

TOPICS IN LINEAR MODELS: METHODS FOR
CLUSTERED, CENSORED DATA AND
TWO-STAGE SAMPLING DESIGNS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Lynn Marie Johnson

May 2013

© 2013 Lynn Marie Johnson

ALL RIGHTS RESERVED

TOPICS IN LINEAR MODELS: METHODS FOR CLUSTERED, CENSORED
DATA AND TWO-STAGE SAMPLING DESIGNS

Lynn Marie Johnson, Ph.D.

Cornell University 2013

In this dissertation, we consider the use of linear models in the presence of clustered, right-censored failure time data. The semiparametric accelerated failure time model is a log-linear model which provides a useful, easy to interpret method for characterizing the relationship between failure time and covariates. Clustered failure time data can be handled in the context of the accelerated failure time model by using marginal estimation methods or by incorporating a random cluster-level frailty term. However, regression parameter estimation for both approaches requires the optimization of a non-smooth objective function. We use an extension of the induced smoothing procedure of Brown and Wang (2006) to construct a marginal estimation procedure that permits fast and accurate computation of regression parameter estimates and standard errors using widely available numerical methods. The regression parameter estimates are shown to be strongly consistent and asymptotically normal and, in addition, the asymptotic distribution of the smoothed estimator is shown to coincide with that obtained without the use of smoothing. In the case of the AFT frailty model, we use an extension of the induced smoothing procedure in conjunction with an EM-type algorithm to construct a procedure which permits simultaneous estimation of the regression parameters, the baseline cumulative hazard, and the parameter indexing a general frailty distribution.

We also consider two-stage sampling designs for linear models. Epidemiological studies frequently involve an important risk factor which is difficult or expensive to measure. When the response variable and a collection of co-

variates are easy to obtain on a large sample of the population, two-stage sampling designs provide a natural framework for using the easily obtained data to identify an informative subsample on which to collect the more difficult to measure covariate. We review traditional two-stage outcome-dependent sampling designs and develop a novel residual-dependent sampling design for this setting. Inverse probability weighted estimators for the sampling designs are presented and asymptotic properties of the estimators are discussed. The proposed residual-dependent sampling design is easy to implement and results in more efficient estimators than the outcome-dependent sampling design in many situations.

BIOGRAPHICAL SKETCH

Lynn Marie Johnson (née Schooley) was born in Berkeley Heights, New Jersey and graduated from Governor Livingston Regional High School. She received a bachelor's degree in mathematics from Virginia Polytechnic Institute and State University in Blacksburg, Virginia and a master's degree in mathematics from the University of Colorado in Boulder, Colorado. While completing a master's degree in biometry at the University of Colorado Health Sciences Center in Denver, Colorado, Lynn worked as a statistical analyst for the Continuous Improvement in Cardiac Surgery Program at the Denver Veterans Affairs Medical Center and for the Division of Vector-Borne Infectious Diseases at the Centers for Disease Control and Prevention in Fort Collins, Colorado.

This dissertation is dedicated to my parents, Richard and Patricia Schooley, whose love and support has made so many wonderful things in my life possible.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Professor Robert L. Strawderman, for his guidance and support throughout my graduate career at Cornell University. His encouragement was paramount to my success. I would also like to thank Professor James G. Booth, Professor Bruce W. Turnbull and Professor Martin T. Wells for their participation on my committee, and Professor Oliver H. Gao for his support and friendship during my time at Cornell University.

I am grateful to the dedicated staff of the Departments of Statistical Science and Biological Statistics and Computational Biology at Cornell University, especially Beatrix Johnson, Todd Cullen, and Diana Drake, for their assistance throughout the years.

Finally, I would like to express my love and appreciation to my wonderful family. To my husband Dan, you are the best part of my day, every day. And to my daughters, Iris and Mae, you have brought so much love and laughter into our lives. Thank you!

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	iv
Acknowledgments	v
Table of Contents	vi
List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 The Semiparametric AFT Model for Clustered Data	1
1.2 A Two-Stage Residual-Dependent Sampling Design	7
2 Marginal Methods for the Semiparametric AFT Model	10
2.1 Methodology and Key Results	10
2.1.1 Notation and assumptions	10
2.1.2 Estimation for clustered data using the Gehan weight . . .	11
2.1.3 Induced smoothing for clustered data	11
2.2 Methods of Variance Estimation	15
2.2.1 Sandwich variance estimator	15
2.2.2 Brown and Wang (2006) procedure for clustered data . . .	16
2.2.3 Resampling variance estimator	16
2.3 Simulation Study	18
2.4 Remarks	21
3 The Semiparametric AFT Frailty Model	23
3.1 Estimation for the AFT Model with General Frailty	23
3.1.1 Notation and assumptions	23
3.1.2 SES algorithm for AFT frailty models	24
3.1.3 Variance estimation	32
3.1.4 Moment-based estimation of θ	34
3.2 Implementation	36
3.2.1 Gamma frailty distribution	36
3.2.2 Inverse Gaussian frailty distribution	39
3.3 Simulation Study	41
3.3.1 Estimator performance and impact of $\hat{\theta}^{(0)}$	42
3.3.2 Impact of smoothing parameter choice	48
3.3.3 Impact of censoring level, cluster size, and sample size . .	49
3.3.4 Two covariate models	51
3.3.5 Robustness to distributional assumptions	52
3.3.6 Variance estimation	53
3.4 Illustrations	53
3.4.1 Litter-matched tumorigenesis	54
3.4.2 Isosorbide dinitrate in angina	54

