	0.9439	0.9486	0.9480	0.9486
SIMEX									
Quadratic	Bias	-0.2038	-0.0115	0.2314	0.0144	-0.2254	-0.0173	0.2301	0.0259
	SE_1	0.0609	0.0708	0.0959	0.1036	0.0817	0.0979	0.1239	0.1334
	SE_2	0.0773	0.0874	0.1109	0.1198	0.1062	0.1087	0.1393	0.1473
	ECP	0.1443	0.8838	0.3232	0.9118	0.2703	0.9158	0.5244	0.9118
Cubic	Bias	-0.1390	0.0100	0.1606	-0.0080	-0.1574	0.0068	0.1655	0.0030
	SE_1	0.0709	0.0745	0.1034	0.1068	0.0955	0.1037	0.1322	0.1378
	SE_2	0.1034	0.0990	0.1346	0.1311	0.1367	0.1240	0.1611	0.1595
	ECP	0.4460	0.8537	0.5662	0.8918	0.5508	0.9034	0.6965	0.9215

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

Table 3.8: *Empirical performance of estimators of the regression parameters associated with continuous X and Z; Number of assessments are $POI(5)$; $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.22)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$; $Z \sim N(0,1)$ and $\boldsymbol{X|Z \sim N(0,1)}$ such that $\boldsymbol{\rho_{XZ} = 0}$.*

		$\boldsymbol{\alpha_X}$		$\boldsymbol{\alpha_Z}$		$\boldsymbol{\beta_X}$		$\boldsymbol{\beta_Z}$	
Method		$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$	$\gamma = 0.5$	$\gamma = 0.8$
Naive	Bias	-0.3686	-0.1536	-0.0143	-0.0109	-0.3973	-0.1659	-0.0344	-0.0140
	SE_1	0.0401	0.0525	0.0546	0.0550	0.0532	0.0724	0.0706	0.0716
	SE_2	0.0416	0.0543	0.0583	0.0581	0.0518	0.0760	0.0730	0.0744
	ECP	0.0000	0.1860	0.9215	0.4300	0.0000	0.3540	0.9095	0.9200
Likelihood									
Known	Bias	0.0063	0.0051	0.0043	-0.0033	-0.0010	0.0121	-0.0007	-0.0005
	SE_1	0.0918	0.0728	0.0621	0.0583	0.1338	0.1039	0.0805	0.0760
	SE_2	0.0901	0.0713	0.0631	0.0605	0.1394	0.1067	0.0798	0.0765
	ECP	0.9474	0.9485	0.9453	0.9320	0.9352	0.9526	0.9636	0.9526
Sample I	Bias	0.0182	0.0108	0.0037	-0.0041	0.0142	0.0179	0.0017	0.0004
	SE_1	0.0932	0.0734	0.0631	0.0585	0.1363	0.1049	0.0821	0.0764
	SE_2	0.1219	0.0883	0.0936	0.0776	0.1715	0.1196	0.1125	0.0885
	ECP	0.8793	0.9145	0.8384	0.8676	0.8873	0.9124	0.8545	0.9063
Sample II	Bias	0.0088	0.0068	0.0008	-0.0038	0.0010	0.0153	-0.0041	-0.0004
	SE_1	0.0920	0.0731	0.0622	0.0584	0.1336	0.1045	0.0806	0.0762
	SE_2	0.0984	0.0745	0.0666	0.0636	0.1415	0.1098	0.0863	0.0805
	ECP	0.9300	0.9381	0.9280	0.9258	0.9259	0.9464	0.9383	0.9464
SIMEX									
Quadratic	Bias	-0.1967	-0.0237	-0.0059	-0.0050	-0.2318	-0.0252	-0.0199	-0.0029
	SE_1	0.0567	0.0656	0.0562	0.0565	0.0763	0.0918	0.0728	0.0738
	SE_2	0.0716	0.0759	0.0623	0.0604	0.0939	0.1029	0.0776	0.0779
	ECP	0.1263	0.8956	0.9165	0.9257	0.1914	0.9118	0.9185	0.9299
Cubic	Bias	-0.1238	-0.0039	-0.0027	-0.0038	-0.1588	0.0020	-0.0134	0.0002
	SE_1	0.0662	0.0687	0.0574	0.0570	0.0901	0.0979	0.0747	0.0748
	SE_2	0.0974	0.0856	0.0661	0.0618	0.1207	0.1158	0.0810	0.0794
	ECP	0.4969	0.8855	0.9145	0.9217	0.4929	0.8978	0.9206	0.9256

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

would be more feasible to obtain in practice.

DISCUSSION

As is clear from the results, the correct maximum likelihood approach performs much better than the naive maximum likelihood approach. The biases are close to 0 and the empirical coverage probabilities are much closer to the nominal level of 0.95 (with both large and small validation samples). The large validation sample results appear to demonstrate improved performance over the small validation study results. This is not surprising because there is more information about the error and covariate distributions with a larger validation sample. SIMEX performs much better for moderate measurement error (γ=0.8) than for major measurement error (γ=0.5). It is a preferred method over the naive maximum likelihood approach. It is important to recognize that SIMEX is an easy way to implement correction for measurement error and it performs best when the measurement error is low. A drawback, however, is the difficulty in specifying the appropriate extrapolation function. When X and Z are uncorrelated, the measurement error in X does not appear to have a significant impact on estimation of the parameters associated with Z. However, when they are correlated, there can be substantial bias introduced. Consistent with the asymptotic bias plots, on average, the magnitudes of α_X and β_X tend to be underestimated by the naive maximum likelihood method; whereas, the magnitudes of α_Z and β_Z tend to be overestimated. Also, the true underlying values of α_0 and β_0 do not appear to affect the estimated bias and coverage probabilities, at least for the parameter configurations investigated.

As in the binary case, these correct maximum likelihood simulations were based on the likelihood function given in (3.8) rather than a continuous version of (3.7). The difference between the estimated standard errors for the SIMEX, naive and correct maximum likelihood estimators did not appear to be as pronounced for the continuous covariate case compared to the binary covariate case. However, based on the parameter configurations investigated in this simulation study, $SE_{naive} < SE_{SIMEX} < SE_{correct}$ with quadratic and cubic SIMEX extrapolants. SIMEX performance may be improved if the extrapolation model fitting process was not automated as in these simulations or another standard error

estimation approach such as the bootstrap was used.

A small number of simulations were also performed by setting the average number of assessments to 20 rather than $\mu = 5$ which resulted in the above results. Table 3.9 summarizes the results comparing two simulations for a particular parameter configuration; one with $\mu = 5$ and one with $\mu = 20$. Upon inspection of the results, there does not appear to be much of a difference in the empirical biases and estimated standard errors for the three approaches between an average of five assessments and twenty assessments.

Table 3.9: *Comparison of the empirical performance of estimators of the regression parameters associated with continuous X and Z when* $\boldsymbol{\mu = 5}$ *and* $\boldsymbol{\mu = 20}$*;* $\alpha_0 = \log(0.2)$, $\beta_0 = \log(0.4)$, $\alpha_X = \beta_X = \log(2)$, $\alpha_Z = \beta_Z = \log(1.25)$*;* $\gamma = 0.5$, $Z \sim N(0,1)$ *and* $X|Z \sim N(1.33Z, 1)$ *such that* $\rho_{XZ} = 0.8$.

		$\boldsymbol{\alpha_X}$		$\boldsymbol{\alpha_Z}$		$\boldsymbol{\beta_X}$		$\boldsymbol{\beta_Z}$	
Method		$\mu = 5$	$\mu = 20$	$\mu = 5$	$\mu = 20$	$\mu = 5$	$\mu = 20$	$\mu = 5$	$\mu = 20$
Naive	Bias	-0.3685	-0.3691	0.4152	0.4204	-0.3851	-0.3877	0.3963	0.3875
	SE_1	0.0440	0.0437	0.0852	0.0853	0.0607	0.0599	0.1187	0.1171
	SE_2	0.0455	0.0474	0.0932	0.0939	0.0636	0.0632	0.1268	0.1218
	ECP	0.0000	0.0000	0.0000	0.0021	0.0000	0.0000	0.0800	0.0951
Likelihood									
Known	Bias	0.0024	0.0023	0.0026	0.0078	-0.0031	-0.0078	0.0241	0.0154
	SE_1	0.0989	0.0983	0.1299	0.1301	0.1408	0.1391	0.1703	0.1690
	SE_2	0.0952	0.1014	0.1302	0.1364	0.1416	0.1414	0.1665	0.1635
	ECP	0.9600	0.9322	0.9520	0.9384	0.9620	0.1271	0.9600	0.9554
Sample I	Bias	0.0173	0.0112	-0.0219	-0.0092	0.0101	-0.0004	0.0026	0.0012
	SE_1	0.1011	0.0998	0.1332	0.1323	0.1438	0.1403	0.1746	0.1718
	SE_2	0.1223	0.1374	0.1742	0.1915	0.1583	0.1622	0.1968	0.1964
	ECP	0.8898	0.8726	0.8637	0.8323	0.9399	0.8917	0.9259	0.9214
Sample II	Bias	0.0111	0.0079	-0.0105	-0.0042	0.0066	-0.0007	0.0108	0.0022
	SE_1	0.1002	0.0993	0.1315	0.1316	0.1427	0.1399	0.1724	0.1708
	SE_2	0.1011	0.1068	0.1379	0.1432	0.1486	0.1486	0.1714	0.1705
	ECP	0.9439	0.9577	0.9419	0.9387	0.9499	0.9345	0.9619	0.9493
SIMEX									
Quadratic	Bias	-0.1996	-0.2024	0.2269	0.2324	-0.2185	-0.2246	0.2323	0.2267
	SE_1	0.0623	0.0619	0.0980	0.0984	0.0870	0.0858	0.1325	0.1311
	SE_2	0.0778	0.0806	0.1179	0.1217	0.1051	0.1048	0.1504	0.1438
	ECP	0.1747	0.1588	0.3594	0.3584	0.3133	0.2983	0.5562	0.5644
Cubic	Bias	-0.1310	-0.1324	0.1519	0.1539	-0.1481	-0.1566	0.1634	0.1593
	SE_1	0.0719	0.0717	0.1054	0.1060	0.1016	0.1003	0.1415	0.1400
	SE_2	0.1064	0.1099	0.1432	0.1507	0.1386	0.1315	0.1741	0.1647
	ECP	0.4819	0.4700	0.6325	0.6052	0.5964	0.5558	0.7329	0.7232

SE_1 and SE_2 : average model-based and empirical standard errors, respectively
ECP: empirical coverage probability (proportion of 95% CI's that include true parameter value)
Sample I and Sample II: small (50) and large (200) validation samples, respectively (SIMEX based on Sample I)
Known: based on using the true parameter values for misclassification and $X|Z$ distributions

3.5 Application: Psoriatic Arthritis Data

This analysis was based on data extracted from the PsA clinic database as of early 2005 but assuming a three-state model similar to Figure 3.2. Extending the methodology presented here to models with a larger number of states is straightforward. As in Gladman et al. (1995), the states were determined based on a clinical assessment; the number of deformed joints. State 1, State 2 and State 3 were defined to represent 0, $1-4$ and $5+$ deformed joints, respectively. For the purposes of this analysis, we will assume that the response (i.e. damaged joint count) is perfectly measured. Gladman et al. (1990) assessed the reliability of the actively inflamed and deformed joint counts based on the American College of Rheumatology (ACR) joint count within the clinic and report these counts as reliable in Gladman et al. (1995). In the available dataset 383 patients entered the PsA Clinic in State 1, 130 in State 2 and 106 in State 3. Along with the demographics of the patients included in these analyses, Table 3.10 presents variables which have been identified as factors potentially associated with PsA progression. The presence of dactylitis and the back measurements were among the variables that were investigated in the reliability study (Gladman et al. 2004). Although information on these and the perfectly measured variables are collected at each clinic visit, we use baseline covariate data only in the regression models. Two models are fit to these data. One includes a binary covariate subject to misclassification which will be fit along with several fixed, precisely measured variables. The second involves a continuous variable subject to error along with several variables assumed to be precisely measured. In both cases, stratification based on state at clinic entry is done by including the indicator variable Z_{S_2} in the second transition intensity. However no interactions are considered in these analyses. The parameters of the misclassification or measurement error process and the conditional covariate distribution were estimated as outlined in Sections 2.6.1 and 2.6.2.

Table 3.10: *Patient demographics and covariates at clinic entry.*

		n = 619
Gender (Z_G)	Women	270
	Men	349
Age at PsA Diagnosis (Z_{AP})	Average	**35.8**
	Range	(9-86)
PsA Duration (years) (Z_{DP})	Average	**7.4**
	Range	(0-47.7)
Number of Effused Joints (Z_E)	Average	**3.2**
	Range	(0-33)
Presence of Dactylitis (W_D)	Yes	214
	No	405
Back Measurements		
Upper Back (W_U)	Average	**2.2**
	Range	(0-9)
Middle Back (W_M)	Average	**3.2**
	Range	(0-7.8)
Lower Back (W_L)	Average	**3.9**
	Range	(0-8.5)

3.5.1 Misclassification in a Binary Covariate

The results for the full model in which covariate effects vary across transitions are summarized in Table 3.11. Likelihood ratio tests based on 5 degrees of freedom were carried out for both the naive and correct maximum likelihood approaches, comparing the full model to a model which assumes common effects across transitions. Both tests suggest that the simpler model is reasonable (p=0.5334 for the naive model, p=0.6101 for the correct model). Table 3.12 summarizes the common effects model. There does not appear to be a substantial difference between the parameter estimates and estimated standard errors across the methods. This is most likely due to the apparent lack of effect for the dactylitis variable. The estimated dactylitis effect and corresponding standard error are larger for the correct likelihood approach compared to the other methods, but the effect does not appear to be significantly different from zero. The SIMEX approach, however, does suggest that the dactylitis effect is significantly different from zero.

For SIMEX, backwards elimination was performed based on the variance approximation of Stefanski & Cook (1995). After the insignificant variables were dropped from each of the naive and correct likelihood models via likelihood ratio tests, results were obtained as summarized in Table 3.13. The dactylitis variable is included in the naive and correct likelihood approaches for comparison purposes even though its effect was not significant under either model. Figure 3.23 contains the SIMEX plots associated with the final model based on the SIMEX procedure. It is interesting to note the trend in the parameter estimates obtained by increasing the degree of misclassification on the already misclassified variable, W. For the most part, the quadratic extrapolant appears to provide the best fit here for both parameter and variance estimates.

In all three approaches, the number of effused joints at clinic entry appears to be associated with the progression of PsA; the larger the number of effused joints at clinic entry, the higher the risk of progressing to the next state. The relative risk estimates of PsA progression with each additional swollen joint at clinic entry are RR=1.0525 [95% CI (1.0208,1.0852)], RR=1.0522 [95% CI (1.0197,1.0857)] and RR=1.0491 [95% CI (1.0255,1.0732)] for the naive and correct likelihood approaches and SIMEX, respectively.

There appears to be little difference between the three estimates, probably because the effect of the misclassified variable does not differ significantly from 0 (at least based on the likelihood approaches).

The SIMEX approach suggests a marginally significant effect of the presence of dactylitis on PsA progression. The resulting relative risk estimate is RR=1.3015 [95% CI (1.0110,1.6756)], suggesting that patients with at least one swollen digit at clinic entry are at a higher risk of developing damaged joints. The analyses in Chapter 2, as well as the naive and correct likelihood approaches here found the dactylitis effect to be insignificant. The SIMEX standard error may be underestimated here since the validity of the variance approximation of Stefanski & Cook (1995) depends on known misclassification probabilities and extrapolation function. Based on R^2_{adj}, the cubic extrapolant appeared to provide the best fit to the variance approximations at each ν. However, as shown in Figure 3.23, the resulting extrapolation function is not monotonic which raises questions regarding its appropriateness here. Under the next best model, a quadratic extrapolant, the standard error was 0.1847. When compared to the estimated dactylitis effect of 0.2635, a p-value of approximately 0.1547 is obtained based on a quadratic extrapolant.

Table 3.11: *Estimates obtained by fitting naive and correct models and applying the SIMEX procedure to the PsA clinic data with a misclassified binary covariate (X_D).*

		Naive			Correct			SIMEX		
Transition	Variable (Parameter)	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
$1 \rightarrow 2$	Intercept (α_0)	-2.9902	0.2990	<0.0001	-3.1824	0.3429	<0.0001	-3.0269	0.2968	<0.0001
$1 \rightarrow 2$	Age at PsA Diagnosis ($\alpha_{Z_{AP}}$)	0.0064	0.0067	0.3342	0.0068	0.0068	0.3221	0.0063	0.0066	0.3335
$1 \rightarrow 2$	PsA duration ($\alpha_{Z_{DP}}$)	0.0204	0.0109	0.0628	0.0199	0.0113	0.0777	0.0174	0.0111	0.1169
$1 \rightarrow 2$	Gender (α_{Z_G})	0.1410	0.1511	0.3512	0.1526	0.1539	0.3220	0.2126	0.1515	0.1611
$1 \rightarrow 2$	Effused joint count (α_{Z_E})	0.0519	0.0159	0.0012	0.0531	0.0161	0.0011	0.0449	0.0149	0.0030
$1 \rightarrow 2$	Presence of dactylitis (α_{X_D})	0.2324	0.1514	0.1254	0.4631	0.2855	0.1054	0.5663	0.1667	0.0007
$2 \rightarrow 3$	Intercept (β_0)	-2.3885	0.3062	<0.0001	-2.8347	0.5894	<0.0001	-2.3900	0.3445	<0.0001
$2 \rightarrow 3$	Age at PsA Diagnosis ($\beta_{Z_{AP}}$)	0.0091	0.0066	0.1704	0.0114	0.0074	0.1252	0.0090	0.0069	0.1954
$2 \rightarrow 3$	PsA duration ($\beta_{Z_{DP}}$)	0.0069	0.0116	0.5536	0.0105	0.0139	0.4508	0.0117	0.0123	0.3414
$2 \rightarrow 3$	Gender (β_{Z_G})	-0.1957	0.1629	0.2302	-0.2352	0.1811	0.1946	-0.2604	0.1654	0.1166
$2 \rightarrow 3$	Effused joint count (β_{Z_E})	0.0372	0.0174	0.0336	0.0393	0.0195	0.0444	0.0396	0.0174	0.0233
$2 \rightarrow 3$	Presence of dactylitis (β_{X_D})	0.0729	0.1713	0.6706	0.6199	0.6647	0.3514	0.1369	0.3031	0.6517
$2 \rightarrow 3$	State 2 clinic entry ($\beta_{Z_{S_2}}$)	-0.1524	0.1653	0.3570	-0.1144	0.1767	0.5176	-0.0683	0.1718	0.6911

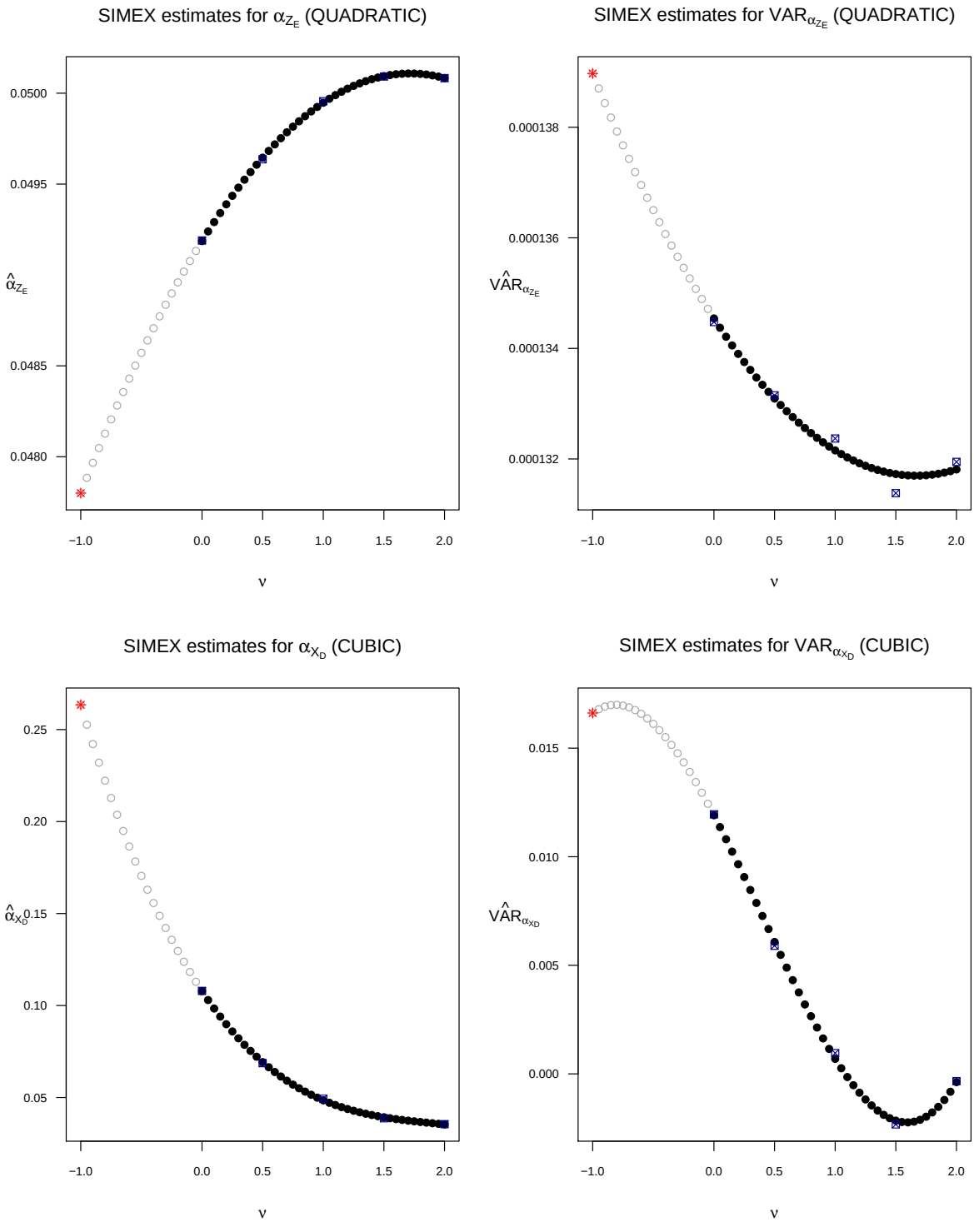
Table 3.12: *Estimates obtained by fitting naive and correct models and applying the SIMEX procedure to the PsA clinic data with a misclassified binary covariate (X_D) assuming common effects across transitions.*

Variable (Parameter)	Naive			Correct			SIMEX		
	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
1^{st} Intercept (α_0)	-2.8696	0.2158	<0.0001	-3.1136	0.2816	<0.0001	-2.8790	0.2255	<0.0001
Age at PsA Diagnosis ($\alpha_{Z_{AP}}$)	0.0078	0.0046	0.0926	0.0088	0.0048	0.0664	0.0070	0.0047	0.1363
PsA duration ($\alpha_{Z_{DP}}$)	0.0140	0.0079	0.0790	0.0145	0.0082	0.0782	0.0101	0.0078	0.1983
Gender (α_{Z_G})	-0.0165	0.1097	0.8805	-0.0227	0.1135	0.8413	0.1189	0.1106	0.2829
Effused joint count (α_{Z_E})	0.0456	0.0118	0.0001	0.0462	0.0123	0.0002	0.0428	0.0120	0.0004
Presence of dactylitis (α_{X_D})	0.1572	0.1125	0.1629	0.4777	0.2719	0.0796	0.3775	0.1458	0.0100
2^{nd} Intercept (β_0)	-2.5667	0.2285	<0.0001	-2.8265	0.3025	<0.0001	-2.5660	0.2366	<0.0001
State 2 clinic entry ($\beta_{Z_{S_2}}$)	-0.1188	0.1622	0.4645	-0.1110	0.1650	0.5014	-0.1600	0.1630	0.3270

Table 3.13: *Final model estimates obtained by fitting naive and correct models and applying the SIMEX procedure to the PsA clinic data with a misclassified binary covariate (X_D) assuming common effects across transitions.*

	Naive			Correct			SIMEX		
Variable (Parameter)	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
1^{st} Intercept (α_0)	-2.5749	0.0980	<0.0001	-2.7362	0.1805	<0.0001	-2.5770	0.0915	<0.0001
Effused joint count (α_{Z_E})	0.0512	0.0156	0.0011	0.0509	0.0160	0.0016	0.0479	0.0116	<0.0001
Presence of dactylitis (α_{X_D})	0.2255	0.1493	0.1318	0.4384	0.2888	0.1299	0.2635	0.1289	0.0415
2^{nd} Intercept (β_0)	-2.0430	0.0783	<0.0001	-2.0441	0.0785	<0.0001	-2.2912	0.1115	<0.0001

Figure 3.23: *SIMEX estimates obtained by applying the SIMEX procedure (based on selection of an extrapolant and the variance approximation procedure of Stefanski & Cook (1995)) to the PsA clinic data with a misclassified binary covariate (X_D) assuming common effects across transitions (Table 3.13).*



3.5.2 Measurement Error in a Continuous Covariate

Here we will fit one error-prone variable, the range of motion variable corresponding to the middle back, along with several other variables assumed to be perfectly measured: gender, age at PsA onset, duration of PsA at clinic entry, the number of effused joints at clinic entry and the extent of the joint damage (i.e. the state) at clinic entry. As in Section 3.5.1, the first model fit was general in the sense that it permits the covariate effects to differ across the two transitions. The second, reduced model that will be fit assumes that the covariate effects are the same for both transitions.

The results for the three estimation approaches; naive maximum likelihood, correct maximum likelihood and SIMEX, are summarized in Tables 3.14 and 3.15. As in the simulation studies, the SIMEX approach involved repeated estimation using naive maximum likelihood ($B = 150$, here) for different multiples of induced measurement error according to $\boldsymbol{\nu} = \{0, 0.5, 1, 1.5, 2\}$. Candidate extrapolation functions that were considered for both the parameter estimates and the variance estimates included linear, quadratic, exponential and nonlinear (rational linear) functions. Error sums of squares and adjusted R^2 ($R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$) were considered to determine the extrapolation function that provided the best fit.

Since the reliability data is external, both the correct likelihood approach and SIMEX treat the measurement error variance as known. The likelihood approach also treats the parameters associated with the conditional covariate distribution as known. If the sampling variability of the estimators of these parameters was incorporated into the correct likelihood and SIMEX analyses (using the bootstrap for example), we would expect that the standard errors would increase. This may prevent us from identifying truly significant variables since their effects would be masked by large standard errors. The corresponding relative risk confidence intervals would also be wider. This supports the use of large supplementary datasets. We would expect the variability associated with the measurement error and covariate distribution parameters to decrease as the size of the supplementary dataset increases.

Likelihood ratio tests were used for the naive and correct likelihood approaches to determine the final models. To determine the final model based on the SIMEX approach, variables were omitted if they did not appear to be significantly different from zero. Estimates from the final models for the three approaches are summarized in Table 3.16 and the corresponding SIMEX plots are displayed in Figures 3.24 and 3.25. Note the effect of the increase in measurement error on parameter estimation in this setting.

Based on the final model, the number of swollen joints has an effect on PsA progression similar to that which was observed in the binary case. In addition, the error-prone variable, X_M, appears to have a significant effect on the second transition, but not the first. Although the naive likelihood approach does suggest that it is significantly different from zero, the magnitude of the effect appears to be underestimated. In terms of relative risk estimates, the naive likelihood approach results in RR=0.8437 [95% CI (0.7186,0.9906)], the correct likelihood approach gives RR=0.5597 [95% CI (0.5278,0.5936)] and for the SIMEX approach, RR=0.7578 [95% CI (0.7195,0.7982)]. Therefore patients who have one additional centimeter of middle back mobility at clinic entry and who have at least one damaged joint are at lower risk of developing a total of five or more damaged joints. The naive likelihood approach appears to understate this risk reduction. The difference that we observe between the correct likelihood approach and SIMEX could be a result of misspecification of the underlying conditional covariate distribution. This would affect only the correct likelihood approach. Sensitivity analyses could be conducted to explore the effects of such misspecification on parameter estimation.

Table 3.14: *Estimates obtained by fitting naive and correct models and applying the SIMEX procedure to the PsA clinic data with an error-prone continuous covariate (X_M).*

		Naive			Correct			SIMEX		
Transition	Variable (Parameter)	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
$1 \rightarrow 2$	Intercept (α_0)	-2.8730	0.4088	<0.0001	-2.7794	0.7852	0.0004	-2.3467	0.1194	<0.0001
$1 \rightarrow 2$	Age at PsA Diagnosis ($\alpha_{Z_{AP}}$)	0.0066	0.0073	0.3628	0.0070	0.0073	0.3314	0.0075	0.0071	0.3113
$1 \rightarrow 2$	PsA duration ($\alpha_{Z_{DP}}$)	0.0211	0.0113	0.0610	0.0206	0.0114	0.0702	0.0207	0.0106	0.0514
$1 \rightarrow 2$	Gender (α_{Z_G})	0.1508	0.1558	0.3331	0.1630	0.1559	0.2958	0.1588	0.1570	0.3123
$1 \rightarrow 2$	Effused joint count (α_{Z_E})	0.0562	0.0154	0.0003	0.0560	0.0156	0.0003	0.0588	0.0152	0.0001
$1 \rightarrow 2$	Middle back ROM (α_{X_M})	-0.0134	0.0724	0.8532	-0.0422	0.1843	0.8189	-0.0326	0.1241	0.7929
$2 \rightarrow 3$	Intercept (β_0)	-1.6191	0.4931	0.0010	0.6380	1.0442	0.5412	-1.8547	1.2763	0.1468
$2 \rightarrow 3$	Age at PsA Diagnosis ($\beta_{Z_{AP}}$)	0.0046	0.0073	0.5260	0.0026	0.0074	0.6980	0.0089	0.0083	0.2841
$2 \rightarrow 3$	PsA duration ($\beta_{Z_{DP}}$)	-0.0027	0.0128	0.8345	-0.0089	0.0129	0.4889	-0.0002	0.0145	0.9890
$2 \rightarrow 3$	Gender (β_{Z_G})	-0.1787	0.1641	0.2762	-0.1938	0.1744	0.2665	-0.2195	0.1631	0.1784
$2 \rightarrow 3$	Effused joint count (β_{Z_E})	0.0350	0.0179	0.0503	0.0389	0.0199	0.0505	0.0344	0.0166	0.0388
$2 \rightarrow 3$	Middle back ROM (β_{X_M})	-0.1551	0.0899	0.0846	-0.7167	0.2505	0.0042	-0.1793	0.1629	0.2710
$2 \rightarrow 3$	State 2 clinic entry ($\beta_{Z_{S_2}}$)	-0.1466	0.1665	0.3786	-0.1313	0.1772	0.4587	-0.2023	0.1667	0.2249

Table 3.15: *Estimates obtained by fitting naive and correct models and applying the SIMEX procedure to the PsA clinic data with an error-prone continuous covariate (X_M) assuming common effects across transitions.*

	Naive			Correct			SIMEX		
Variable (Parameter)	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
1^{st} Intercept (α_0)	-2.5708	0.3088	<0.0001	-2.0215	0.6798	0.0029	-2.4625	0.4572	<0.0001
Age at PsA Diagnosis ($\alpha_{Z_{AP}}$)	0.0064	0.0051	0.2065	0.0064	0.0052	0.2178	0.0062	0.0046	0.1786
PsA duration ($\alpha_{Z_{DP}}$)	0.0112	0.0083	0.1766	0.0099	0.0084	0.2394	0.0099	0.0083	0.2348
Gender (α_{Z_G})	0.0003	0.1116	0.9979	0.0129	0.1127	0.9089	0.0129	0.1132	0.9093
Effused joint count (α_{Z_E})	0.0457	0.0117	<0.0001	0.0459	0.0119	0.0001	0.0440	0.0119	0.0002
Middle back ROM (α_{X_M})	-0.0525	0.0550	0.3398	-0.1939	0.1607	0.2276	-0.0611	0.0664	0.3575
2^{nd} Intercept (β_0)	-2.2301	0.3236	<0.0001	-1.6866	0.6849	0.0138	-1.8119	0.4744	0.0002
State 2 clinic entry ($\beta_{Z_{S_2}}$)	-0.1680	0.1649	0.3083	-0.1529	0.1664	0.3582	-0.1432	0.1678	0.3938

Table 3.16: *Final model estimates obtained by fitting naive and correct models and applying the SIMEX procedure to the PsA clinic data with an error-prone continuous covariate (X_M).*

		Naive			Correct			SIMEX		
Transition	Variable (Parameter)	Estimate	SE	P-value	Estimate	SE	P-value	Estimate	SE	P-value
$1 \rightarrow 2$	Intercept (α_0)	-2.4859	0.0853	0.0001	-2.4850	0.0858	<0.0001	-2.5261	0.1071	<0.0001
$1 \rightarrow 2$	Effused joint count (α_{Z_E})	0.0555	0.0151	0.0002	0.0552	0.0151	0.0003	0.0552	0.0154	0.0004
$2 \rightarrow 3$	Intercept (β_0)	-1.5982	0.2823	<0.0001						
$2 \rightarrow 3$	Effused joint count (β_{Z_E})	0.0386	0.0178	0.0304	0.0406	0.0191	0.0337	0.0459	0.0240	0.0567
$2 \rightarrow 3$	Middle back ROM (β_{X_M})	-0.1699	0.0819	0.0380	-0.5803	0.0300	<0.0001	-0.2773	0.0265	<0.0001

Figure 3.24: *Final model estimates of parameters corresponding to the first transition obtained by applying the SIMEX procedure (based on selection of an extrapolant and the variance approximation procedure of Stefanski & Cook (1995)) to the PsA clinic data with an error-prone continuous covariate (X_M).*