3.4.3	Recurrent time to infection following catheterization . . .	56
3.5	Remarks	57
4	A Two-Stage Residual-Dependent Sampling Design	60
4.1	Two-Stage Sampling Designs	60
4.1.1	Outcome-Dependent Sampling	61
4.1.2	Residual-Dependent Sampling	62
4.2	Estimation	68
4.2.1	Outcome-Dependent Sampling	70
4.2.2	Residual-Dependent Sampling	71
4.3	Simulation Study	73
4.4	Remarks	83
A	Proofs for Chapter 2	85
B	Proofs for Chapter 3	96
C	Theorems and Proofs for Chapter 4	97
	Bibliography	102

LIST OF FIGURES

3.1	Mean squared error of the baseline cumulative hazard	46
3.2	Proportion of sample at risk	47

LIST OF TABLES

2.1	Marginal method simulation results	20
3.1	Frailty model simulation results	43
3.2	Impact of smoothing parameter	49
3.3	Impact of censoring level and cluster size	50
3.4	Two covariate model	51
3.5	Impact of true baseline and frailty distributions when fitting a gamma frailty model	52
3.6	Weighted bootstrap standard error estimates	53
3.7	Illustration 3.4.2 results	56
4.1	Increasing stage one effect	76
4.2	Increasing stage two effect	77
4.3	Increasing stage one and stage two effects	78
4.4	Correlated stage one and stage two covariates	79
4.5	Impact of strata definitions	80
4.6	ODS design including stage one covariate	81
4.7	Standard error estimates	82

CHAPTER 1

INTRODUCTION

In this dissertation, we consider semiparametric log-linear models for censored data with a specific focus on clustered observations and two-stage sampling designs for linear models. We begin by providing a detailed introduction to each topic.

1.1 The Semiparametric AFT Model for Clustered Data

The need to analyze failure time data possibly subject to right-censoring arises in a number of fields, including medicine, economics, epidemiology, demography, and engineering. Semiparametric regression models are commonly used for characterizing the relationship between failure time and covariates, with the Cox proportional hazards regression model (Cox, 1972) being used almost exclusively in practice. The accelerated failure time (AFT) model (e.g., Kalbfleisch and Prentice, 2002) provides a useful but infrequently used alternative to the Cox proportional hazards model. Letting \bar{T}_i and X_i respectively denote the failure time and vector of observed covariates for observation i ($i = 1, \dots, n$), the AFT model specifies that $\log \bar{T}_i = X_i' \beta + \epsilon_i$, where the error terms are independent and identically distributed with an unspecified distribution. The regression coefficient β has a nice interpretation and a variety of simple estimators are available when $\bar{T}_1, \dots, \bar{T}_n$ are fully observed. The infrequent use of this model in applications involving censored failure time data may be attributed in large part to the computational challenges that arise in both regression parameter and covariance matrix estimation.

In the presence of censoring, the observed data for subject i can be described

by the triplet (T_i, Δ_i, X_i) where $T_i = \min(\bar{T}_i, C_i)$, $\Delta_i = I(\bar{T}_i \leq C_i)$, and C_i denotes the censoring time for subject i . Tsiatis (1990) proposes to estimate β using a weighted estimating equation of the form

$$W_n^*(\beta) = \sum_{i=1}^n w_i(\beta) \Delta_i \left[X_i - \frac{\sum_{j=1}^n X_j I\{e_j(\beta) \geq e_i(\beta)\}}{\sum_{j=1}^n I\{e_j(\beta) \geq e_i(\beta)\}} \right], \quad (1.1)$$

where $e_i(\beta) = \log(T_i) - X_i' \beta$ and $w_i(\cdot)$ are nonnegative weight functions ($i = 1, \dots, n$). Due to the fact that β appears in this expression only inside indicator functions, $W_n^*(\beta)$ is not a continuous function of β and a solution to $W_n^*(\beta) = 0$ typically does not exist. Parameter estimates may instead be obtained by minimizing $\|W_n^*(\beta)\|$, where $\|v\|$ denotes $(v'v)^{1/2}$ for a vector v . However, this minimization problem may admit several solutions and, because $W_n^*(\beta)$ is not necessarily monotone in β , the resulting set of minimizers is not even guaranteed to be convex. Hence, despite the existence of a consistent and asymptotically normal sequence of generalized solutions (e.g., Tsiatis, 1990), identifying this sequence can be challenging in practice.

Fyngenson and Ritov (1994) show that using the Gehan weight function $w_i(\beta) = \sum_{j=1}^n I\{e_j(\beta) \geq e_i(\beta)\}$ ($i = 1, \dots, n$) leads to the monotone estimating equation

$$W_n(\beta) = \sum_{i=1}^n \sum_{j=1}^n \Delta_i (X_i - X_j) I\{e_i(\beta) - e_j(\beta) \leq 0\}. \quad (1.2)$$

Recognizing that $W_n(\beta)$ is the gradient of the convex objective function

$$O_n(\beta) = \sum_{i=1}^n \sum_{j=1}^n \Delta_i \{e_j(\beta) - e_i(\beta)\} I\{e_i(\beta) - e_j(\beta) \leq 0\}, \quad (1.3)$$

a regression parameter estimate may be obtained by minimizing $O_n(\beta)$ with respect to β . The resulting set of solutions is convex and thus easier to locate than in the general case. However, even in this comparatively nice setting, the associated lack of smoothness continues to present computational challenges. Numerous methods have been proposed for finding parameter estimates derived

from (1.2) and (1.3). To date, the most promising methods for regression parameter estimation based on the Gehan-weighted estimating equation include the use of linear programming techniques (e.g., Jin et al., 2003) and the implementation of smoothing methods; see, for example, Brown and Wang (2005, 2006), Heller (2007), and Song et al. (2007). However, while such methods can be implemented with relative ease, the computational burden can be high, especially with large datasets.

Estimating the covariance matrix of the regression parameter estimate obtained under the AFT model remains a challenging problem. Fyngenson and Ritov (1994) show that the regression parameter estimate derived from (1.3) is asymptotically normal with a covariance matrix that involves the hazard function of the unspecified error distribution. Direct estimation of the covariance matrix thus requires an estimate of this hazard function. Tsiatis (1990) suggests kernel-based estimation, whereas Fyngenson and Ritov (1994) suggest a form of numerical differentiation. Both have proven to be unstable choices in the presence of censored data and several authors have since tackled this problem in other ways; see, for example, Jones (1997) and Jin et al. (2003). Jin et al. (2003) propose to randomly reweight the Gehan log-rank objective function (1.3) and then minimize the resulting perturbed objective function. Repeating this process a large number of times, the covariance matrix may then be estimated using the empirical covariance matrix of these parameter estimates. This interesting and useful approach eliminates the need to estimate the indicated hazard function. However, the computationally intensive nature of this procedure quickly becomes unwieldy, particularly with large datasets. Huang (2002), Strawderman (2005), and Jin et al. (2006b) propose useful alternatives in three related problems.

Several authors have recently proposed useful smoothing methods for non-smooth estimating equations arising in the AFT model; see, for example, Brown and Wang (2005, 2006), Heller (2007), and Song et al. (2007). Each of these smoothing methods leads to a continuously differentiable objective or estimating function that can be dealt with using standard numerical methods. Of direct relevance to this dissertation are the works of Brown and Wang (2006) and Heller (2007). Building on the work of Brown and Wang (2005), Brown and Wang (2006) propose the use of “induced smoothing” for the Gehan estimating equation (1.2). This method involves solving the equation $E_Z\{W_n(\beta + \Gamma_n Z)\} = 0$, where $W_n(\cdot)$ is given in (1.2), Z is a continuous, mean zero normal random vector independent of all of the data, and Γ_n is a sequence of matrices converging to zero with elements $\Gamma_{n,ij} = O_p(n^{-1/2})$. The smoothed estimating equation $E_Z\{W_n(\beta + \Gamma_n Z)\}$ reduces to

$$\tilde{W}_n(\beta) = \sum_{i=1}^n \sum_{j=1}^n \Delta_i(X_i - X_j) \Phi \left\{ \frac{e_j(\beta) - e_i(\beta)}{r_{n,ij}} \right\}, \quad (1.4)$$

where $r_{n,ij}^2 = (X_i - X_j)' \Sigma_n (X_i - X_j)$, $\Sigma_n = \Gamma_n^2$, and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. In a related vein, Heller (2007) directly approximates the indicator function $I(u \leq 0)$ in $W_n(\beta)$ with $1 - \Upsilon(u/h)$, where $\Upsilon(\cdot)$ denotes a “local distribution function” satisfying certain conditions and the fixed scalar parameter h is used control the accuracy of approximation. The resulting estimating equation,

$$W_n^{**}(\beta) = \sum_{i=1}^n \sum_{j=1}^n \Delta_i(X_i - X_j) \Upsilon \left\{ \frac{e_j(\beta) - e_i(\beta)}{h} \right\}, \quad (1.5)$$

has the same structure as (1.4). In fact, upon taking $\Upsilon(\cdot)$ to be the standard normal distribution function $\Phi(\cdot)$, (1.5) is essentially a special case of (1.4), utilizing a fixed bandwidth h in place of the covariate-dependent bandwidth $r_{n,ij}$. Heller (2007) also proposes a robust version of (1.5) having a bounded influence func-

tion. A potential difference between (1.4) and (1.5) lies in the ability of the former to employ a smoothing parameter that respects the scaling and covariance structure of the solution sequence. Brown and Wang (2006) claim but do not prove that the sequence of solutions obtained under (1.4) has the same asymptotic distribution as that obtained in the absence of smoothing. Heller (2007) proves that the solution sequence obtained under (1.5) is consistent and asymptotically normal, provided that h satisfies $nh \rightarrow \infty$ and $nh^4 \rightarrow 0$ as $n \rightarrow \infty$. Interestingly, Heller (2007) further proves that (1.2) and (1.5) are asymptotically equivalent but does not establish the equivalence result posited in Brown and Wang (2006).

The problem of regression parameter estimation under the AFT model with correlated survival data has also been considered. For example, Lin and Wei (1992), Lee et al. (1993), and Jin et al. (2006a) consider the setting in which failure times are grouped into clusters, such that observations within a cluster may be correlated but observations in distinct clusters may be considered independent. Each propose a marginal method for rank-based estimation of regression parameters, avoiding the need to model the correlation structure among observations. Jin et al. (2006a) also devise a suitable extension of the resampling procedure for covariance matrix estimation proposed in Jin et al. (2003). Pan (2001), Zhang and Peng (2007), and Xu and Zhang (2010) instead propose AFT gamma frailty models, handling the dependence among failure times within a cluster using an additive cluster-level random effect; see also Strawderman (2006) for related work in the case of a recurrent event outcome. Each relies on an EM-type algorithm for estimation of the regression parameters, gamma frailty parameter and cumulative hazard function of the unspecified error distribution. As in the independent data case, regression parameter estimation

requires solving a discontinuous estimating equation. Pan (2001), Zhang and Peng (2007), and Xu and Zhang (2010) consider extensions of existing methods (e.g., grid searches, linear programming techniques) whereas Strawderman (2006) introduces a variation on the Levenberg-Marquardt algorithm, a Newton-type method, for use in nonsmooth settings in an effort to improve computational efficiency. These various methods suffer from estimation and computational challenges that equal or exceed those experienced in the case of independent failure time data.