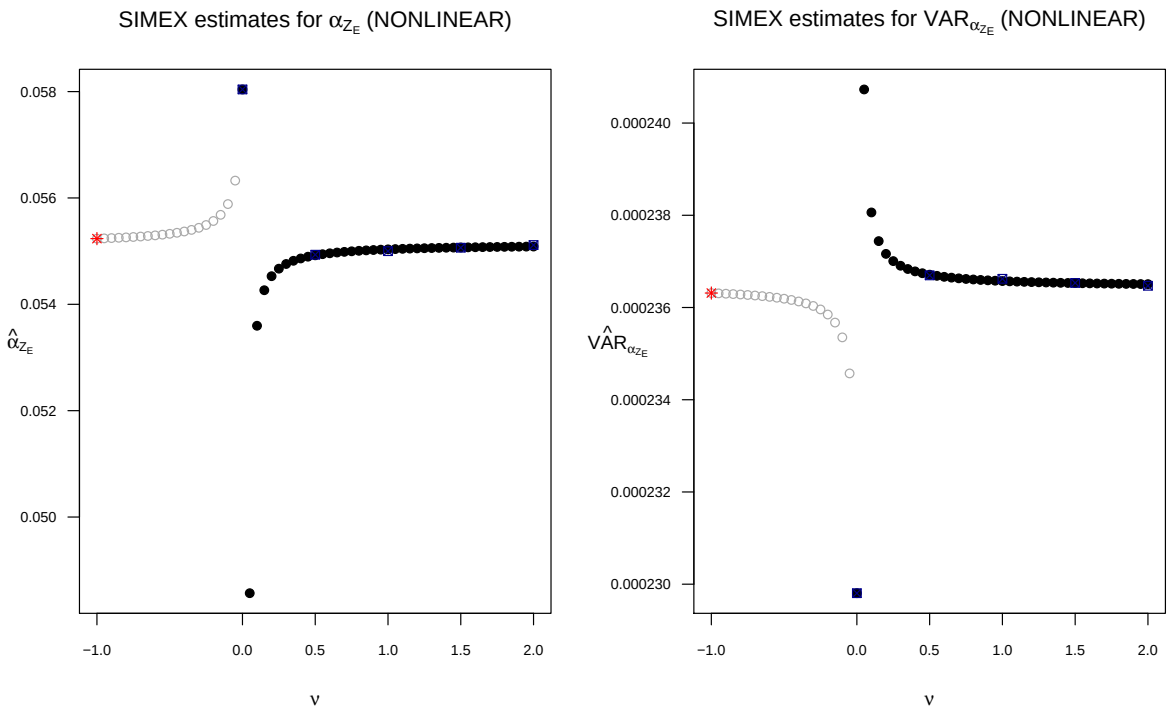
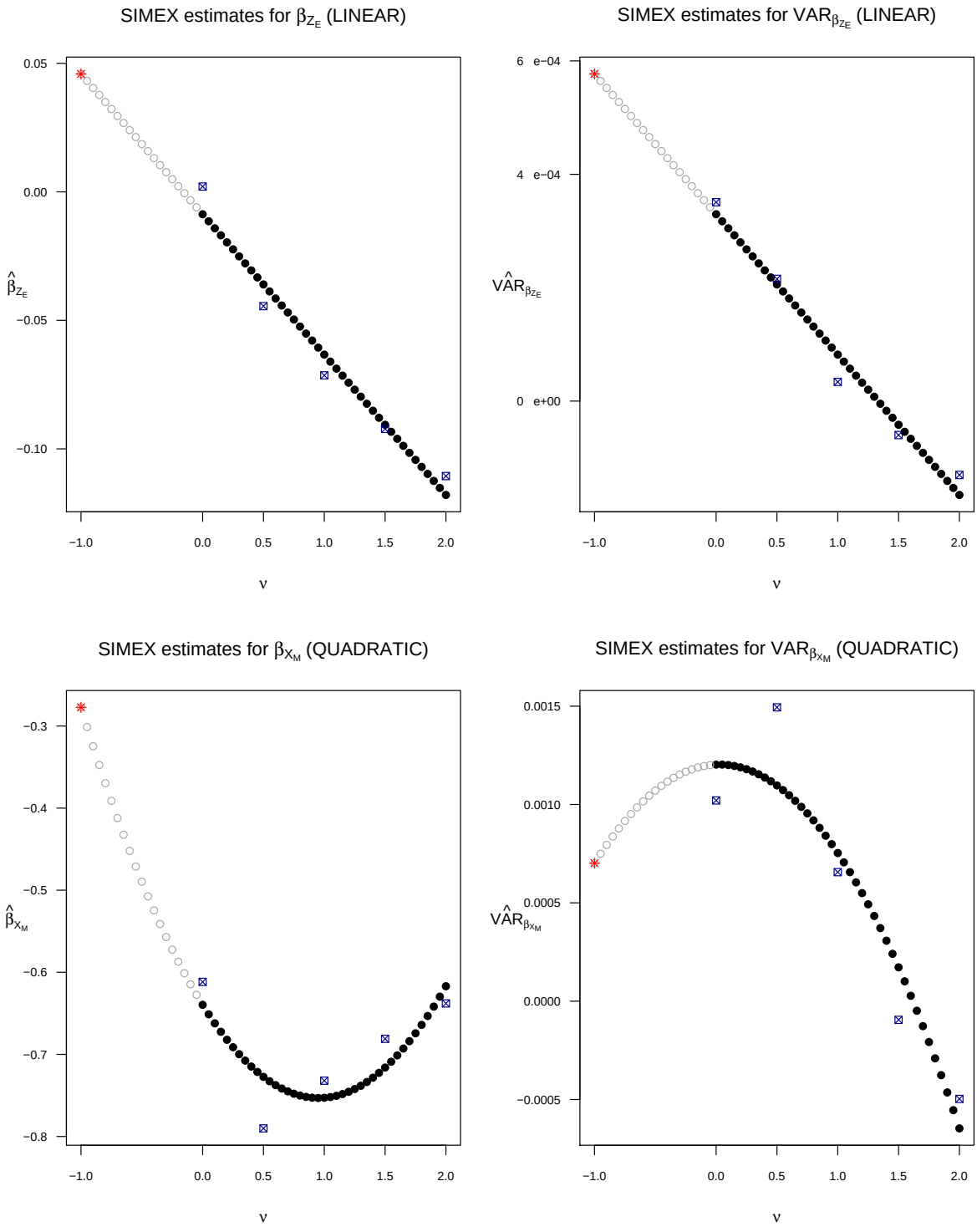


Figure 3.25: *Final model estimates of parameters corresponding to the second transition obtained by applying the SIMEX procedure (based on selection of an extrapolant and the variance approximation procedure of Stefanski & Cook (1995)) to the PsA clinic data with an error-prone continuous covariate (X_M).*



Chapter 4

Current Status Data with a Susceptible Fraction

4.1 Overview

In the analysis of lifetime data, individuals who do not experience the event of interest by the end of the study are typically treated as having right-censored event times. However, if a subgroup of the individuals will never experience the event of interest (i.e. there is a non-susceptible fraction of the population), their event times are undefined, but are often taken to be infinite. This was briefly discussed in Section 1.4. Another complication arises when it is very difficult or costly in terms of time or money to assess individuals repeatedly over time as discussed in Chapters 2 and 3. If this is the case, a single follow-up assessment is sometimes planned, leading to event times that are either left-censored or right-censored. Such data are called type I interval-censored data; or sometimes current status data (Sun 2006). This chapter is concerned with estimation of the parameters associated with the probability of experiencing an event (i.e. being susceptible), as well as the lifetime distribution for the susceptible subpopulation. Lam & Xue (2005) considered a similar problem and proposed a semi-parametric mixture model involving a logistic model for the event probability and a semi-parametric accelerated failure time model for the event time distribution for the susceptible group. They allowed for covariates to affect both components (i.e. event probability and event time distribution) and used sieve maximum likelihood to

obtain estimates of the parameters (Lam & Xue 2005). For the event time distribution here, standard parametric and piecewise constant hazards models will be considered along with a nonparametric approach. The methods developed will be applied to data arising from a series of studies involving orthopedic surgery patients.

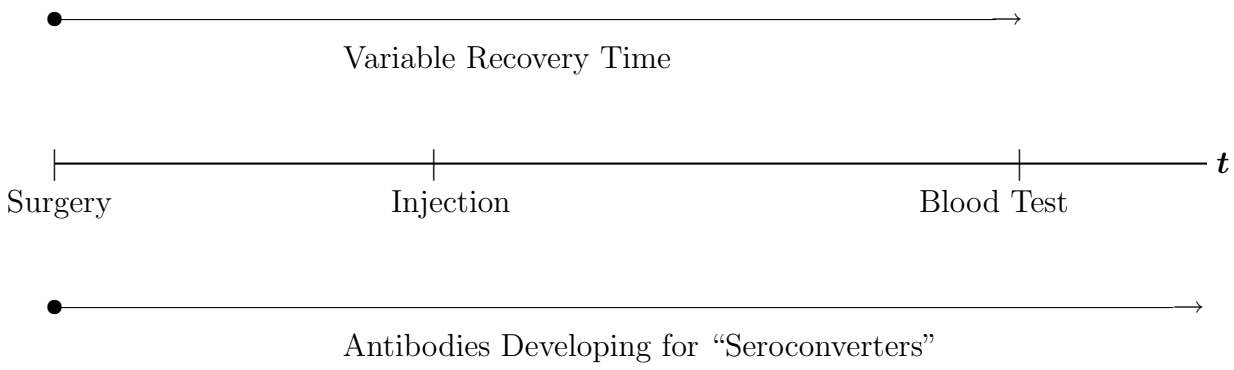
4.2 Motivating Study

Patients undergoing orthopedic surgery such as hip or knee replacement are at increased risk of developing thrombosis or potentially fatal blood clots. To prevent the formation of these blood clots heparin-based blood thinners are currently administered to patients undergoing these surgeries. Unfortunately some patients (reported to be approximately 5%) develop an adverse reaction to surgery and treatment known as *Heparin-induced thrombocytopenia* (HIT). This is characterized by the development of antibodies of the IgG class and a rapid drop in platelet counts which increases the risk of bleeding. A series of international orthopedic surgical trials were recently conducted looking into alternative medications for the prevention of blood clots: in North America, the Pentamaks study which involved knee replacement (Bauer et al. 2001), the Pentathlon study which involved hip replacement (Turpie et al. 2002), and in Europe, the Ephesus (Lassen et al 2002) and Pentifra (Eriksson et al. 2001) studies, both involving hip surgeries. Only the Pentifra study dealt with hip surgery due to fractures. The primary objective of these studies was to evaluate the relative performance of a new anticoagulant (Fondaparinux) versus the standard drug therapy (low molecular weight heparin-based enoxaparin) in the prevention of venographically-documented thrombosis. Some of the patients treated with the heparin-based drug enoxaparin will experience seroconversion and it is also of interest to understand the factors associated with such a response.

Antibodies usually develop, if they do at all, between five and ten days after surgery. In these studies, injections were not given at the same time for all patients. Some patients received their first dose of medication prior to surgery, while others received it after surgery. Patients recovered in the hospital and blood tests were conducted upon discharge to assess seroconversion status. Figure 4.1 illustrates the scenario for a subject receiving the medi-

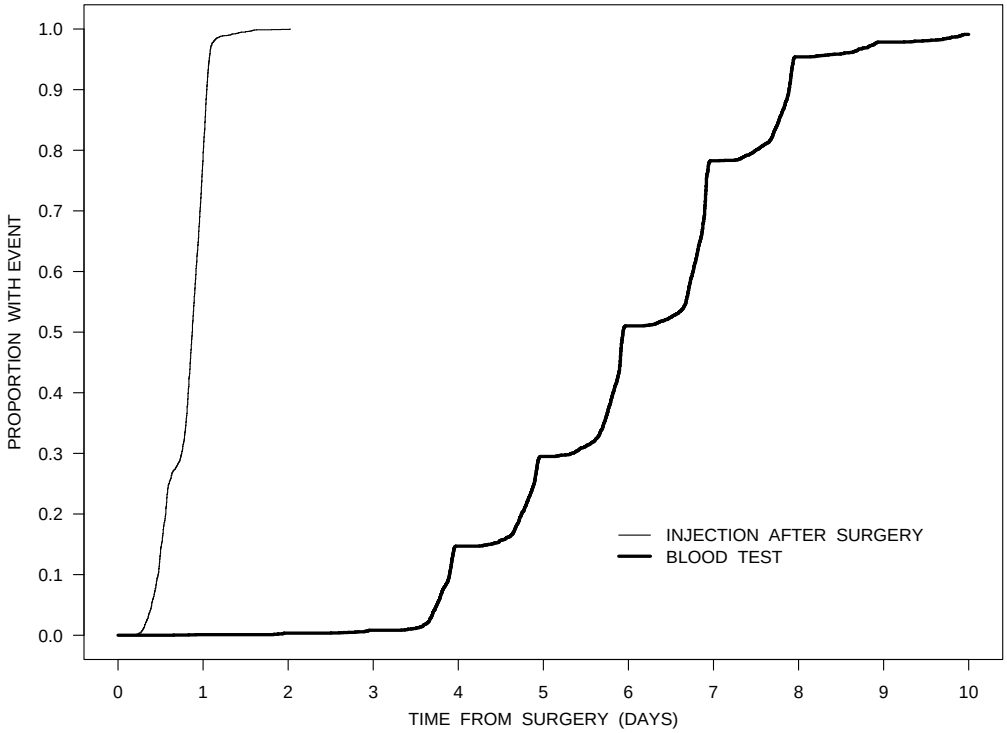
cation post-surgery. Here it is of interest to characterize the probability of seroconversion in patients following surgery and heparin-based anticoagulation therapy so we focus on the 3150 patients (1904 women, 1246 men) who received enoxaparin in the four studies. Most of the patients underwent hip surgery (88.4%) while the others were having knee surgery and the first injection was administered prior to surgery for 67.9% of the patients. Figure 4.2 displays the empirical distributions of the injection times and the discharge (i.e. blood test) times with respect to the surgery times. The median time between surgery and the first postsurgical injection was 0.517 days and the median recovery period following surgery was 5.934 days. The irregular shape of the empirical cumulative distribution function for the time to the blood test reflects the fact that patients were not discharged, and hence blood samples were not taken, during the night.

Figure 4.1: *An illustration of the underlying process over time t for an arbitrary subject.*



Interest primarily lies in whether or not patients develop HIT antibodies rather than when these antibodies develop. In other words, it is of interest to investigate factors related to the probability that an individual will experience the event (seroconversion) rather than related to the timing of the event. In the following sections, we first consider issues related to model misspecification. This is motivated by the fact that early analyses of this data were based on naive models involving a binary analysis of the seroconversion status at the time of testing. This analysis fails to address the fact that individuals recovering from

Figure 4.2: *Empirical distributions of time from surgery to i) injection after surgery and ii) blood sample, for the 3150 patients receiving enoxaparin.*



surgery and tested early do not have as much time to develop antibodies as individuals who were tested much later following surgery. Alternative analyses involve the use of standard current status models which assume all subjects will eventually seroconvert. Instead we propose a simple latent class model which gives estimates of parameters more closely related to the question of primary interest. An EM algorithm is proposed for parameter estimation, and profile likelihood intervals are used for the construction of confidence intervals. This method of estimation is assessed via simulation and applied to the motivating data from the orthopedic surgery studies.

4.3 Statistical Methodology

The following notation will be used throughout this discussion. Let

- $X_i = \begin{cases} 1, & \text{if patient } i \text{ is a seroconverter} \\ 0, & \text{otherwise} \end{cases}$,
- π = probability of seroconversion for a one sample problem such that $P(X_i = 1) = \pi$ and $P(X_i = 0) = 1 - \pi$,
- S_i = time to seroconversion ($S_i \to \infty$ if $X_i = 0$),
- $\mathcal{F}_S(\cdot)$ = survival function of time to seroconversion for subpopulation of patients who will experience this event (i.e. those with $X_i = 1$),
- B_i = random variable which represents the time from surgery to the blood test for individual i,
- $W_i = \begin{cases} 1, & \text{if seroconversion occurred for individual } i \text{ by time } B_i \\ 0, & \text{otherwise} \end{cases}$, and
- Z_i = a covariate of interest.

Note then, that X_i is unobserved because of the inspection scheme and $W_i = I(S_i < B_i)$. Interest lies in identifying prognostic variables for seroconversion and estimating their effects. To this end, we consider logistic regression models such as

$$\log\left(\frac{\pi(Z_i)}{1-\pi(Z_i)}\right) = \psi_0 + \psi_1 Z_i. \tag{4.1}$$

4.3.1 Model Misspecification

Since we are interested in modeling the probability of seroconversion and the seroconversion status has been determined at hospital discharge, it might be tempting to fit a naive model, treating W_i as the true binary response, ignoring the seroconversion time distribution. For a one sample problem (i.e. no covariates), based on White (1982), we can solve (2.6) to obtain expressions relating the limiting values of the naive estimator to the

parameter(s) of the "true" distribution.

For the purpose of this discussion, we will consider assessment times that follow a $GAM(\gamma_1, \gamma_2)$ with mean $\mu = \gamma_1\gamma_2$, variance $\phi = \gamma\gamma_2^2$ and p.d.f.

$$g_B(b^*) = \frac{(b^*)^{\gamma_1 - 1}\exp(-b^*/\gamma_2)}{\Gamma(\gamma_1)\gamma_1^{\gamma_2}}. \tag{4.2}$$

To avoid unrealistic situations with extremely large inspection times however, if $b^* > 1$, we set the inspection time to 1, $B = \min(B^*, 1)$. We consider exponentially distributed seroconversion times (i.e. $S_i \sim EXP(\lambda)$) and let $\rho = P(S_i < B_i | X_i = 1)$. We consider a naive analysis based on the assumption that $W_i \sim BIN(1, \pi^*)$. Then if $\boldsymbol{X} = (X_1, X_2, ..., X_n)'$, $\boldsymbol{B} = (B_1, B_2, ..., B_n)'$ and (X_i, B_i) are i.i.d.,

$$\begin{aligned}
E\left[S_{naive}\left(\pi^*\right); \pi, \lambda, \gamma_1, \gamma_2\right] &= E\left[\sum_{i=1}^{n}\left(\frac{W_i}{\pi^*} - \frac{1-W_i}{1-\pi^*}\right)\right] \\
&= E\left[\sum_{i=1}^{n}\left(\frac{W_i}{\pi^*(1-\pi^*)} - \frac{1}{1-\pi^*}\right)\right] \\
&= E_{B_i}\left\{\sum_{i=1}^{n} E_{W_i|B_i}\left(\frac{W_i}{\pi^*(1-\pi^*)} - \frac{1}{1-\pi^*}|B_i\right)\right\} \\
&= E_{B_i}\left[\sum_{i=1}^{n}\left(\frac{E_{W_i|B_i}\left(W_i|B_i\right)}{\pi^*(1-\pi^*)} - \frac{1}{1-\pi^*}\right)\right] \\
&= nE_{B_i}\left[\frac{P\left(W_i = 1|B_i\right)}{\pi^*(1-\pi^*)} - \frac{1}{1-\pi^*}\right] \\
&= nE_{B_i}\left[\frac{P\left(S_i < B_i|X_i = 1\right)P(X_i = 1)}{\pi^*(1-\pi^*)} - \frac{1}{1-\pi^*}\right] \\
&= nE_{B_i}\left[\frac{\left(1-\exp(-\lambda B_i)\right)\pi}{\pi^*(1-\pi^*)} - \frac{1}{1-\pi^*}\right] \\
&= n\int_0^{\tau} g_B(u)\left[\frac{\left(1-\exp(-\lambda u)\right)\pi}{\pi^*(1-\pi^*)} - \frac{1}{1-\pi^*}\right]du \\
&\quad + n\int_{\tau}^{\infty} g_B(u)\left[\frac{\left(1-\exp(-\lambda\tau)\right)\pi}{\pi^*(1-\pi^*)} - \frac{1}{1-\pi^*}\right]du \\
&= \frac{\pi}{\pi^*(1-\pi^*)}\left\{\int_0^{\tau} g_B(u)\left[1-\exp(-\lambda u)\right]du + \left[1-\exp(-\lambda\tau)\right]\int_{\tau}^{\infty} g_B(u)du\right\} \\
&\quad - \frac{1}{1-\pi^*}.
\end{aligned}$$

Setting this to 0 and solving for π^* gives

$$\pi^* = \pi \left[1 - (1 - G(\tau)) \exp(-\lambda\tau) - \left(\frac{1}{(1/\gamma_2 + \lambda)\,\gamma_2} \right)^{\gamma_1} H(\tau) \right], \tag{4.3}$$

where $G(\cdot)$ is the c.d.f. of the $GAM(\gamma_1, \gamma_2)$ distribution and $H(\cdot)$ is the c.d.f. of the $GAM(\gamma_1, 1/\left(1/\gamma_2 + \lambda\right))$ distribution.

The parameters γ_1 and γ_2 are associated with the inspection time distribution as in (4.2). Given a specific value for $\phi = \gamma_1\gamma_2^2$, we can calculate $\mu = \gamma_1\gamma_2$ for a certain ρ. Then, these values can be used to calculate the asymptotic bias for a given π. Figure 4.3 illustrates the asymptotic bias, $\pi^* - \pi$, for different values of π with π^* given in (4.3). Based on this plot, the naive estimator for π appears to underestimate the true value of π, which is as expected since treating W_i as the true seroconversion status will incorrectly classify the response as zero for those who did not develop antibodies before their assessment time. The magnitude of this bias appears to increase with the true underlying value of π. This is not surprising since the bias is proportional to π in (4.3). The bias appears to decrease in severity as $\rho = P(S_i < B_i | X_i = 1)$ increases. This is reasonable since the higher ρ, the more likely seroconversion is to occur prior to assessment. Therefore, as ρ increases, a larger number of responses will be correctly classified, leading to smaller asymptotic bias in the naive estimator for π. Interestingly, it can be shown that the expression in the square brackets in (4.3) is simply $\rho = P(S_i < B_i | X_i = 1)$ in this case (see (4.11)).

The asymptotic bias in the estimator for a covariate effect on the probability of seroconversion can be derived in a similar way. Suppose the true underlying model is given as in Section 4.3 and a logistic regression model is assumed for π^* such that

$$\text{logit}(\pi^*(Z_i)) = \psi_0^* + \psi_1^* Z_i.$$

Noting that

$$\pi^*(Z_i = 0) = \exp(\psi_0^*)/\left(1 + \exp(\psi_0^*)\right)$$

and

$$\pi^*(Z_i = 1) = \exp(\psi_0^* + \psi_1^*)/\left(1 + \exp(\psi_0^* + \psi_1^*)\right),$$

based on (4.3), we conclude that

$$\psi_1^* = \log\left\{\frac{\frac{\pi(\psi|Z_i=1)}{1-\pi(\psi|Z_i=1)\left[1-(1-G(\tau))\exp(-\lambda\tau)-\left(\frac{1}{(1/\gamma_2+\lambda)\gamma_2}\right)^{\gamma_1}H(\tau)\right]}}{\frac{\pi(\psi|Z_i=0)}{1-\pi(\psi|Z_i=0)\left[1-(1-G(\tau))\exp(-\lambda\tau)-\left(\frac{1}{(1/\gamma_2+\lambda)\gamma_2}\right)^{\gamma_1}H(\tau)\right]}}\right\} \tag{4.4}$$

Figure 4.4 illustrates the asymptotic bias in the naive estimator for the covariate effect given by $\psi^* - \psi$ for different values for ψ with ψ^* according to (4.4). As was the case in (4.3), the expressions in the square brackets in (4.4) is $\rho = P(S_i < B_i|X_i = 1)$. Again, the magnitude of the asymptotic bias increases as the true underlying value of ψ_1 increases. The direction of this asymptotic bias depends on the sign of the true covariate effect. However, in both cases, the naive estimator underestimates the magnitude of the true covariate effect. This provides compelling evidence of the need to fit models such as the proposed latent class current status model if there is good scientific rationale for such a formulation. There is no asymptotic bias present when the true underlying covariate effect is zero. Clearly, ignoring the seroconversion time distribution in a naive analysis can lead to substantial asymptotic bias in the estimators associated with the seroconversion probability, especially when a small proportion of individuals in the susceptible sub-population develop antibodies before their assessment times.

Figure 4.3: *Asymptotic bias in the naive estimator for the probability of experiencing the event* ($\rho = P(S_i < B_i | X_i = 1)$ *and* $\phi = 0.5$).

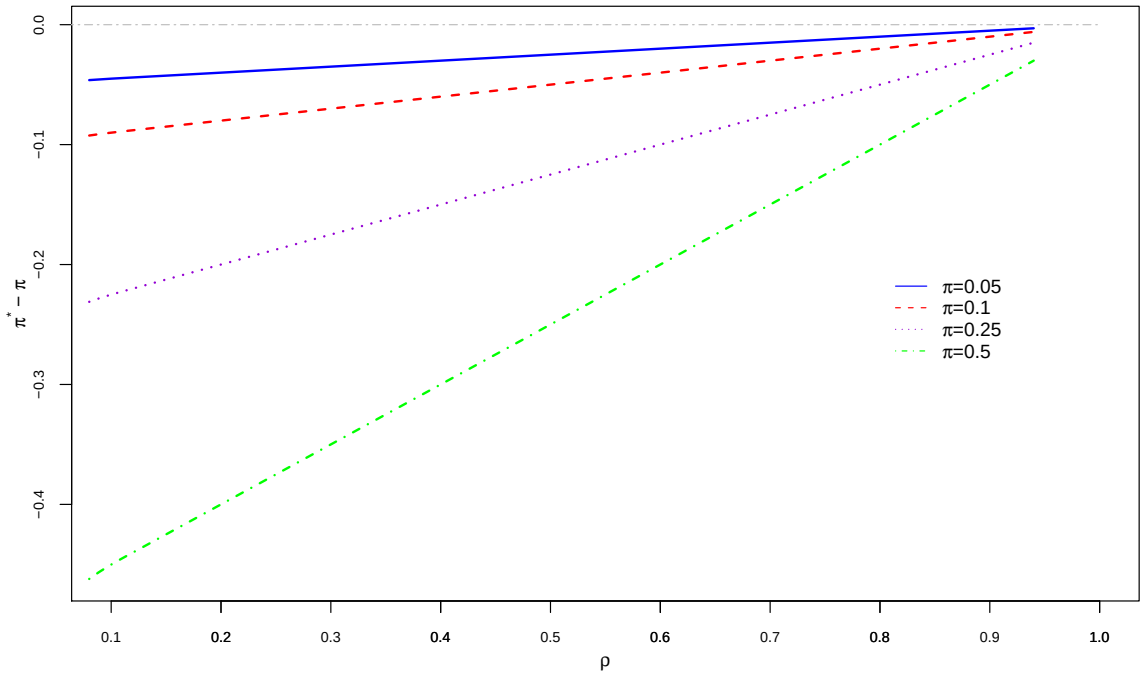
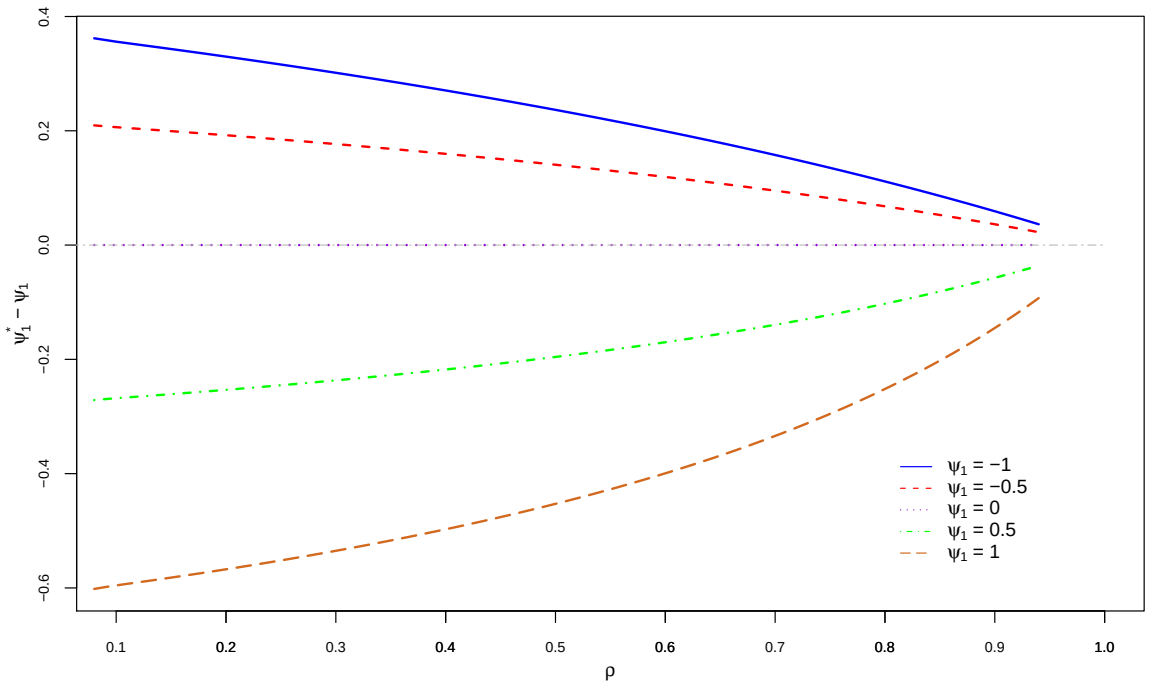


Figure 4.4: *Asymptotic bias in the naive estimator for a covariate effect on the probability of experiencing the event* ($\rho = P(S_i < B_i | X_i = 1)$ *and* $\phi = 0.5$).



4.3.2 Likelihood with a Non-susceptible Fraction

In the absence of covariates the observed data for individual i is (b_i, w_i). The variable X_i is called a *latent variable* because it is unobserved for many individuals. If $W_i = 1$, then by the definition of X_i, we know that it must be 1 but for those with $W_i = 0$, the true value of X_i is unknown. To proceed with the likelihood approach, the following probability expressions are required. For the purposes of this formulation, we consider the one sample problem, but note that extensions to deal with covariates are straightforward.

A likelihood contribution from a subject testing positive is proportional to

$$\begin{aligned} P(W_i = 1|B_i = b_i) &= P(S \leq b_i|X_i = 1)P(X_i = 1) + 0 \times P(X_i = 0) \\ &= (1 - \mathcal{F}_S(b_i)) \times \pi + 0 \times (1 - \pi), \end{aligned}$$

but for an individual testing negative it is

$$\begin{aligned} P(W_i = 0|B_i = b_i) &= P(S > b_i|X_i = 1)P(X_i = 1) + 1 \times P(X_i = 0) \\ &= \mathcal{F}_S(b_i) \times \pi + 1 \times (1 - \pi). \end{aligned}$$

Assuming the inspection times (times of blood test) are uninformative, the likelihood function can be constructed based on these probabilities alone and is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[(1 - \mathcal{F}_S(b_i))\,\pi\right]^{w_i} \left[\mathcal{F}_S(b_i)\pi + (1 - \pi)\right]^{1-w_i}. \tag{4.5}$$

A "complete data" likelihood function can be constructed by including x_i in the data, so if we observe (x_i, b_i, w_i), for $i = 1, 2, ..., n$, we write

$$\mathcal{L}_c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[(1 - \mathcal{F}_S(b_i))\pi\right]^{x_i w_i} \left\{\left[\mathcal{F}_S(b_i)\pi\right]^{x_i} \left[1 - \pi\right]^{1-x_i}\right\}^{1-w_i}, \tag{4.6}$$

and if $l_c = \log \mathcal{L}_c$,

$$\begin{aligned} l_c(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \{x_i w_i \left[\log\left(1 - \mathcal{F}_S(b_i)\right) + \log \pi\right] \\ &\quad + x_i(1 - w_i)\left[\log \mathcal{F}_S(b_i) + \log \pi\right] + (1 - x_i)\log\left(1 - \pi\right)\}. \end{aligned}$$

Since this involves "missing data" (the x_i's here), the natural approach is to apply the EM Algorithm (Dempster et al. 1977).

4.3.3 An EM Algorithm for Missing X_i

To illustrate how the EM algorithm can be applied to obtain maximum likelihood estimates for this problem, we consider the one sample problem and let $\boldsymbol{\theta} = \{\pi, \mathcal{F}_S(\cdot)\}$. At the r^{th} iteration we denote the estimate of $\boldsymbol{\theta}$ obtained by maximization, $\hat{\boldsymbol{\theta}}^{(r)}$, and write $\mathcal{F}_S(s; \hat{\boldsymbol{\theta}}^{(r)})$ as $\hat{\mathcal{F}}_S^{(r)}$. Then the EM algorithm proceeds as follows:

1. *Expectation Step (E-Step)*
Because the complete data log-likelihood function is linear in X_i we can write $Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\boldsymbol{r-1})}\right) \doteq l_c(\boldsymbol{\theta})|_{x_i = x_i^{(r)}}$,

$$\begin{aligned} x_i^{(r)} &\doteq E\left(X_i | B_i = b_i, W_i = w_i; \hat{\boldsymbol{\theta}}^{(r-1)}\right) \\ &= P\left(X_i = 1 | B_i = b_i, W_i = w_i; \hat{\boldsymbol{\theta}}^{(r-1)}\right) \\ &= w_i \times P\left(X_i = 1 | B_i = b_i, W_i = 1; \hat{\boldsymbol{\theta}}^{(r-1)}\right) + (1 - w_i) \times P\left(X_i = 1 | B_i = b_i, W_i = 0; \hat{\boldsymbol{\theta}}^{(r-1)}\right) \\ &= w_i + (1 - w_i) \times \frac{\hat{\pi}^{(r-1)} \hat{\mathcal{F}}_S^{(r-1)}(b_i)}{\hat{\pi}^{(r-1)} \hat{\mathcal{F}}_S^{(r-1)}(b_i) + (1 - \hat{\pi}^{(r-1)})}. \end{aligned} \tag{4.7}$$

2. *Maximization Step (M-Step)*
Obtain $\hat{\boldsymbol{\theta}}^{(\boldsymbol{r})}$ through maximization of $Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\boldsymbol{r-1})}\right)$ with respect to $\boldsymbol{\theta}$ for $r = 1, 2, ...$

Steps 1 and 2 are repeated until convergence is reached (i.e. when the difference between successive parameter estimates drops below a specified tolerance).

Fortunately, the function we need to maximize in Step 2, $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r-1)}\right) = l_{c_1}(\pi)|_{x_i = x_i^{(r)}} + l_{c_2}(\mathcal{F}_S(\cdot))|_{x_i = x_i^{(r)}}$, breaks down into two familiar problems. We need to maximize $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r-1)}\right)$ with respect to θ, where

$$l_{c_1}(\pi)|_{x_i = x_i^{(r)}} = \sum_{i=1}^{n} \left\{ x_i^{(r)} \log \pi + (1 - x_i^{(r)}) \log(1 - \pi) \right\}, \tag{4.8}$$

and

$$l_{c_2}(\mathcal{F}_S(\cdot))|_{x_i=x_i^{(r)}} = \sum_{i=1}^{n} x_i^{(r)} \left\{ w_i \log(1 - \mathcal{F}_S(b_i)) + (1 - w_i) \log \mathcal{F}_S(b_i) \right\}. \tag{4.9}$$

The expression (4.8) is simply the familiar binomial log-likelihood function with working response $x_i^{(r)}$, for which there are many software packages available. The specific form of (4.9) depends on the model assumed for the seroconversion time distribution for the sub-population of susceptible patients. If we assume a Weibull model for the seroconversion time distribution (i.e. $S \sim WEI(\lambda, \kappa)$), the survivor function would be $\mathcal{F}_S(b) = \exp\left[-(\lambda b)^{\kappa}\right]$. Since

$$\log\left(-\log\left(\mathcal{F}_S(b_i)\right)\right) = \kappa \log \lambda + \kappa \log b_i, \tag{4.10}$$

if we let $v_i = 1 - w_i$, then (4.9) is simply

$$l_{c_2}(\mathcal{F}_S(b_i; \lambda, \kappa))|_{x_i=x_i^{(r)}} = \sum_{i=1}^{n} x_i^{(r)} \left\{ v_i \log(\mathcal{F}_S(b_i)) + (1 - v_i) \log(1 - \mathcal{F}_S(b_i)) \right\}.$$

This is also a binomial log-likelihood function with weights given by $x_i^{(r)}$, $i = 1, 2, ..., n$. To incorporate covariates, we could assume a proportional hazards Weibull regression model, in which case, the linear predictor would be added to the expression given in (4.10). Existing software could then be used to fit a binary regression model for W_i with the complementary log-log link and weights given by $x_i^{(r)}$. In addition to any covariates that appear in the assumed proportional hazards model, a supplementary covariate given by $\log b_i$ should be included when fitting this model. When κ is assumed to be one (i.e. $S \sim EXP(\lambda)$), $\log b_i$ should be treated as an offset rather than a covariate.