In this dissertation, we develop methods of estimation for the semiparametric AFT model in the presence of clustered survival data using both marginal methods and a general frailty model. In the second chapter, we extend the smoothing procedure of Brown and Wang (2006) to the problem of marginal estimation of the regression parameter in the presence of clustered data. We prove that the resulting estimator is consistent and asymptotically normal in both the independent and correlated data settings. We further establish the equivalence of these limiting distributions with those arising in the unsmoothed case, providing rigorous justification of the equivalence claim made in Brown and Wang (2006) for the case of independent failure times and its extension to the setting of clustered data. Several possible methods of covariance matrix estimation are evaluated, among them a generalization of the Brown and Wang (2006) procedure and a modification of the resampling procedure due to Jin et al. (2006a). A useful consequence of developing the extended Brown and Wang (2006) estimator is an easy-to-compute sandwich estimator that avoids the need for resampling. The proposed methods substantially ease the computational burden of previously proposed methods for parameter and covariance matrix estimation. These results have previously appeared in Johnson and Strawderman (2009).

The third chapter of this dissertation proposes an estimation procedure for a general semiparametric AFT frailty model that combines the EM algorithm for estimating equations (Elashoff and Ryan, 2004) and induced smoothing procedure developed in the second chapter. The resulting EM-type procedure is referred to as the Smoothing Expectation and Substitution (SES) algorithm and permits simultaneous estimation of the regression parameter, the baseline cumulative hazard, and the parameter indexing a general frailty distribution. A novel method of moments framework for the frailty parameter is introduced and used to construct several possible method-of-moments estimators, including a generalized method of moments estimator. Standard error estimation is considered using an adaptation of the weighted bootstrap methodology studied in Ma and Kosorok (2005). The details for implementing the proposed algorithm, computing standard the error estimates and calculating the moment-based estimators is provided assuming that the frailties follow a gamma distribution; we also provide implementation details assuming that frailties follow an inverse Gaussian distribution. These results have previously appeared in Johnson and Strawderman (2012).

1.2 A Two-Stage Residual-Dependent Sampling Design

Epidemiological studies frequently involve an important risk factor which is difficult or expensive to measure. When the response variable and a collection of covariates are easy to obtain on a large sample of the population, there is an inherent interest in using the easily obtained data to identify an informative subsample on which to collect the more difficult to measure covariate. Two-stage sampling designs, in which information collected on a large sample of the population at stage one is used to select a stage two sample on which the

remaining data will be collected, provide a natural framework for developing such studies.

Two-stage outcome-dependent designs use data collected on the response variable at stage one to select the second stage sample. The case-control study, in which the disease status of subjects is determined in the first stage and exposure is measured for a predetermined number of cases and controls at the second stage, is a familiar example of a two-stage outcome-dependent sampling design (Wild, 1991). The case-control study involves a binary response variable but two-stage outcome-dependent designs for continuous response variables have also been considered. For continuous response variables, subjects selected at stage one are typically stratified based on their observed responses. Then a predetermined number of subjects is selected from each strata for the stage two sample (e.g., Lawless et al., 1999; Zhou et al., 2002; Weaver and Zhou, 2005). Ideally, strata containing the most informative subjects will be oversampled. Covariate data collected at stage one may also be considered in the stratification procedure (e.g., Lawless et al., 1999).

Two-stage outcome-dependent designs using stratified sampling have been shown to improve the efficiency of inference in many situations. However, determining how to define the strata and strata-specific sample sizes to improve efficiency is a difficult problem in general. Incorporating data collected on stage one covariates into the sampling design can make these determinations decidedly more complex.

In the fourth chapter of this dissertation, we provide a review of traditional two-stage outcome-dependent samplings designs for linear models with continuous response variables. We develop a novel two-stage residual-dependent sampling design in which the residuals from a linear regression model fit to

the stage one data are used to identify which subjects are most informative and hence should be included in the stage two sample. Inverse probability weighted estimators for the stage two model are presented and the asymptotic properties of the estimators are discussed. The proposed residual-dependent sampling design results in more efficient estimators than the outcome-dependent sampling design in many situations.

2.1 Methodology and Key Results

2.1.1 Notation and assumptions

Consider a random sample of n independent clusters with K_i members in the i th cluster. Let \bar{T}_{ik} and C_{ik} denote the failure time and censoring time for the k th member of the i th cluster, and let X_{ik} denote the corresponding $p \times 1$ vector of covariates. We assume that $(\bar{T}_{i1}, \dots, \bar{T}_{iK_i})'$ and $(C_{i1}, \dots, C_{iK_i})'$ are independent conditional on the covariates $(X_{i1}, \dots, X_{iK_i})'$. Let the survival data for the k th member of i th cluster be denoted $W_{ik} = (\log T_{ik}, \Delta_{ik}, X_{ik})'$ where $T_{ik} = \min(\bar{T}_{ik}, C_{ik})$ and $\Delta_{ik} = I(\bar{T}_{ik} \leq C_{ik})$.

We assume that the marginal distribution of T_{ik} follows the accelerated failure time (AFT) model

$$\log \bar{T}_{ik} = X'_{ik} \beta_0 + \epsilon_{ik},$$

where β_0 is a $p \times 1$ vector of unknown regression parameters contained in a compact subset \mathbb{B} of \mathbb{R}^p and $(\epsilon_{i1}, \dots, \epsilon_{iK_i})'$ ($i = 1, \dots, n$) are independent random error vectors. Within each cluster i , the error terms $\epsilon_{i1}, \dots, \epsilon_{iK_i}$ may be correlated; however, as in Jin et al. (2006a, §4), we assume that these error terms are exchangeable with a common, unknown marginal distribution. That is, for any $i, j = 1, \dots, n$ and $K \leq \min(K_i, K_j)$, the vectors $(\epsilon_{i1}, \dots, \epsilon_{iK})'$ and $(\epsilon_{j1}, \dots, \epsilon_{jK})'$ have the same distribution. Evidently, the case of independent failure time data follows as a special case of the above model upon setting $K_i = 1$ for all i .

2.1.2 Estimation for clustered data using the Gehan weight

Let $e_{ik}(\beta) = \log(T_{ik}) - X'_{ik}\beta$. Under the assumptions of §2.1.1, the relevant extension of (1.3) to the clustered data setting may be written

$$L_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \{e_{jl}(\beta) - e_{ik}(\beta)\} I \{e_{ik}(\beta) - e_{jl}(\beta) \leq 0\}; \quad (2.1)$$

see, for example, Jin et al. (2006a, §4). Observe that $L_n(\beta)$ is a continuous convex function for $\beta \in \mathbb{B}$ and thus differentiable almost everywhere. The derivative of the objective function with respect to β , or $S_n(\beta) = \nabla L_n(\beta)$, is the discontinuous function

$$S_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} (X_{ik} - X_{jl}) I \{e_{ik}(\beta) - e_{jl}(\beta) \leq 0\}. \quad (2.2)$$

Let $\hat{\beta}_n = \operatorname{argmin}_{\beta \in \mathbb{B}} L_n(\beta)$. The solution to this minimization problem may not be unique; however, the convexity of $L_n(\beta)$ implies that the set of minimizers on \mathbb{B} is convex (e.g., Fygenon and Ritov, 1994). The lack of smoothness makes minimization of $L_n(\beta)$ computationally challenging, particularly with multiple covariates. However, under regularity conditions to be described later, the results of Jin et al. (2006a, Theorem 5) imply that there exists a sequence of solutions that is strongly consistent for β_0 and, in addition, such that $n^{1/2}(\hat{\beta}_n - \beta_0)$ converges in distribution to a $N(0, A^{-1}\Omega A^{-1})$ random vector, where $\Omega = \lim_{n \rightarrow \infty} \operatorname{var} \{n^{1/2}S_n(\beta_0)\}$ and $A = \nabla S_0(\beta_0)$ for $S_0(\beta) = \lim_{n \rightarrow \infty} S_n(\beta)$. An explicit formula for A is provided in (A.1). In addition to the numerical challenges that arise in computing the solution $\hat{\beta}_n$, variance estimation is difficult due to the dependence on A and the fact that $S_n(\beta)$ is not differentiable in β .

2.1.3 Induced smoothing for clustered data

Brown and Wang (2005) propose an “induced smoothing” method for approximating discontinuous but monotone estimating functions using contin-

uously differentiable functions. Assuming independent failure time observations, Brown and Wang (2006) apply this smoothing method to the problem of estimating the regression parameter in the AFT model, using (1.4) in place of (1.2). As shown below, the extension of this methodology to the problem of estimating β in the clustered data setting under the assumptions of §2.1.1 is straightforward.

Let Z be a $N(0, I_p)$ random vector independent of the data, where I_p denotes the $p \times p$ identity matrix. Let Γ be a $p \times p$ matrix such that $\|\Gamma\| = O(1)$ and $\Gamma^2 = \Sigma$, where Σ is some symmetric, positive definite matrix. Then, similarly to Brown and Wang (2005, 2006), a smoothed score function may be constructed by adding the random perturbation $n^{-1/2}\Gamma Z$ to the argument of the score function $S_n(\beta)$ in (2.2) and then taking the expectation with respect to Z . Specifically, with $\tilde{S}_n(\beta) = E_Z \{S_n(\beta + n^{-1/2}\Gamma Z)\}$, an easy calculation shows

$$\tilde{S}_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} (X_{ik} - X_{jl}) \Phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \right\} \right], \quad (2.3)$$

where $r_{ikjl}^2 = (X_{ik} - X_{jl})' \Sigma (X_{ik} - X_{jl})$. When $K_i = K_j = 1$ for $i, j = 1, \dots, n$, this estimating equation reduces to (1.4). Alternatively, one might work directly with the smoothed objective function $\tilde{L}_n(\beta) = E_Z \{L_n(\beta + n^{-1/2}\Gamma Z)\}$. Let $\phi(\cdot)$ denote the standard normal density function. Then, using standard results for normal random variables and integration by parts, we have

$$\tilde{L}_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \left[\{e_{jl}(\beta) - e_{ik}(\beta)\} H_{ikjl}^{(n)}(\beta) + \frac{r_{ikjl}}{n^{1/2}} h_{ikjl}^{(n)}(\beta) \right], \quad (2.4)$$

where r_{ikjl} is defined above,

$$H_{ikjl}^{(n)}(\beta) = \Phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \right\} \right], \quad h_{ikjl}^{(n)}(\beta) = \phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \right\} \right]. \quad (2.5)$$

A straightforward calculation shows that $\nabla \tilde{L}_n(\beta) = \tilde{S}_n(\beta)$.

Let $\tilde{\beta}_n = \operatorname{argmin}_{\beta \in \mathbb{B}} \tilde{L}_n(\beta)$. The smoothed objective function, $\tilde{L}_n(\beta)$, is convex and continuously differentiable and standard numerical methods can be used

to efficiently compute $\tilde{\beta}_n$. Alternatively, $\tilde{\beta}_n$ can be found as the multivariate root of $\tilde{S}_n(\beta)$.

The asymptotic results, summarized below and proved in Appendix A, rely on the following regularity conditions.

A1. The parameter space \mathbb{B} containing β_0 is a compact subset of \mathbb{R}^p .

A2. $\sum_{k=1}^{K_i} \|X_{ik}\| + K_i$ is bounded almost surely by a nonrandom constant ($i = 1, \dots, n$).

A3. The assumptions of §2.1.1 hold with $\text{var}(\epsilon_{11}) < \infty$.

A4. The matrix $A = \nabla S_0(\beta_0)$, where $S_0(\beta) = \lim_{n \rightarrow \infty} S_n(\beta)$, exists and is non-singular.