4.3.4 Relative Efficiency

We now consider the precision of the estimators based on latent class models with current status data and explore the factors that influence this precision. This is important to help identify settings where it is and is not sensible to consider models for a non-susceptible fraction with current status data. To carry out this investigation we derive expressions for the relative efficiency of the estimators based on the Fisher information matrix. To obtain

the Fisher information, the expectation of $S(\theta)S'(\theta)$ with respect to (W, B) is required. We consider the one sample problem (i.e. no covariates) and assume $P(X_i = 1) = \pi$ and that the seroconversion time follows an $EXP(\lambda)$ distribution with $P(S_i > s|X_i = 1; \lambda) = \exp(-\lambda s)$. With $\boldsymbol{\theta} = (\pi, \lambda)'$, the i^{th} individual's contribution to the observed data log-likelihood function (suppressing the subscript i) is

$$l(\boldsymbol{\theta}) = w \log\left(1 - \exp(-\lambda b)\right) + w \log \pi + (1 - w) \log\left[\exp(-\lambda b)\pi + (1 - \pi)\right].$$

Based on the reparametrization $\theta_1 = \text{logit}(\pi)$ and $\theta_2 = \log \lambda$ to avoid parameter constraints, the score function is $\left(S_1(\boldsymbol{\theta}), S_2(\boldsymbol{\theta})\right)'$, where

$$S_1(\boldsymbol{\theta}) = wa\left(\boldsymbol{\theta}, b\right) + b\left(\boldsymbol{\theta}, b\right)$$

and

$$S_2(\boldsymbol{\theta}) = wg\left(\boldsymbol{\theta}, b\right) + h\left(\boldsymbol{\theta}, b\right),$$

with

$$\begin{aligned}
a\left(\boldsymbol{\theta}, b\right) &= (1 - \pi) + \frac{\pi(1 - \pi)(1 - \exp(-\lambda b))}{\exp(-\lambda b)\pi + (1 - \pi)} \\
b\left(\boldsymbol{\theta}, b\right) &= \frac{\pi(1 - \pi)(\exp(-\lambda b) - 1)}{\exp(-\lambda b)\pi + (1 - \pi)} \\
g\left(\boldsymbol{\theta}, b\right) &= b\lambda \exp(-\lambda b)\left[\frac{1}{1 - \exp(-\lambda b)} + \frac{\pi}{\exp(-\lambda b)\pi + (1 - \pi)}\right] \\
h\left(\boldsymbol{\theta}, b\right) &= -\frac{b\pi\lambda \exp(-\lambda b)}{\exp(-\lambda b)\pi + (1 - \pi)}.
\end{aligned}$$

To construct the $S(\theta)S'(\theta)$ matrix, expressions for $S_1^2(\boldsymbol{\theta})$, $S_2^2(\boldsymbol{\theta})$, and $S_1(\boldsymbol{\theta})S_2(\boldsymbol{\theta})$ are required:

$$\begin{aligned}
S_1^2(\boldsymbol{\theta}) &= w^2 a^2\left(\boldsymbol{\theta}, b\right) + 2wa\left(\boldsymbol{\theta}, b\right) b\left(\boldsymbol{\theta}, b\right) + b^2\left(\boldsymbol{\theta}, b\right) \\
S_2^2(\boldsymbol{\theta}) &= w^2 g^2\left(\boldsymbol{\theta}, b\right) + 2wg\left(\boldsymbol{\theta}, b\right) h\left(\boldsymbol{\theta}, b\right) + h^2\left(\boldsymbol{\theta}, b\right) \\
S_1(\boldsymbol{\theta})S_2(\boldsymbol{\theta}) &= w^2 a\left(\boldsymbol{\theta}, b\right) g\left(\boldsymbol{\theta}, b\right) + w\left[a\left(\boldsymbol{\theta}, b\right) h\left(\boldsymbol{\theta}, b\right) + b\left(\boldsymbol{\theta}, b\right) g\left(\boldsymbol{\theta}, b\right)\right] + b\left(\boldsymbol{\theta}, b\right) h\left(\boldsymbol{\theta}, b\right).
\end{aligned}$$

Then, the Fisher information matrix, $\mathcal{I}(\boldsymbol{\theta})$, is the expectation with respect to W, B of $S(\theta)S'(\theta)$ or $\mathcal{I}(\boldsymbol{\theta}) = E_B\left\{E_{W|B}\left[S(\boldsymbol{\theta})S'(\boldsymbol{\theta})|B\right]\right\}$. It has the following entries.

$$\begin{aligned}
\mathcal{I}(\boldsymbol{\theta})_{[1,1]} &= \int_0^\tau g_B(u)\left\{\pi\left[1-\exp(-\lambda u)\right]\left[a^2\left(\boldsymbol{\theta},u\right)+2a\left(\boldsymbol{\theta},u\right)b\left(\boldsymbol{\theta},u\right)\right]+b^2\left(\boldsymbol{\theta},u\right)\right\}du \\
&+ \left[1-G(\tau)\right]\left\{\pi\left[1-\exp(-\lambda\tau)\right]\left[a^2\left(\boldsymbol{\theta},\tau\right)+2a\left(\boldsymbol{\theta},\tau\right)b\left(\boldsymbol{\theta},\tau\right)\right]+b^2\left(\boldsymbol{\theta},\tau\right)\right\} \\
\mathcal{I}(\boldsymbol{\theta})_{[2,2]} &= \int_0^\tau g_B(u)\left\{\pi\left[1-\exp(-\lambda u)\right]\left[g^2\left(\boldsymbol{\theta},u\right)+2g\left(\boldsymbol{\theta},u\right)h\left(\boldsymbol{\theta},u\right)\right]+h^2\left(\boldsymbol{\theta},u\right)\right\}du \\
&+ \left[1-G(\tau)\right]\left\{\pi\left[1-\exp(-\lambda\tau)\right]\left[g^2\left(\boldsymbol{\theta},\tau\right)+2g\left(\boldsymbol{\theta},\tau\right)h\left(\boldsymbol{\theta},\tau\right)\right]+h^2\left(\boldsymbol{\theta},\tau\right)\right\} \\
\mathcal{I}(\boldsymbol{\theta})_{[1,2]} &= \mathcal{I}(\boldsymbol{\theta})_{[2,1]} \\
&= \int_0^\tau g_B(u)\left\{\pi\left[1-\exp(-\lambda u)\right]\left[a\left(\boldsymbol{\theta},u\right)g\left(\boldsymbol{\theta},u\right)+a\left(\boldsymbol{\theta},u\right)h\left(\boldsymbol{\theta},u\right)+b\left(\boldsymbol{\theta},u\right)g\left(\boldsymbol{\theta},u\right)\right]+b\left(\boldsymbol{\theta},u\right)h\left(\boldsymbol{\theta},u\right)\right\}du \\
&+ \left[1-G(\tau)\right]\left\{\pi\left[1-\exp(-\lambda\tau)\right]\left[a\left(\boldsymbol{\theta},\tau\right)g\left(\boldsymbol{\theta},\tau\right)+a\left(\boldsymbol{\theta},\tau\right)h\left(\boldsymbol{\theta},\tau\right)+b\left(\boldsymbol{\theta},\tau\right)g\left(\boldsymbol{\theta},\tau\right)\right]+b\left(\boldsymbol{\theta},\tau\right)h\left(\boldsymbol{\theta},\tau\right)\right\},
\end{aligned}$$

where $G(\cdot)$ is the c.d.f. of a $GAM(\gamma_1, \gamma_2)$ random variable. The parameters associated with this assessment time distribution depend on $\rho = P(S < B|X = 1; \lambda, \mu, \phi)$ which is the probability of testing positive for the sub-population of individuals who will develop antibodies at some point. Consider the maximum observation time $\tau = 1$, and let λ be the solution to $1 - \mathcal{F}_S(\tau; \lambda) = 0.95$ which is $\lambda = -\log 0.05$. This ensures that 95% of the susceptible sample would be expected to seroconvert over the course of the study. In addition, let $B^* \sim GAM(\gamma_1, \gamma_2)$ with mean $\mu = \gamma_1\gamma_2$ and variance $\phi = \gamma_1\gamma_2^2$. Since the inspection times are $B = \min(B^*, 1)$, ρ is:

$$\rho = \int_0^\tau \left[1-\exp(-\lambda u)\right] g(u;\mu,\phi)du + \left[1-\exp(-\lambda\tau)\right]\left[1-G(\tau;\mu,\phi)\right]. \tag{4.11}$$

Based on specified values of ϕ and ρ, (4.11) can be solved for μ using numerical integration. Then these parameter values can be used to evaluate the Fisher information which, when inverted, provides asymptotic variances to be used to calculate asymptotic relative efficiencies. The parameter values given by $\boldsymbol{\nu} = (\pi, \rho, \phi)'$ characterize a specific configuration. If we denote $\boldsymbol{\nu}_0 = (\pi_0, \rho_0, \phi_0)'$ as the reference parameter configuration, then the asymptotic relative efficiency of the estimator for different values of $\boldsymbol{\nu} = (\pi, \rho, \phi)'$ compared to the estimator with $\boldsymbol{\nu}_0 = (\pi_0, \rho_0, \phi_0)'$ is the ratio of asymptotic variances given by

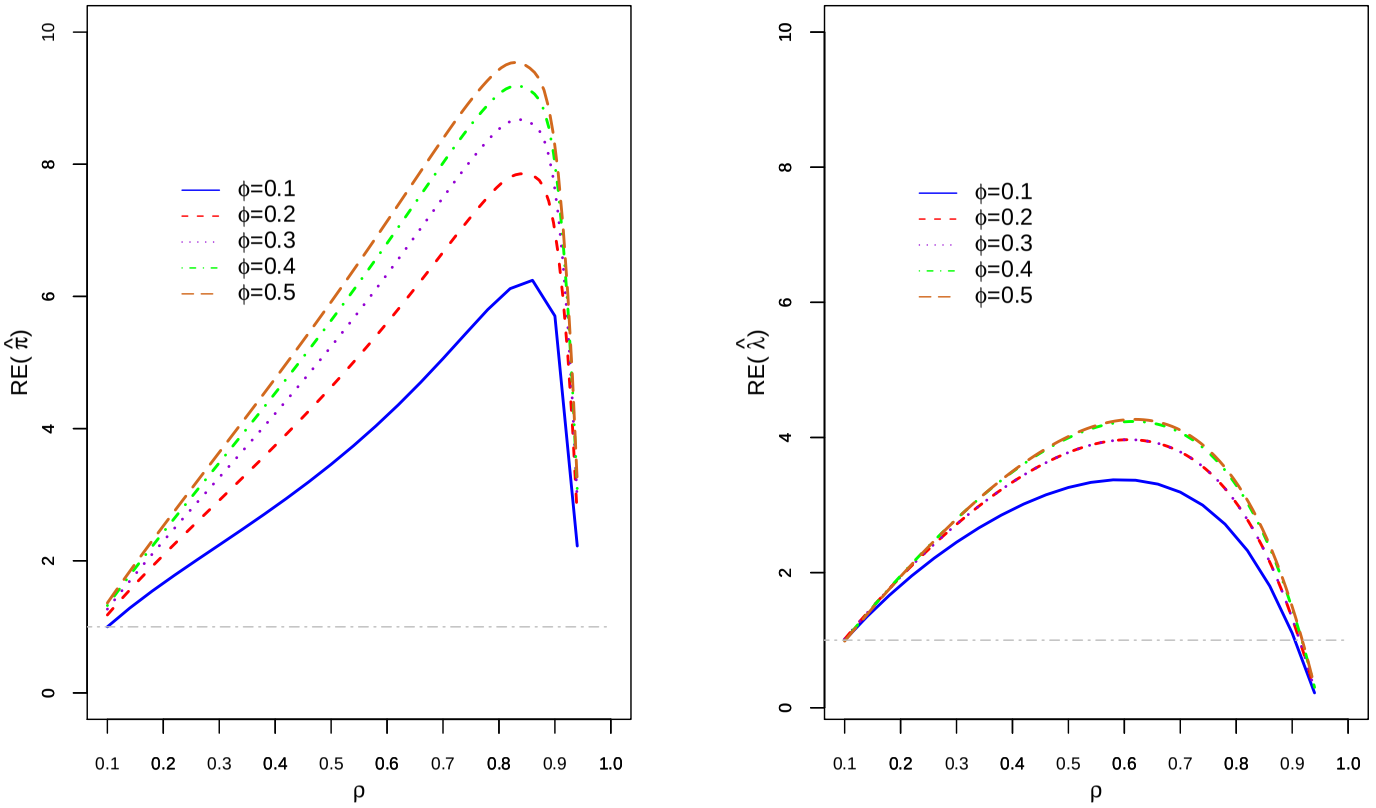
$$R.E.(\hat{\pi}) = \frac{asvar\left(\sqrt{n}(\hat{\pi}-\pi);\boldsymbol{\nu}_0\right)}{asvar\left(\sqrt{n}(\hat{\pi}-\pi);\boldsymbol{\nu}\right)}$$

and

$$R.E.(\hat{\lambda}) = \frac{asvar\left(\sqrt{n}(\hat{\lambda}-\lambda);\boldsymbol{\nu}_0\right)}{asvar\left(\sqrt{n}(\hat{\lambda}-\lambda);\boldsymbol{\nu}\right)}.$$

Figures 4.5 and 4.6 display asymptotic relative efficiencies of estimators for π and λ based on the latent class current status model. In both plots, the reference parameter configuration is $\boldsymbol{\nu}_0 = (0.1, 0.1, 0.1)'$. Figure 4.5 compares relative efficiencies for different values of the variance of the inspection time distribution, ϕ; whereas, Figure 4.6 compares these values for different values of π. Based on the plots, it appears that estimators are least efficient when the probability of testing positive is extreme (i.e. either very low or very high) for the susceptible sub-population. It seems that the true underlying value for π has more of an impact on the R.E. than the value for ϕ, although this may not be the case for $\phi > 0.5$. There appears to be considerable increases in sensitivity of R.E. to ρ as π is increased. Figure 4.5 suggests that imposing variation in the inspection times will increase efficiency, while 4.6 suggests that the most efficient estimators are obtained when the inspection times are distributed such that a moderate proportion of susceptible individuals are observed to test positive.

Figure 4.5: *Relative efficiency of the estimators for* π *and* λ *for different values of* ϕ *(*$\boldsymbol{\nu}_0 = (\pi, \rho, \phi) = (0.1, 0.1, 0.1)$*; with* $n = 3150$*,* $asvar(\hat{\pi}; \boldsymbol{\nu}_0) = 0.001253$ *and* $asvar(\hat{\lambda}; \boldsymbol{\nu}_0) = 3.408129$*).*



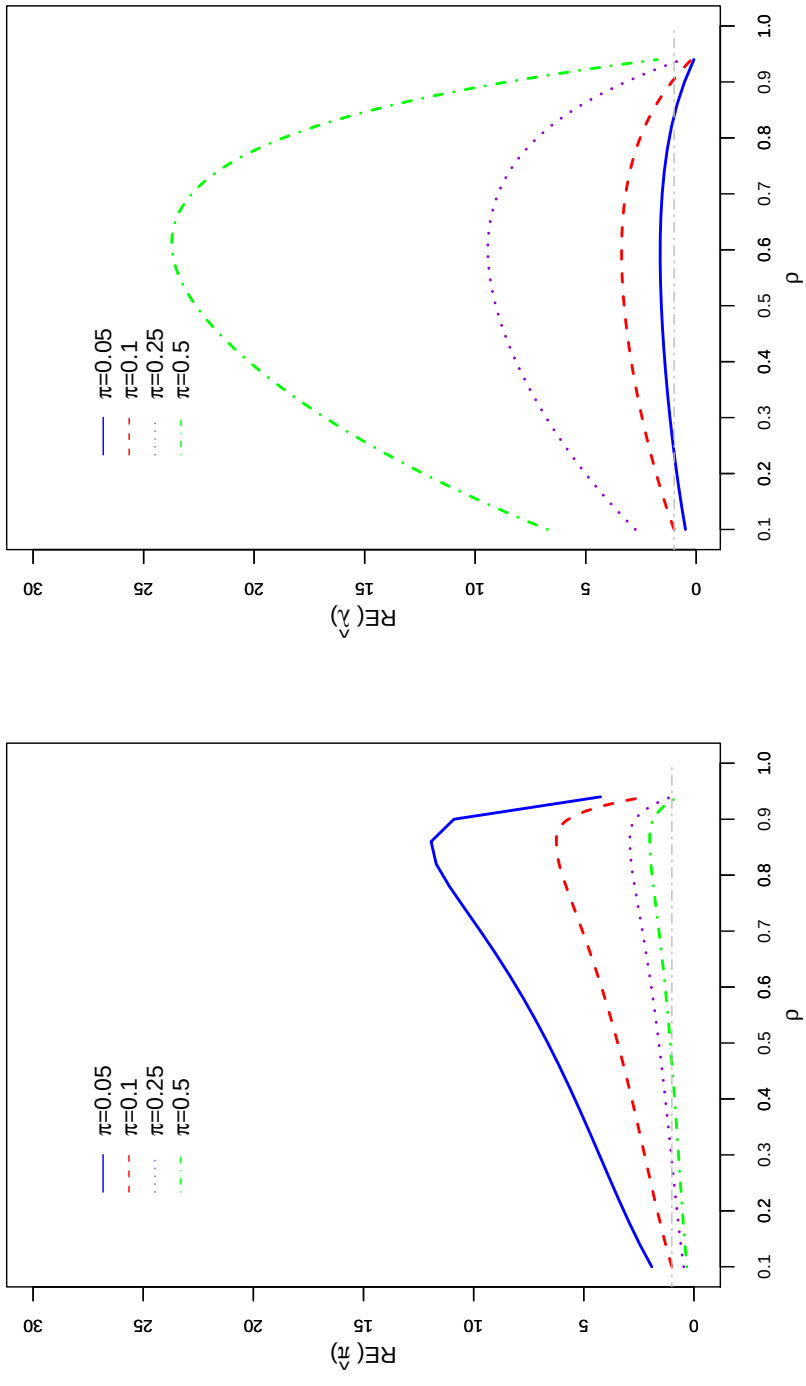


Figure 4.6: *Relative efficiency of the estimators for* π *and* λ *for different values of* π $(\boldsymbol{\nu}_0 = (\pi, \rho, \phi) = (0.1, 0.1, 0.1)$*; with* $n = 3150$*,* $asvar(\hat{\pi}; \boldsymbol{\nu}_0) = 0.001253$ *and* $asvar(\hat{\lambda}; \boldsymbol{\nu}_0) = 3.408129)$.