A5. Let $f_0(\cdot)$ denote the marginal density associated with model error term ϵ_{11} and let $\lambda_0(\cdot)$ denote its corresponding hazard function. Then, $f_0(\cdot)$ and $f'_0(\cdot)$ are bounded functions on \mathbb{R} with

$$\int_{\mathbb{R}} \left\{ \frac{f'_0(t)}{f_0(t)} \right\}^2 f_0(t) dt < \infty.$$

A6. The marginal distribution of C_{rs} is absolutely continuous and has a bounded density $g_{rs}(\cdot)$ on \mathbb{R} ($r = 1, \dots, n; s = 1, \dots, K_r$).

As indicated in the statements of Theorems 2.1.1 and 2.1.2 below, $\Sigma = \Gamma^2$ is assumed to be a symmetric and positive definite matrix with $\|\Gamma\| < \infty$. Conditions A1, A2, A4, A5 and A6 are standard and ensure consistency and asymptotic normality of the unsmoothed Gehan estimator (e.g., Tsiatis, 1990; Ying, 1993; Jin et al., 2006a). Condition A3 implies $|\text{cov}(\epsilon_{ik}, \epsilon_{il})| \leq \text{var}(\epsilon_{11})$ ($i = 1, \dots, n; k, l = 1, \dots, K_i$); hence, the covariances between all error terms within a cluster are bounded.

Theorem 2.1.1 *Let $\Sigma = \Gamma^2$ be any symmetric and positive definite matrix with $\|\Gamma\| < \infty$. Under conditions A1-A4, $\tilde{\beta}_n$ is a strongly consistent estimator of β_0 .*

Theorem 2.1.2 *Let $\Sigma = \Gamma^2$ be any symmetric and positive definite matrix with $\|\Gamma\| < \infty$. Under conditions A1-A6, $n^{1/2}(\tilde{\beta}_n - \beta_0)$ converges in distribution to $N(0, \Psi)$, where $\Psi = A^{-1}\Omega A^{-1}$, $\Omega = \lim_{n \rightarrow \infty} \text{var} \{n^{1/2}S_n(\beta_0)\}$ and $A = \nabla S_0(\beta_0)$ is defined in (A.1).*

The above results provide theoretical justification for the proposed smoothing procedure when estimating regression parameters under the marginal AFT model with clustered failure time data. Importantly, the matrices A and Ω in Theorem 2.1.2 are defined in terms of (2.2), demonstrating that the limiting distribution of $n^{1/2}(\tilde{\beta}_n - \beta_0)$ coincides with that for $n^{1/2}(\hat{\beta}_n - \beta_0)$, where $\hat{\beta}_n$ is obtained via the unsmoothed objective function (2.1). Since justification for the independent data case follows directly from the above theorems upon setting $K_i = 1$ ($i = 1, \dots, n$). Theorems 2.1.1 and 2.1.2 also provide rigorous justification for the claims made in Brown and Wang (2006).

Remark 2.1.3 *The above results hold for a general smoothing matrix Γ that satisfies certain minimal conditions. Brown and Wang (2006) propose an iterative procedure for estimating β_0 , in which $\Sigma = \Gamma^2$ is updated at each iteration using successive estimates of Ψ . One implementation of this procedure in the clustered data setting is provided in §2.2.2.*

Remark 2.1.4 *The smoothing bandwidth employed in (2.3) and (2.4) is $O(n^{-1/2})$, where n denotes the number of independent clusters. In the absence of clustering, Heller (2007) recommends the choice $h = \hat{\sigma}n^{-0.26}$ in (1.5), where $\hat{\sigma}$ is an estimate of the residual variance obtained using a minimizer of the unsmoothed equation (1.3). The selection $h = O(n^{-0.26})$ is motivated as that which provides the “quickest rate of convergence while satisfying the bandwidth constraint $nh^4 \rightarrow 0$.” In asymptotic terms, Theorem 2.1.2 suggests that such oversmoothing is unnecessary.*

2.2 Methods of Variance Estimation

2.2.1 Sandwich variance estimator

The sandwich form of the covariance matrix of $n^{1/2}(\tilde{\beta}_n - \beta_0)$ in Theorem 2.1.2 suggests a natural estimator provided that suitable estimates of both A and Ω can be found. In the independent data case, Brown and Wang (2006) suggest estimating A with $\tilde{A}_n = \nabla \tilde{S}_n(\tilde{\beta}_n)$; Theorems 2.1.1 and 2.1.2 imply that this remains a consistent estimator in the clustered data setting. Brown and Wang (2006) further suggest several estimates of Ω , including the asymptotic variance of $n^{1/2}S_n(\beta_0)$ provided in Jin et al. (2003) and an estimator of Ω based on the U-statistic structure of the estimating function (1.4). However, neither estimator of Ω properly accounts for the correlation between observations within a cluster. Lee et al. (1993) show that the asymptotic variance of $n^{1/2}S_n(\beta_0)$ in the clustered data case can be consistently estimated via $\hat{\Omega}_n = \hat{\Omega}_n(\hat{\beta}_n)$, where

$$\hat{\Omega}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{l=1}^{K_i} \{\hat{\xi}_{ik}(\beta)\}^{\otimes 2},$$

$v^{\otimes 2} = vv'$ for any vector v , and

$$\hat{\xi}_{ik}(\beta) = \sum_{j=1}^n \sum_{l=1}^{K_j} \left(\frac{\Delta_{ik}}{n} (X_{ik} - X_{jl}) I\{e_{ik}(\beta) < e_{jl}(\beta)\} - \frac{\Delta_{jl}}{n} \left[\frac{\sum_{r=1}^n \sum_{s=1}^{K_r} (X_{ik} - X_{rs}) I\{e_{rs}(\beta) \geq e_{jl}(\beta)\}}{\sum_{m=1}^n \sum_{k=1}^{K_m} I\{e_{mk}(\beta) \geq e_{jl}(\beta)\}} \right] I\{e_{ik}(\beta) \geq e_{jl}(\beta)\} \right).$$