4.3.5 Piecewise Constant Hazards Models

An alternative to adopting a standard parametric model for the seroconversion time is to assume a flexible piecewise constant proportional hazards model. Although we will show that the expectation step of the EM algorithm is a little more complicated than it was before, unlike typical parametric models, this type of model does not require strong assumptions regarding the underlying distribution. Moreover there is greater flexibility in the degree of robustness of the model: the greater the number of pieces, the more robust the method. The complete data in this case is given by (S_i, X_i), $i = 1, 2, ..., n$. Let $a_0 = 0 < a_1 < a_2 < ... < a_K = \infty$ denote the cut-points and suppose there are K pieces to the hazard function so that

$$\lambda_S(s) = \lambda_k, \text{ if } s \in (a_{k-1}, a_k], \tag{4.12}$$

for $k = 1, 2, ..., K$. The cumulative hazard function $\Lambda_S(s)$ is then

$$\Lambda_S(s) = \sum_{k=1}^{K} c_k(s)\lambda_k,$$

where

$$c_k(s) = \max\left(0, \min\left(s - a_{k-1}, a_k - a_{k-1}\right)\right),$$

and we may write $\mathcal{F}_S(s; \lambda) = \exp(-\Lambda_S(s))$. The complete data log-likelihood function is then given by

$$l_c\left(\theta\right) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K} E\left[X_i I\left(S_i \in (a_{k-1}, a_k]\right)\right] \log \lambda_k - \sum_{k=1}^{K} E\left[X_i c_k(S_i)\right] \lambda_k \right\}. \tag{4.13}$$

Based on this log-likelihood function, the EM algorithm proceeds as outlined below.

1. *Expectation Step (E-Step)*

$$
\begin{aligned}
Q\left(\theta;\hat{\boldsymbol{\theta}}^{(r-1)}\right) &\doteq E\left\{\sum_{i=1}^{n} X_i \sum_{k=1}^{K}\left[I\left(S_i \in (a_{k-1}, a_k]\right)\log\lambda_k - c_k\left(S_i\right)\lambda_k\right]|W_i, B_i;\boldsymbol{\theta}\right\} \\
&= E\left\{E\left[\sum_{i=1}^{n} X_i \sum_{k=1}^{K}\left[I\left(S_i \in (a_{k-1}, a_k]\right)\log\lambda_k - c_k\left(S_i\right)\lambda_k\right]|W_i, B_i, X_i;\boldsymbol{\theta}\right]|W_i, B_i\right\} \\
&= E\left\{\sum_{i=1}^{n} X_i \left[\sum_{k=1}^{K} E\left(I\left(S_i \in (a_{k-1}, a_k]\right)|W_i, B_i, X_i\right)\log\lambda_k\right.\right. \\
&\quad \left.\left. - E\left(c_k\left(S_i\right)|W_i, B_i, X_i\right)\lambda_k|W_i, B_i;\boldsymbol{\theta}\right]\right\} \\
&= \sum_{i=1}^{n} x_i^{(r)}\left[\sum_{k=1}^{K} E\left(I\left(S_i \in (a_{k-1}, a_k]\right)|W_i, B_i, X_i = 1;\lambda^{(r-1)}\right)\log\lambda_k\right. \\
&\quad \left. - E\left(c_k\left(S_i\right)|W_i, B_i, X_i = 1;\lambda^{(r-1)}\right)\lambda_k\right].
\end{aligned}
$$

When $W_i = 1$ and $B_i > a_{k-1}$

$$
\begin{aligned}
E\left(c_k\left(S_i\right)|W_i = 1, B_i, X_i = 1;\Lambda^{(r-1)}\right) &= E\left(\int_{a_{k-1}}^{\min(B_i,a_k)} I\left(S_i > u\right)du|W_i = 1, X_i = 1, B_i\right) \\
&= \int_{a_{k-1}}^{\min(B_i,a_k)} P\left(S_i > u|W_i = 1, X_i = 1, B_i\right)du \\
&= \int_{a_{k-1}}^{\min(B_i,a_k)} \frac{P\left(u < S_i < B_i|X_i = 1, B_i\right)}{P\left(S_i < C_i|X_i = 1, B_i\right)}du \\
&= \int_{a_{k-1}}^{\min(B_i,a_k)} \frac{\mathcal{F}_S(u) - \mathcal{F}_S(B_i)}{1 - \mathcal{F}_S(B_i)}du \\
&= \min(B_i, a_k) - a_{k-1} - \int_{a_{k-1}}^{\min(B_i,a_k)} \frac{1 - \exp(-\Lambda_S^{(r-1)}(u))}{1 - \exp(-\Lambda_S^{(r-1)}(B_i))}du,
\end{aligned}
$$

and

$$\begin{aligned} & E\left[I\left(S_i \in (a_{k-1}, a_k]; \Lambda^{(r-1)}\right) | W_i = 1, B_i, X_i = 1\right] \\ = \; & E\left[I\left(a_{k-1} < S_i < a_k\right) | W_i = 1, X_i = 1, B_i\right] \\ = \; & \frac{P\left(a_{k-1} < S_i < \min(B_i, a_k) | X_i = 1, B_i\right)}{P\left(S_i < \min(B_i, a_k) | X_i = 1, B_i\right)} \\ = \; & \frac{\mathcal{F}_S(a_{k-1}) - \mathcal{F}_S(\min(B_i, a_k))}{1 - \mathcal{F}_S(\min(B_i, a_k))} \\ = \; & \frac{\exp(-\Lambda_S^{(r-1)}(a_{k-1})) - \exp(-\Lambda_S^{(r-1)}(\min(B_i, a_k)))}{1 - \exp(-\Lambda_S^{(r-1)}(\min(B_i, a_k)))}. \end{aligned}$$

If $B_i < a_{k-1}$ then the inspection time occurred prior to the lower endpoint of the interval $(a_{k-1}, a_k]$. Since $W_i = 1$, seroconversion was observed to occur prior to a_{k-1} so

$$E\left(c_k\left(S_i\right) | W_i = 1, B_i, X_i = 1; \Lambda^{(r-1)}\right) = E\left[I\left(S_i \in (a_{k-1}, a_k]\right) | W_i = 1, B_i, X_i = 1; \Lambda^{(r-1)}\right] = 0.$$

Based on similar steps, when $W_i = 0$ and $B_i \geq a_k$ then

$$E\left(c_k\left(S_i\right) | W_i = 0, B_i, X_i = 1; \Lambda^{(r-1)}\right) = \int_{\max(B_i, a_{k-1})}^{a_k} \frac{\exp(-\Lambda_S^{(r-1)}(u))}{\exp(-\Lambda_S^{(r-1)}(B_i))} du$$

and

$$E\left[I\left(S_i \in (a_{k-1}, a_k]\right) | W_i = 0, B_i, X_i = 1; \Lambda^{(r-1)}\right] = \frac{\exp(-\Lambda_S^{(r-1)}(\max(B_i, a_{k-1}))) - \exp(-\Lambda_S^{(r-1)}(a_k))}{\exp(-\Lambda_S^{(r-1)}(B_i))}.$$

If $B_i > a_k$, then the inspection time (and the seroconversion time since $W_i = 0$ and $X_i = 1$) occurred after the upper endpoint of the interval so individual i was at risk for the entire interval resulting in

$$E\left(c_k\left(S_i\right) | W_i = 0, B_i, X_i = 1; \Lambda^{(r-1)}\right) = a_k - a_{k-1}$$

and

$$E\left[I\left(S_i \in (a_{k-1}, a_k]\right) | W_i = 0, B_i, X_i = 1; \Lambda^{(r-1)}\right] = 0.$$

In addition, $x_i^{(r)}$ is as given in (4.7).

2. *Maximization Step (M-Step)*
Obtain $\hat{\boldsymbol{\theta}}^{(\boldsymbol{r})}$ through maximization of $Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\boldsymbol{r-1})}\right)$ with respect to $\boldsymbol{\theta}$ for $r = 1, 2, ...$ This can be achieved using ordinary software for fitting exponential regression models; (4.13) has the form of a sum of contributions to a log-likelihood for a series of exponential models. Covariates can be introduced to indicate the "piece" of the piecewise constant hazard function.

Steps 1 and 2 are repeated until convergence is reached (i.e. when the difference between successive parameter estimates drops below a specified tolerance).

4.3.6 EM with Nonparametric Estimation of $\mathcal{F}_S(\cdot)$

The term l_{c_2} in (4.9) appears to be a weighted version of (1.33). It is reasonable then, that a modified version of PAVA could be applied to find the nonparametric maximum likelihood estimate of $F_S(\cdot)$, and therefore $\mathcal{F}_S(\cdot)$, in l_{c_2} at each iteration. There may be some identifiability issues when estimating $F_S(\cdot)$ due to the fact that only a proportion of the population will experience the event of interest. Tail adjustments may be required in such settings to ensure identifiability (Farewell 1977; Taylor 1995). The idea, however, would work basically as follows.

A nonparametric estimate of $\mathcal{F}_S(\cdot)$ can be obtained by noting that (4.9) is like a weighted version of (1.33), so optimizing (4.9) may be carried out by adapting the usual isotonic regression approach (Sun 2006). Similar to the steps outlined in Section 1.4, let $B_{(1)} < B_{(2)} < ... < B_{(J)}$ denote the J unique ordered inspection times and let $r_j = \sum_{i=1}^{n} I(B_i = B_{(j)})W_i$ be the number of individuals with inspection time $B_{(j)}$ who test positive. Individuals testing positive are known to be seroconverters but those testing negative will have $x_i^{(r)} < 1$. The "effective number at risk" at the j^{th} inspection time is then $\widehat{\mu}_j^{(r)} = \sum_{i=1}^{n} I(B_i = B_{(j)})(W_i + (1 - W_i)x_i^{(r)})$ and so (4.9) can be optimized by the isotonic regression of $(r_1/\widehat{\mu}_{(1)}^{(r)}, \ldots, r_J/\widehat{\mu}_{(J)}^{(r)})'$ with weights $(\widehat{\mu}_{(1)}^{(r)}, \ldots, \widehat{\mu}_{(J)}^{(r)})'$ to give

$$\widehat{F}^{(r)}(B_{(j)}) = \max_{u \leq j} \min_{v \geq j} \left(\frac{\sum_{\ell=u}^{v} r_\ell}{\sum_{\ell=u}^{v} \widehat{\mu}_{(\ell)}^{(r)}} \right) \tag{4.14}$$

To ensure identifiability in the nonparametric setting, as in the case with right-censored data (Taylor 1995), it is necessary to force $\widehat{F}_S(\cdot)$ to increase to one at some point, Υ. This can be achieved by putting a point mass at Υ so that $\widehat{F}_S(\Upsilon) = 1$. The literature on immunological response following exposure to low molecular weight heparin suggests that this occurs within 10 days of exposure.

Likelihood ratio statistics can be used to carry out tests of significance of covariate effects in the binary response model for X_i. Let ψ_j denote the coefficient of Z_{ij} in this model. Profile maximum likelihood estimates for $\mathcal{F}_S(s)$ and $\boldsymbol{\psi}$ are can be obtained by carrying out a slightly modified EM algorithm. If ψ_{j0} is a particular value of ψ_j, let $\mathcal{F}_S^{\psi_{j0}}(s)$ and $\psi^{\psi_{j0}}$ denote the maximum likelihood estimates when ψ_j is constrained to equal ψ_{j0}; these are obtained by treating $\psi_{j0}Z_{ij}$ as an offset in the maximization of a version of (4.8) that incorporates covariates (i.e. with π replaced with $\pi(\boldsymbol{\psi})$). The profile likelihood ratio pivotal is

$$LRS(\psi_{j0}) = -2\log\left(\frac{L(\widehat{\mathcal{F}}_S^{\psi_{j0}}(\cdot), \widehat{\psi^{\psi_{j0}}})}{L(\widehat{\mathcal{F}}_S(\cdot), \widehat{\psi})}\right) \tag{4.15}$$

so the p-value for testing $H_0 : \psi_j = \psi_{j0}$ versus $H_A : \psi_j \neq \psi_{j0}$ is $P(\chi_1^2 > LRS(\psi_{j0}))$. Similarly, a 95% confidence interval for ψ_j is defined as $\{\psi_j : LRS(\psi_j) < \chi_1^2(0.95))\}$ where $\chi_1^2(0.95)$ is the 95th percentile of the chi-square distribution with 1 degree of freedom.

4.4 Simulation Study

A small simulation study was conducted to assess the finite sample performance of estimators based on the EM algorithm. The one sample setting was considered. Data were generated in the following manner.

- Consider $\pi = 0.1, 0.25$ to represent low and moderate susceptible proportions of the population.
- Let the maximum assessment time be $\tau = 1$.
- For the susceptible subpopulation, the event times, s_i were generated from a $WEI(\lambda, \kappa)$ with λ and κ determined as follows:

  - Initially, we set $\kappa = 1$ so that there is no trend in the lifetime distribution for the susceptible sub-population. To represent an increasing trend in the distribution, $k = 1.25$ will also be investigated.
  - For each parameter configuration, λ was selected such that the probability of experiencing the event of interest by time τ was 0.95 (i.e. $\mathcal{F}_S(\tau) = 0.05$).

- The inspection time distribution was based on $B^* \sim GAM(\gamma_1, \gamma_2)$.
  - Let $VAR(B^*) = \phi = \gamma_1\gamma_2^2 = 0.1$ and 0.5 to represent low to moderate variation in the inspection times.
  - Based on the value of ϕ, the mean of B^*, $\mu = \gamma_1\gamma_2$, and therefore, both γ_1 and γ_2 are determined by solving $\rho = P(S < B|X = 1) = 0.5, 0.75$ where ρ is given by (4.11).
  - We let $b_i = \min(1, b_i^*)$.

- The observed data were then recorded as (b_i, w_i), where $w_i = 1$ if $s_i < b_i$ and 0 otherwise, $i = 1, 2, ..., 2000$.

Figure 4.7 gives a plot of the densities and cumulative distribution for a particular setting. Table 4.1 summarizes results based on fitting a single sample latent class model assuming an exponential lifetime distribution to data generated as above with $\kappa = 1$. In other words, we are considering the situation when the model is specified correctly. The EM algorithm was implemented with the tolerance criteria set to 1×10^{-4}. Depending on the configuration, the average number of iterations required to reach convergence at this tolerance level ranged between 85 and 310. Overall, the estimators based on this model appear to perform reasonably well. Histograms and normal probability plots were generated to identify possible outliers and evidence of non-normality. For most of the parameter configurations, no outliers were present and the plots did not suggest departures from a normal distribution. It is worth noting that the trends seen in the standard errors are broadly consistent with what we would expect from the asymptotic calculations conducted in Section 4.3.4. Specifically the standard errors are smaller when ϕ is larger and when π is larger, all other parameters being equal.

Estimators based on this latent class model did not fare so well for a couple of the parameter configurations investigated. On average, larger numbers of iterations were required to reach convergence for these configurations as well. Although not shown here, the empirical biases and standard errors were quite a bit larger when $\pi = 0.05$ compared to when $\pi > 0.05$. However, the sample size was taken to be $n = 2000$ so when $\pi = 0.05$, we would expect approximately 100 individuals in each dataset to be at risk to experience the event. Of that group, only some will have experienced the event of interest by their inspection times. Depending on the relationship between the lifetime and the inspection time distributions (see Figure 4.7 for example), there may be a relatively small number of individuals testing positive (i.e. with $w_i = 1$) in any given dataset. In this case, there may not be enough information present to estimate (π, λ). It appears that large sample sizes are essential to successful estimation under this model. This is especially the case when π is small.

Figure 4.7: *True underlying event and inspection time distributions when* $P(S < 1) = 0.95$, $\rho = 0.75$, *and* $\phi = 0.5$.

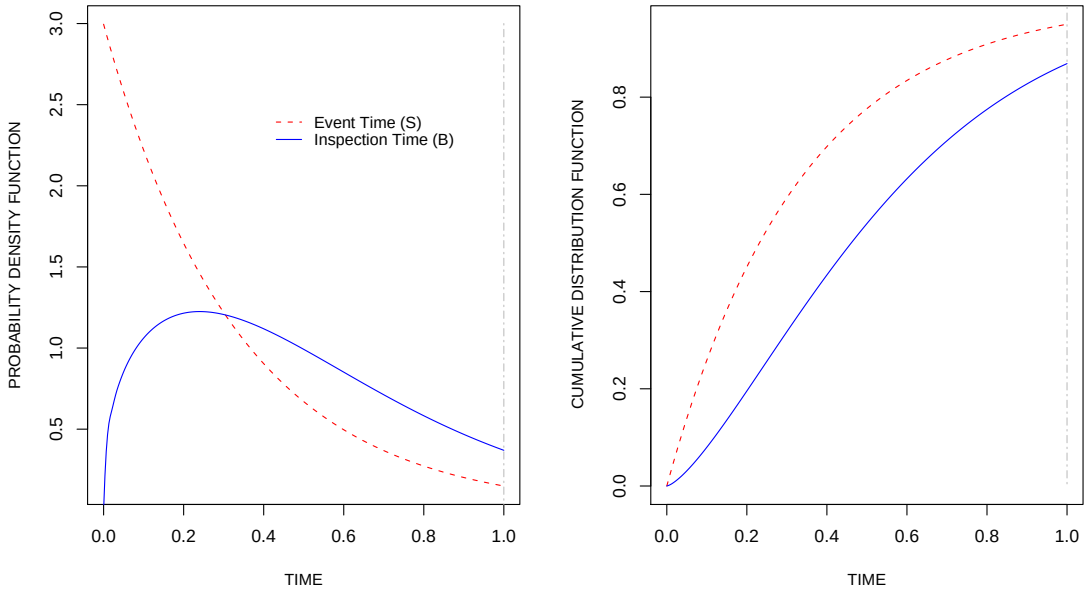


Table 4.1: *Simulation results evaluating the performance of a single sample latent class current status model assuming an exponential lifetime distribution; Number of simulations = 500, sample size = 2000 and* $\kappa = 1$.

Configuration			Exponential ($S \sim EXP(\lambda)$) logitπ		Exponential ($S \sim EXP(\lambda)$) $\log\lambda$	
π	ϕ	ρ	BIAS	SE†	BIAS	SE†
0.25	0.1	0.75	0.0282	0.1512	-0.0167	0.2624
0.25	0.1	0.5	0.0067	0.1938	0.0199	0.2571
0.25	0.5	0.75	0.0181	0.1241	0.0113	0.2488
0.25	0.5	0.5	0.0041	0.1445	-0.0055	0.2318
0.1	0.1	0.75	0.0839	0.3206	0.0685	0.5585
0.1	0.1	0.5	0.0448	0.3692	0.0160	0.5159
0.1	0.5	0.75	0.0438	0.2100	0.0576	0.4299
0.1	0.5	0.5	0.0271	0.2355	-0.0002	0.4172

†SE is the empirical standard error.