Conditions A1-A5 ensure that $\hat{\Omega}_n$ is a consistent estimator of Ω ; with the addition of condition A6, Ψ can be consistently estimated in the clustered data setting using

$$\hat{\Psi}_n = \tilde{A}_n^{-1} \hat{\Omega}_n \tilde{A}_n^{-1}. \quad (2.6)$$

2.2.2 Brown and Wang (2006) procedure for clustered data

As suggested in Brown and Wang (2006), an iterative procedure can be used to simultaneously estimate the regression parameters and their covariance matrix. Denoting $\tilde{A}_n(\beta) = \nabla \tilde{S}_n(\beta)$, the proposed procedure consists of the following steps in the clustered data setting:

1. Set $i = 0$ and initialize $\hat{\Sigma}_{(0)}$ such that $\|\hat{\Sigma}_{(0)}\| = O(1)$; for example, $\hat{\Sigma}_{(0)} = I_p$.
2. Set $i = i + 1$ and solve $\tilde{S}_n(\beta) = 0$ for $\tilde{\beta}_{(i)}$ using $\Gamma = (\hat{\Sigma}_{(i-1)})^{1/2}$ in equation (2.3).
3. Using $\tilde{\beta}_{(i)}$, calculate $\tilde{A}_{(i)} = \tilde{A}_n(\tilde{\beta}_{(i)})$ and $\hat{\Omega}_{(i)} = \hat{\Omega}(\tilde{\beta}_{(i)})$.
4. Compute $\hat{\Sigma}_{(i)} = \tilde{A}_{(i)}^{-1} \hat{\Omega}_{(i)} \tilde{A}_{(i)}^{-1}$.
5. Repeat steps 2–4 until convergence of both $\tilde{\beta}_{(i)}$ and $\hat{\Sigma}_{(i)}$ is achieved to a specified tolerance.

In our experience, convergence of this algorithm typically occurs with relatively few iterations, the value of $\hat{\Sigma}_{(*)}$ at convergence being very close to $\hat{\Psi}_n$ in (2.6).

Remark 2.2.1 *The above procedure makes use of a data-dependent smoothing parameter. The proofs of Theorems 2.1.1 and 2.1.2 assume that the matrix Γ is known; however, since $\|\hat{\Gamma}_{(*)} - (A^{-1}\Omega A^{-1})^{1/2}\| = O_p(n^{-1/2})$, replacing Γ by $\hat{\Gamma}_{(*)}$ does not alter these asymptotic results.*

2.2.3 Resampling variance estimator

Jin et al. (2006a, §4) propose a useful resampling method for estimating Ψ in the presence of correlated data. This method, which can be motivated by the conditional multiplier central limit theorem (e.g., Martinussen and Scheike, 2006, p.

43), involves randomly reweighting the Gehan log-rank objective function (2.1) and then minimizing the resulting perturbed objective function. Specifically, let

$$L_n^*(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} Z_i Z_j \Delta_{ik} \{e_{jl}(\beta) - e_{ik}(\beta)\} I\{e_{ik}(\beta) - e_{jl}(\beta) \leq 0\}$$

where Z_1, \dots, Z_n are independent positive random variables with $E(Z_i) = \text{var}(Z_i) = 1$ ($i = 1, \dots, n$). Let $\hat{\beta}_n^* = \text{argmin}_{\beta \in \mathbb{B}} L_n^*(\beta)$. Jin et al. (2006a, Theorem 5) prove that, conditional on the data $(W_{ik}; k = 1, \dots, K_i; i = 1, \dots, n)'$, the limiting distribution of $n^{1/2}(\hat{\beta}_n^* - \hat{\beta}_n)$ converges almost surely to the limiting distribution of $n^{1/2}(\hat{\beta}_n - \beta_0)$. Thus, the distribution of $\hat{\beta}_n$ can be approximated by repeatedly generating random samples Z_1, \dots, Z_n and then minimizing $L_n^*(\beta)$ to obtain realizations of $\hat{\beta}_n^*$. The covariance matrix of $\hat{\beta}_n$ can be approximated directly by the empirical covariance matrix of the realizations of $\hat{\beta}_n^*$.

Jin et al. (2006a, §4) work directly with the unsmoothed Gehan objective function and utilize linear programming methods in combination with resampling in order to obtain regression parameter and covariance matrix estimates. Specifically, linear programming is used to minimize $L_n(\beta)$, obtaining the estimated regression parameter $\hat{\beta}_n$; it is then applied repeatedly in minimizing each of the realizations of $L_n^*(\beta)$ generated for the purposes of covariance matrix estimation. The use of linear programming methods can be avoided by randomly reweighting the smoothed objective function $\tilde{L}_n(\beta)$ in (2.4). Such an approach allows for standard numerical methods to be used for minimization, resulting in the potential for computational savings with larger datasets. With Z_1, \dots, Z_n defined as above, let

$$\tilde{L}_n^*(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} Z_i Z_j \Delta_{ik} \left[\{e_{jl}(\beta) - e_{ik}(\beta)\} H_{ikjl}^{(n)}(\beta) + \frac{r_{ikjl}}{n^{1/2}} h_{ikjl}^{(n)}(\beta) \right],$$

where $H_{ikjl}^{(n)}(\beta)$ and $h_{ikjl}^{(n)}(\beta)$ are defined in (2.5), and define $\tilde{\beta}_n^* = \text{argmin}_{\beta \in \mathbb{B}} \tilde{L}_n^*(\beta)$.