Table 4.1 summarizes results based on the correct assumption that the distribution follows an exponential distribution. A small number of simulations were performed for the parameter configuration defined by $\pi = 0.25$, $\phi = 0.5$ and $\rho = 0.5$ again with $\kappa = 1$. Based on additional 75 datasets, the empirical biases and standard errors for logitπ and $\log\lambda$ were 0.0287 (SE=0.1453) and -0.0345 (SE=0.2354) assuming an exponential model and 0.0808 (SE=0.2583) and -0.1077 (SE=0.4101) under a Weibull model. The parameter κ is also estimated when the Weibull model it fit to the data. The empirical bias associated with κ was observed to be 0.0115 with a standard error of 0.1542. Based on this small numerical investigation, the estimated biases and standard errors appear to be larger under the Weibull model as compared to the exponential model when the true distribution is exponential.

We will now turn our attention to piecewise constant models. Three pieces were used and the cut-points (a_1 and a_2) were chosen to be the 33.33% and 66.67% percentiles of

the true underlying lifetime distribution. To ensure there were data available to estimate each of the pieces, the variance, ϕ was chosen such that $P(B < a_1)$, $P(a_1 < B < a_2)$ and $P(B > a_2)$ were greater than 0.1 based on a prespecified average assessment time of 0.6 (μ=0.6). The results based on two parameter configurations are summarized in Table 4.2. The empirical biases and standard errors are higher for the piecewise model than for the correctly specified Weibull model. A more conservative choice of the EM algorithm tolerance (1×10^{-4}, here) may result in smaller biases and standard errors. However, the piecewise approach seems to require a larger number of iterations than those based on parametric model to give converging solutions. When $\kappa = 1.25$, the empirical bias may also be reduced by increasing the number of pieces. A larger number of pieces would provide a better approximation to the true underlying hazard, but if the data are limited over some of the intervals imposed by the cut-points, estimation of the pieces will be difficult. Based on the simulations summarized in this section, it is clear that a large sample size is necessary to successfully estimate parameters associated with a latent class current status model.

Table 4.2: *Simulation results evaluating the performance of a single sample latent class current status model assuming Weibull and piecewise constant lifetime distributions; Number of simulations = 350, sample size = 5000,* $\pi = 0.25$ *and* $P(S < 1) = 0.95$.

Configuration				Weibull ($S \sim WEI(\lambda, \kappa)$)						Piecewise	
				logitπ		$\log \lambda$		κ		logitπ	
μ	ϕ	ρ	κ	BIAS	SE†	BIAS	SE†	BIAS	SE†	BIAS	SE†
0.6	0.2158	0.7	1	-0.0004	0.0925	0.0109	0.1666	-0.0002	0.0116	0.0334	0.1722
0.6	0.6115	0.5	1.25	0.0184	0.1274	-0.0153	0.1791	0.0194	0.1314	0.0959	0.1363

†SE is the empirical standard error.

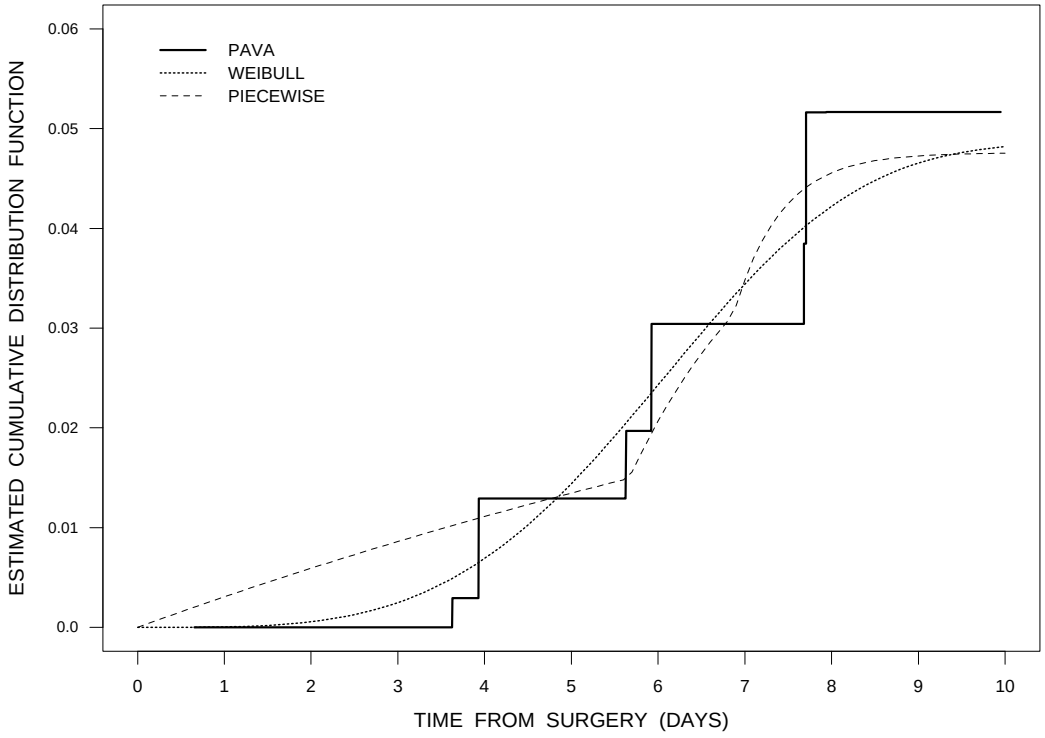
4.5 Application: Orthopedic Surgery Data

For the purposes of this application, the time of surgery is taken to be the origin. Both the time to seroconversion distribution and the probability of seroconversion will be estimated based on these data. Ignoring covariates for now, Weibull and piecewise constant models were fit to obtain $\hat{\pi}$ and $\hat{\mathcal{F}}_S(t;\hat{\lambda},\hat{\kappa})$. Cumulative distribution function estimates based on these models, $\hat{\pi}\left(1-\hat{\mathcal{F}}_S(t;\hat{\lambda},\hat{\kappa})\right)$, along with a nonparametric estimate are displayed in Figure 4.8. These estimates appear to agree over the region for which there is a reasonable amount of data. As Figure 4.2 demonstrates, most of the patients are discharged from the hospital before the eight-day mark so there are little data available to estimate the distribution after that time.

A latent class current status model was fit to the data arising from the four orthopedic surgery studies, the results of which are presented in this section. Weibull, nonparametric and piecewise constant models were considered for the seroconversion time distribution in the latent class current status model and the seroconversion probability was modeled by a logistic distribution. The constraint $\mathcal{F}_S(10) = 0$ was imposed to facilitate nonparametric estimation of $\mathcal{F}_S(\cdot)$. The Pool-Adjacent-Violators algorithm (PAVA) was used to fit the isotonic regression for the nonparametric approach in R (Raubertas 1994). No covariates were included in the model for the seroconversion distribution. It was the probability of seroconversion that was of interest in this application rather than the time to seroconversion. Also, since the proportion of individuals susceptible to HIT is so low, even though the sample size is large here, there would only be a small amount of data available to estimate the effects of covariates on both the probability of seroconversion and the time to seroconversion distributions.

Table 4.3 summarizes results for these data based on Weibull and nonparametric modeling of $\mathcal{F}_S(\cdot)$. The confidence intervals presented in this table are based on profile likelihood pivotals and the p-values using likelihood ratio statistics as described in Section 4.3.6. Results based on a piecewise constant hazards model for $\mathcal{F}_S(\cdot)$ will be reported throughout the text. Three pieces were used for the piecewise constant models and the cut-points were determined by the estimated 33.33% and 66.67% percentiles of the seroconversion time

Figure 4.8: *Estimates of the cumulative distribution function for the seroconversion time based on a latent class model with Weibull and piecewise constant hazard functions and a nonparametric estimate.*



distribution. Three covariates were considered: the timing of the first injection (before versus after), the location of surgery (hip versus knee) and the gender of the patient (male versus female). The covariate effects represent log odds ratios since they are included in the logistic model for the seroconversion probability rather than the seroconversion time distribution, in which case they would represent relative risks.

The models tended to give similar estimates. There is little effect of the timing of the first injection or gender on the odds of seroconversion based on the Weibull and nonparametric models. This was also found to be the case based on the piecewise constant

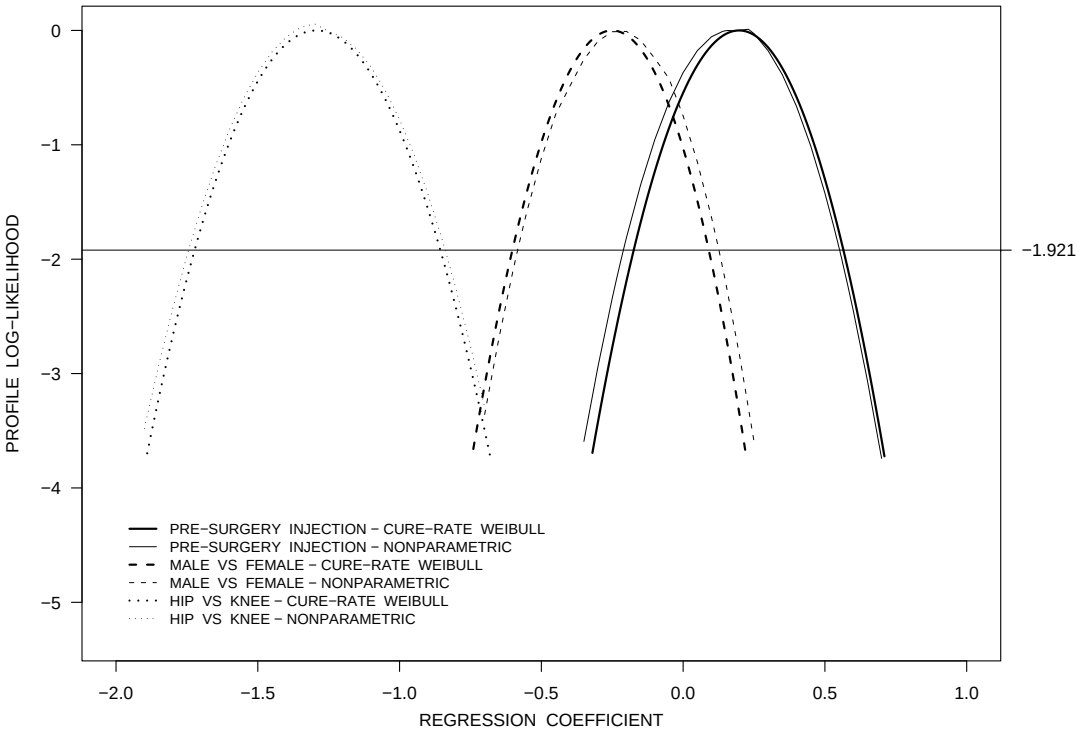
model, under which the estimates of the log odds ratios were 0.207 (95% CI (-0.165,0.576); p=0.438) for injection prior to surgery and -0.248 (95% CI (-0.600,0.092); p=0.313) for gender. The only variable that appeared to significantly influence the risk of seroconversion was the location of the surgery (hip versus knee). Hip surgery patients experienced a lower risk of seroconversion than those in for knee surgery. Estimates for the odds ratio were OR=0.274 (95% CI (0.179, 0.425); $p < 0.001$) based on a Weibull model and for the nonparametric model, OR=0.279 (95% CI (0.175, 0.432); $p < 0.001$). Again, the piecewise constant model resulted in a similar estimate of OR=0.285 (95% CI (0.186, 0.440); $p < 0.001$). The profile likelihood plots for the Weibull and nonparametric seroconversion time distributions are given in Figure 4.9. The horizontal line determines the profile likelihood-based 95% confidence intervals reported in Table 4.3.

Table 4.3: *Estimates of the covariate effects on seroconversion probability based on a latent class current status model.*

	Weibull			Nonparametric		
	EST	95 % CI†	$p-$value	EST	95 % CI†	$p-$value
Prior injection	0.197	(-0.175, 0.566)	0.460	0.163	(-0.211, 0.553)	0.544
Hip surgery	-1.293	(-1.723, -0.856)	<0.001	-1.278	(-1.745, -0.839)	0.001
Male	-0.250	(-0.603, 0.091)	0.310	-0.223	(-0.584, 0.126)	0.388

†CI based on profile likelihood (see Figure 4.9).

Figure 4.9: *Profile likelihood estimates for Weibull and nonparametric models of* $\mathcal{F}_S(s)$.



Chapter 5

Concluding Remarks

5.1 Overview

Throughout this thesis the effects of several types of incomplete life history data on parameter estimation were investigated. It was demonstrated that conducting inference based on simpler, naive models can result in seriously biased estimators and incorrect standard errors which often lead to inaccurate conclusions. Alternative approaches were proposed for these problems and the performance of some of the resulting estimators was shown to be superior to those based on naive models. There is considerable need for extensions of this work as is evident by the following topics summarized by chapter.

5.1.1 Interval-censored Lifetime Data with Mismeasured Covariates

The findings in the simulation studies of Chapter 2 suggested that significant biases can result from naive analyses of interval-censored data with both continuous mismeasured covariates and binary misclassified covariates. Bias reductions can be obtained from corrected likelihood-based analyses and the SIMEX procedure. When large validation datasets are available the corrected likelihood methods work very well with the coverage probability being within the acceptable range of the nominal level in most cases. The SIMEX procedure, while attractive from a coding standpoint, did not perform as well; when there was

minor measurement error or misclassification it sometimes performed acceptably when a large validation study was available, but it was sufficiently unpredictable that it cannot be recommended for use as implemented in the settings of the simulation study.

Piecewise constant baseline hazard (PCBH) models are considered to be robust in many applications. In this context, for the most part these models appeared to give similar results as Weibull models. However, to explore the robustness of the PCBH models more fully here, it would be a valuable exercise to extend the simulation studies to investigate their performance when data are generated from a model other than Weibull. The effect of varying the number of baseline hazard pieces would also be an interesting extension.

Measurement Error for Current Status Data

Further investigation is needed on the effects of misclassification and measurement error when only current status data are available. Table 2.5 summarizes results involving a misclassified binary covariate with current status data for a particular parameter configuration. It would be interesting to extend this to the case of continuous measurement error and perform a more extensive simulation study investigating the impact of mismeasured covariates on the lifetime distribution parameters (λ, κ) as well as the covariate effects for different distributions for the inspection times. To help with the planning of future studies involving current status data, it would also be useful to explore the optimal choice of inspection times.

As (4.10) indicates, the likelihood function based on current status data can be expressed as a generalized linear model (GLM). Specifically, a proportional hazards Weibull regression model lifetime distribution is equivalent to a binary regression model with a complementary log-log link. The literature on covariate mismeasurement in generalized linear models would therefore provide some insight into the effects of mismeasured covariates on estimation as well as possible approaches of handling this problem.

Misclassified Covariates and States

It would be interesting to investigate methodology that addresses mismeasured covariates and misclassification of states simultaneously. Rosychuk & Thompson (2003) consider

a misclassified two-state model in the absence of covariates allowing for two transitions ($0 \rightarrow 1$ and $1 \rightarrow 0$). The behavior of maximum likelihood estimators are investigated and two bias-correction methods are proposed and implemented. Methodology such as that presented in their paper could be extended to consider covariates with and without error in addition to misclassified states.

Bayesian Methods

The approaches implemented in this chapter tended to be computationally burdensome. The Bayesian approach should be investigated as another possible method to deal with mismeasured covariates in progressive multi-state models using the software package WINBUGS. Gustafson (2004) describes this approach for other settings involving the mismeasurement of covariates.

5.1.2 Interval-censored Three-state Data with Mismeasured Covariates

The findings in Chapter 3 were broadly similar to those of Chapter 2. The unique aspect of this setting was the bias induced in estimates of regression coefficients of error-free covariates in transition intensities with no covariates measured with error. Here we found this bias tended to be modest but could be reduced further by use of likelihood methods with a large validation study. The reliability data available for the psoriatic arthritis dataset was very small and the empirical studies suggest that it may be too small to place much confidence in the results of the corrected analyses, either by likelihood or SIMEX approaches.

Validation Studies vs. Reliability Studies

It would be useful to investigate optimal design strategies for selection of validation samples, as well as to compare the utility of reliability studies versus validation studies. In settings where there is no gold standard, it is easier to conduct reliability studies, but there are few guidelines on the optimal design of studies aiming to estimate an intraclass correlation coefficient, or misclassification rates from latent class analyses such as those discussed

for the psoriatic arthritis study. This information would be helpful at the planning stage of future mismeasured covariate problems involving multi-state models.

Misclassified Covariates and States in Multi-State Models

In the motivating application for both Chapters 2 and 3, the states were defined by the number of damaged joints determined by clinical assessment. In my research to date these have been treated as being precisely measured. However, these counts have been demonstrated to vary between physicians on the same patient (Gladman et al. 2004). Therefore, in addition to the presence of error in covariates, there is also error in the response. In other words, the observed states are misclassified versions of the true underlying states. Based on the literature, a mixture modeling approach involving hidden Markov models, or models where the true states of the Markov chain are unobserved, can be taken to deal with the misclassified state problem (Bureau et al. 2003). A more complicated problem, also motivated by the PsA application is one where the misclassification of states and covariate mismeasurement are considered simultaneously.

More Complex Measurement Error Models

Extending this work to accommodate more complex state structures such as progressive models with more states or non-progressive models would also be useful. Extensions to more complex mismeasurement models, possibly involving dependence on other covariates, $\boldsymbol{Z}$, and considering misclassification in discrete covariates and measurement error in continuous covariates simultaneously represents practical areas worthy of development. The challenge in this setting is the need to develop models for the joint distribution of many covariates.

Mismeasured Time-Dependent Covariates

Misclassification and measurement error in fixed covariates were considered in this work. Extension to time-varying variables is a much more complex problem if their values may be influenced by the PsA progression process, but it is worth examining. Model misspecification other than incorrect usage of W in place of X was not considered here. In the

simulation studies, the form of the model used to conduct inference was the same as that used to generate the data, and these models were all Markov models. It would be interesting to consider applications to semi-Markov models and to fit and evaluate piecewise constant baseline intensities models in this setting.

5.1.3 Current Status Data with a Susceptible Fraction

The work in Chapter 4 was motivated by the need to analyze data from several orthopedic studies on seroconversion rates following orthopedic surgery and exposure to blood thinning medication. The findings included that covariate effects can be seriously biased when naive models are fit to the observable status indicators at the time of inspection. Relative efficiency plots indicate the settings when information is maximized for a given sample size and provide rough guidelines on the implications of different inspection time distributions. Two EM algorithms were described including one which facilitated estimation with parametric and nonparametric estimates of the seroconversion time distribution, and a more involved version which gave estimates under a piecewise constant hazards model. For the motivating problem, there is little interest in fitting covariates in the seroconversion time distribution, but there is some appeal to the piecewise constant approach because it would facilitate fitting covariate effects in proportional hazards models.

Comprehensive Simulation Study

It would be useful to extend the simulation studies to compare the performance of the three models; logistic model for the probability of seroconversion (depending on covariates) with nonparametric, Weibull and piecewise constant hazards models for the time to seroconversion distribution. The range of the susceptible fractions investigated via simulation was selected based on the motivating example (i.e. it has been reported that approximately 5% of patients experience HIT following surgery and injection). In other applications, this fraction may be much larger. Therefore, a simulation study investigating a larger range of possible susceptible fractions would represent more situations that may arise in practice and would provide details regarding the relative utility of these models over different fraction values.

Covariates in Seroconversion Time Model

It is possible in principle to put covariates in the model for the event time distribution as well as the model for seroconversion. There can be serious identifiability issues even in settings with right-censored data, and with current status data these may be more challenging. One strategy is to put covariates in one of the two component models but not both of them. This may address some of the computing challenges as well. In the motivating problem there is little interest in characterizing the seroconversion time distribution or related covariate effects, but it must be dealt with to ensure valid inferences as discussed in Chapter 4.

Bivariate Current Status Data

The primary objective of the orthopedic studies was to examine the incidence of deep vein thrombosis (DVT). There were two ways of measuring this outcome. One was based on careful radiographic examination of the patients over 5-11 days following surgery, which detected both symptomatic and asymptomatic clots, and the other was based on contacting the patients 49 days after surgery to ask them if they had any symptoms since surgery (this outcome was therefore based only on symptomatic clots).

The former method of assessment could be viewed as corresponding to a current status observation scheme since the status of patients with respect to DVT is assessed at the examination time. As in the case of seroconversion there is little interest in the actual time a DVT develops, but more in whether such a DVT develops, and one could consider using a bivariate version of the latent class model to examine the association between seroconversion and the development of DVT; it would be expected that these would be negatively correlated since seroconversion increases risk of thrombocytopenia (a decrease in platelet counts) and clots are less likely to occur with lower platelet counts.

Current Status Observation of Covariates

The variable W, the observed seroconversion status, was considered as a response in Chapter 4. However, it may be of interest to consider whether the true seroconversion status

(X) affects the distribution of another outcome of interest, such as the status of patients at 49 days with respect to symptomatic DVT mentioned above. In this case one might form a logistic regression model with a single misclassified binary covariate (W) in addition to several covariates to control for other risk factors. An EM algorithm (similar to that described in Chapter 4) can be used to obtain maximum likelihood estimates in this situation, or one could consider adapting mean score methods as suggested by Reilly & Pepe (1995).

Bibliography

[1] Aitkin, M. and Rocci, R. A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, 12:163–174, 2002.

[2] Andersen, P. K., Borgan, O., Gill, R.D., and Keiding, N. *Statistical Models Based on Counting Processes*. Springer-Verlag, 1993.

[3] Andersen, P. K. and Keiding, N. Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11:91–115, 2002.

[4] Armstrong, B. The effects of measurement errors on relative risk regressions. *American Journal of Epidemiology*, 132(6):1176–1184, 1990.

[5] Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26:641–647, 1955.

[6] Bauer, K. A., Eriksson, B. I., Lassen, M. R., and Turpie, A. G. G. Fondaparinux compared with enoxaparin for the prevention of venous thromboembolism after elective major knee surgery. *New England Journal of Medicine*, 345:1305–1310, 2001.

[7] Broyden, C. G. A new double-rank minimization algorithm. *Notices of the American Mathematical Society*, 16:670, 1969.

[8] Buonaccorsi, J. P., Laake, P., and Veierød, M. B. On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, 61:831–836, 2005.

[9] Bureau, A., Shiboski, S., and Hughes, J. P. Application of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine*, 22:441–462, 2003.

[10] Buzas, J. F. Unbiased scores in proportional hazards regression with covariate measurement error. *Journal of Statistical Planning and Inference*, 67:247–257, 1998.

[11] Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91(433):242–250, 1996.

[12] Carroll, R. J., Ruppert, D., and Stefanski, L. A. *Measurement Error in Nonlinear Models*. Chapman & Hall, 1995.

[13] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, 2^{nd} edition, 2006.

[14] Carroll, R. L. Measurement error in epidemiologic studies. *Encyclopedia of Biostatistics*, 3:2491–2519, 1998.

[15] Commenges, D. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11:167–182, 2002.

[16] Cook, J. and Stefanski, L. A. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994.

[17] Cox, D. R. Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–220, 1972.

[18] Cox, D. R. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[19] DeGrottola, V. and Tu, X. M. Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014, 1994.

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

[21] Eriksson, B. I., Bauer, K. A., Lassen, M. R., and Turpie, A. G. G. Fondaparinux compared with enoxaparin for the prevention of venous thromboembolism after hip-fracture surgery. *New England Journal of Medicine*, 345:1298–1394, 2001.

[22] Evans, M. and Swartz, T. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 10(3):254–272, 2000.

[23] Farewell, V. T. A model for a binary variable with time censored observations. *Biometrika*, 64:43–46, 1977.

[24] Farewell, V. T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046, 1982.

[25] Farewell, V. T. Mixture models in survival analysis: are they worth the risk? *The Canadian Journal of Statistics*, 14(3):257–262, 1986.

[26] Fletcher, R. A new approach to variable metric methods. *Computer Journal*, 13:317–322, 1970.

[27] National Psoriasis Foundation. About psoriatic arthritis: the basics. www.psoriasis.org/about/psa/basics.php, 2004. (Accessed 3 Nov, 2005).

[28] Fuller, W. A. *Measurement Error Models*. Wiley, 1987.

[29] Gentleman, R. C., Lawless, J. F., Lindsey, J. C., and Yan, P. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, 13:805–821, 1994.

[30] Gladman, D. G., Cook, R. J., Schentag, C., and et al. The clinical assessment of patients with psoriatic arthritis: results of a reliability study of the Spondyloarthritis Research Consortium of Canada. *The Journal of Rheumatology*, 31(6):1126–1131, 2004.

[31] Gladman, D. G., Farewell, V. T., Buskila, D., and et al. Reliability of measurements of active and damaged joints in psoriatic arthritis. *The Journal of Rheumatology*, 17:62–64, 1990.

[32] Gladman, D. G., Farewell, V. T., and Nadeau, C. Clinical indicators of progression in psoriatic arthritis: multivariate relative risk model. *The Journal of Rheumatology*, 22:675–679, 1995.

[33] Goldfarb. D. A family of variable metric methods derived by variational means. *Mathematics of Computation*, 24:23–26, 1970.

[34] Gong, G., Whittemore, A. S., and Grosser, S. Censored survival data with misclassified covariates: a case study of breast-cancer mortality. *Journal of the American Statistical Association*, 85:20–28, 1990.

[35] Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 2:215–231, 1974.

[36] Gorfine, M., Hsu, L., and Prentice, R. L. Nonparametric correction for covariate measurement error in a stratified Cox model. *Biostatistics*, 5:75–87, 2004.

[37] Greene, W. F. and Cai, J. Measurement error in covariates in the marginal hazards model for multivariate failure time data. *Biometrics*, 60:987–996, 2004.

[38] Grüger, J., Kay, R., and Schumacher, M. The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47(2):595–605, 1991.

[39] Gustafson, P. *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall, 2004.

[40] Henderson, R., Diggle, P., and Dobson, A. Joint modeling of longitudinal measurements and event time data. *Biostatistics*, 4:465–480, 2000.

[41] Hougaard, P. Multi-state models: a review. *Lifetime Data Analysis*, 5:239–264, 1999.

[42] Hu, C. and Lin, D. Y. Cox regression with covariate measurement error. *The Scandinavian Journal of Statistics*, 29:637–655, 2002.

[43] Hu, P., Tsiatis, A. A., and Davidian, M. Estimating the parameters in the cox model when covariate variables are measured with error. *Biometrics*, 54:1407–1419, 1998.

[44] Huang, X., Stefanski, L., and Davidian, M. Latent-model robustness in structural measurement error models. *Biometrika*, 93:53–64, 2006.

[45] Husted, J. A., Tom, B. D., Farewell, V. T., Schentag, C. T., and Gladman, D. D. Description and prediction of physical functional disability in psoriatic arthritis. *Arthritis & Rheumatology*, 53(3):404–409, 2005.

[46] Jennrich, R. I. and Bright, P. B. Fitting systems of linear differential equations using computer generated exact derivatives. *Technometrics*, 18:385–392, 1976.

[47] Jewell, N. P. and van der Laan, M. J. Current status data: review, recent developments and open problems. U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 113, 2002.

[48] Kalbfleisch, J. D. and Lawless, J. F. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985.

[49] Kalbfleisch, J. D. and Lawless, J. F. Some statistical methods for panel life history data. *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time*, pages 185–192, October 1989.

[50] Kalbfleisch, J. D. and Lawless, J. F. Analysis of life history data. University of Waterloo Lecture Notes, November 1999.

[51] Kalbfleisch, J. D. and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2002.

[52] Kaplan, E. L. and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[53] Kelley, W. N., Harris, E. D., Ruddy, S., and Sledge, C. B., editors. *Textbook of Rheumatology*. W. B. Saunders, 1981.

[54] Kong, F. H. and Gu, M. Consistent estimation in Cox proportional hazards model with covariate measurement errors. *Statistica Sinica*, 9:129–135, 1999.

[55] Küchenhoff, H. and Carroll, R. J. Segmented regression with errors in predictors: semiparametric and parametric methods. *Statistics in Medicine*, 16:169–188, 1997.

[56] Küchenhoff, H., Lederer, W., and Lesaffre, E. Asymptotic variance estimation for the Misclassification SIMEX. Working Papers, Ludwig-Maximilian-University, Munich, Germany, 2006.

[57] Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. A general method for dealing with misclassification in regression: the Misclassification SIMEX. *Biometrics*, 62(1):85–96, 2005.

[58] Kulich, M. and Lin, D. Y. Additive hazards regression with covariate measurement error. *Journal of the American Statistical Association*, 95:238–248, 2000.

[59] Lam, K. F. and Xue, H. A semiparametric regression cure model with current status data. *Biometrika*, 92(3):573–586, 2005.

[60] Lassen, M. R., Bauer, K. A., Eriksson, B. I., and Turpie, A. G. G. Postoperative fondaparinux versus preoperative enoxaparin for prevention of venous thromboembolism in elective hip-replacement surgery: a randomized double-blind comparison. *Lancet*, 359:1715–1720, 2002.

[61] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, 2 edition, 2003.

[62] Lawless, J. F. and Zhan, M. Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *The Canadian Journal of Statistics*, 26(4):549–565, 1998.

[63] Lee, C. C. The min-max algorithm and isotonic regression. *The Annals of Statistics*, 11(2):467–477, 1983.

[64] Lee, E. T. and Wang, J. W. *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, 2003.

[65] Li, C. and Taylor, J. M. G. A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21:3235–3247, 2002.

[66] Li, Y. and Lin, X. Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *Journal of the American Statistical Association*, 98(461):191–203, 2003.

[67] Maller, R. and Zhou, X. *Survival Analysis with Long-Term Survivors*. John Wiley & Sons, 1996.

[68] Martinez, J. M. Practical quasi-Newton methods for solving nonlinear systems. *Journal of Computational and Applied Mathematics*, 124:97–121, 2000.

[69] Matthews, D.E. and Cook, R. The analysis of survival data. University of Waterloo Lecture Notes, Winter 2005.

[70] Nakamura, T. Corrected score function for errors-in-variables models: methodology and application to generalized linear models. *Biometrika*, 77:127–137, 1990.

[71] Nakamura, T. Proportional hazards model with covariates subject to measurement error. *Biometrics*, 48:829–838, 1992.

[72] Toronto General Hospital University Health Network. HIV overview: natural history. www.tthhivclinic.com/overview_home.htm. (Accessed 8 Nov, 2005).

[73] Pan, W., Lin, X., and Zeng, D. Structural inference in transition measurement error models. *Biometrics*, 62:402–412, 2006.

[74] Peng, Y. and Dear, K. B. G. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243, 2000.

[75] Pepe, M. S. and Fleming, T. R. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86:108–113, 1991.

[76] Pepe, M. S., Self, S. G., and Prentice, R. L. Further results on covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine*, 8:1167–1178, 1989.

[77] Prentice, R. L. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69:331–342, 1982.

[78] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical Recipes in* C^{++}. Cambridge University Press, 2 edition, 2002.

[79] Raboud, J. M. The effects of errors in measurement in survival analysis. Ph.D. Thesis, University of Toronto, 1991.

[80] Raubertas, R. F. S-news: reponse to John Steward's message regarding isotonic regression. www.biostat.wustl.edu/archives/html/s-news/2005-02/msg00143.html, 1994. (Accessed 27 Apr, 2007).

[81] Reeves, G. K. and Cox, D. R. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine*, 17:2157–2177, 1998.

[82] Reilly, M. and Pepe, M. S. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314, 1995.

[83] Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors and not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

[84] Ross, S. M. *Introduction to Probability Models*. Academic Press, New York, 5^{th} edition, 1993.

[85] Rosychuk, R. J. and Thompson, M. E. Bias correction of two-state latent Markov process parameter estimates under misclassification. *Statistics in Medicine*, 22:2035–2055, 2003.

[86] SAS, http://support.sas.com/documentation/onlinedoc/sas9doc.html. *SAS OnlineDoc®* 9.1.3.

[87] Satten, G. A. Estimating the extent of tracking in interval-censored chain-of-events data. *Biometrics*, 55:1228–1231, 1999.

[88] Schafer, D. W. and Purdy, K. G. Likelihood analysis for errors-in-variables regression with replicate measurements. *Biometrika*, 83:813–824, 1996.

[89] Schoenberg, R. Optimization with the quasi-Newton method. 2001.

[90] Shanno, D. F. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24:145–160, 1970.

[91] Shiboski, S. C. Generalized additive models for current status data. *Lifetime Data Analysis*, 4:29–50, 1998.

[92] Song, X., Davidian, M., and Tsiatis, A. A. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58:742–753, 2002.

[93] Song, X. and Huang, Y. On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics*, 61:702–714, 2005.

[94] Stefanski, L. A. and Carroll, R. J. Covariate measurement error in logistic regression. *The Annals of Statistics*, 13:1335–1351, 1985.

[95] Stefanski, L. A. and Carroll, R. J. Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, 74:703–716, 1987.

[96] Stefanski, L. A. and Cook, J. R. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256, 1995.

[97] Sun, Jianguo. *The Statistical Analysis of Interval-censored Failure Time Data*. Springer-Verlag, 2006.

[98] Sy, J. P. and Taylor, J. M. G. Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[99] Taylor, J. M. G. Semi-parametric estimation in failure time mixture models. *Biometrics*, 51:899–907, 1995.

[100] Tsiatis, A. A. and Davidian, M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88:447–458, 2001.

[101] Turnbull, B. W., Jiang, W., and Clark, L. C. Regression models for recurrent event data: parametric random effects models with measurement error. *Statistics in Medicine*, 16:853–864, 1997.

[102] Turpie, A. G. G., Bauer, K. A., Eriksson, B. I., and Lassen, M. R. Postoperative fondaparinux versus postoperative enoxaparin for prevention of venous thromboembolism after elective hip-replacement surgery: a randomised double-blind trial. *Lancet*, 359:1721–1726, 2002.

[103] Wang, C. Y., Hsu, L., Feng, Z. D., and Prentice, R. L. Regression calibration in failure time regression. *Biometrics*, 53:131–145, 1997.

[104] Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93(441):249–261, 1998.

[105] White, H. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–26, 1982.

[106] Wulfsohn, M. S. and Tsiatis, A. A. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339, 1997.

[107] Xie, S. X., Wang, C. Y., and Prentice, R. L. A risk set calibration method for failure time regression by using a covariate reliability sample. *Journal of the Royal Statistical Society, Series B*, 63:855–870, 2001.

[108] Xu, J. and Zeger, S. L. Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics*, 50:375–387, 2001.

[109] Yi, G. Y. and Cook, R. J. Errors in the measurement of covariates. *Encyclopedia of Biostatistics, 2^{nd} Edition*, 3:1741–1748, 2005.

[110] Yi, G. Y. and Lawless, J. F. A corrected likelihood method for the proportional hazards model with covariates subject to measurement error. *Journal of Statistical Planning and Inference*, 2007. In press.

[111] Zhou, H. and Pepe, M. Auxiliary covariate data in failure time regression analysis. *Biometrika*, 82:139–149, 1995.

[112] Zhou, H. and Wang, C. Y. Failure time regression with continuous covariates measured with error. *Journal of the Royal Statistical Society, Series B*, 62:657–665, 2000.

[113] Zucker, D. M. A pseudo-partial likelihood method for semiparametric survival regression with covariate errors. *Journal of the American Statistical Association*, 100:1264–1277, 2005.

[114] Zucker, D. M. and Spiegelman, D. Inference for the proportional hazards model with misclassified discrete-valued covariates. *Biometrics*, 60:324–334, 2004.