Theorems 2.1.1 and 2.1.2 imply that an argument identical to the one given in

Jin et al. (2006a, Theorem 5) can be used to show that the limiting distribution of $n^{1/2}(\tilde{\beta}_n^* - \tilde{\beta}_n)$ converges almost surely to the limiting distribution of $n^{1/2}(\tilde{\beta}_n - \beta_0)$. The covariance matrix of $\tilde{\beta}_n$ can then be approximated exactly as described above, using the simulated realizations of $\tilde{\beta}_n^*$ in place of the β_n^* s.

2.3 Simulation Study

Two simulation studies were carried out to assess the performance of $\tilde{\beta}_n$ as well as to evaluate the covariance matrix estimators described in §2.2. The proposed simulation studies are modeled after that described in Jin et al. (2006a, §5), allowing for a direct comparison between their simulation results and those to be summarized below.

Specifically, for each cluster, we use the algorithm of Johnson (1987, §10.1) to generate two failure times from the bivariate Gumbel distribution

$$F(t_1, t_2) = F_1(t_1)F_2(t_2) [1 + \theta \{1 - F_1(t_1)\} \{1 - F_2(t_2)\}]$$

where $-1 \leq \theta \leq 1$, $F_k(\cdot)$ is the cumulative distribution function for an exponential random variable with hazard function $\lambda_k = \exp(\beta_1 X_{1k} + \beta_2 X_{2k})$, X_{1k} is Bernoulli(0.5), and X_{2k} is standard normal truncated at ± 2 ($k = 1, 2$). All covariates are generated independently and the correlation between \bar{T}_1 and \bar{T}_2 is $\theta/4$. The resulting failure time model is a special case of the AFT model of §2.1.1 with true regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$. Censoring times are independently generated from a Uniform(0, τ) distribution, where τ is selected to achieve a desired level of censoring. Similarly to Jin et al. (2006a, §5), we consider the cases $\theta = 0$ and $\theta = 1$, 50 clusters of size two, and censoring percentages of 0%, 25%, and 50%.

Two different estimation methods are considered. Method 1 refers to the iter-

ative method of §2.2.2 for simultaneously estimating the regression parameters and covariance matrix. Method 2 refers to estimating the regression parameter by minimizing the smoothed objective function (2.4) with the fixed choice $\Sigma = I_2$. Within Method 2, we consider estimating the covariance matrix using the resampling-based variance estimate of §2.2.3 and also using the sandwich variance estimator (2.6). All simulations were conducted in R and use the built-in package NLM for optimization (R Development Core Team, 2005); the simulation code is available upon request.

Table 2.1 summarizes the results of two simulation studies. Each row of the table is based on the same 1000 simulated datasets. In the first simulation study, the semiparametric AFT model of §2.1.1 is fit using the covariates X_{1k} and X_{2k} . We report the results for the estimation of the regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$ and associated standard errors using Methods 1 and 2. The second simulation study repeats the first simulation study, fitting a model that uses the covariates $X_{1k}^* = X_{1k}$ and $X_{2k}^* = X_{2k}/500$. The underlying failure time model is identical to that used in the first simulation study, the true regression parameters now being $\beta_1^* = \beta_1 = 1$ and $\beta_2^* = 500\beta_2 = 250$. However, in contrast to the first simulation study, the magnitudes of X_{2k}^* and β_2^* are quite different from those of X_{1k}^* and β_1^* . The results for β_1^* (not shown) are very similar to those reported in Table 2.1 for β_1 ; hence, we only report the results for β_2^* . The intent of the second study is to investigate the impact of using the fixed smoothing parameter $\Sigma = I_2$ versus the data-dependent smoothing parameter of §2.2.2, a choice that better reflects the covariance structure and scaling of the regression parameter.

Considering only β_1 and β_2 , the relative biases are observed to be small, comparable in magnitude, and generally increase with the censoring percentage. In

Table 2.1: Marginal method simulation results

Regression		Method 1				Method 2			
Parameter	θ	Censoring %	RBIAS	RSE	RSEE1	RBIAS	RSE	RSEE2A	RSEE2B
$\beta_1 = 1$	0	0	0.32	0.25	0.25	0.21	0.25	0.25	0.25
		25	3.91	0.28	0.28	3.41	0.28	0.28	0.28
		50	4.30	0.35	0.36	2.22	0.34	0.36	0.35
	1	0	0.91	0.26	0.25	0.78	0.25	0.25	0.25
		25	1.58	0.27	0.28	1.12	0.27	0.27	0.27
		50	5.52	0.37	0.36	3.38	0.36	0.36	0.35
$\beta_2 = 0.5$	0	0	0.74	0.28	0.26	0.78	0.28	0.26	0.26
		25	1.89	0.30	0.30	1.58	0.30	0.30	0.30
		50	5.13	0.38	0.38	3.40	0.38	0.38	0.38
	1	0	1.23	0.26	0.26	1.20	0.26	0.26	0.26
		25	1.99	0.30	0.30	1.69	0.28	0.30	0.30
		50	4.40	0.40	0.38	2.70	0.38	0.38	0.38
$\beta_2^* = 250$	0	0	0.28	0.27	0.27	0.00	0.27	0.26	0.29
		25	2.99	0.29	0.30	1.96	0.29	0.29	0.35
		50	5.77	0.38	0.38	2.76	0.37	0.38	0.46
	1	0	5.80	0.26	0.27	2.40	0.26	0.26	0.30
		25	1.56	0.29	0.29	5.60	0.29	0.29	0.35
		50	5.25	0.37	0.38	2.20	0.36	0.37	0.46

Results based on 1000 replications and $n = 50$ pairs for regression parameter and standard error estimates obtained using the induced smoothing methodology. Results for β_1 and β_2 are based on an accelerated failure time model that depends on covariates X_1 and X_2 ; results for β_2^* are based on an AFT model that depends on covariates $X_1^* = X_1$ and $X_2^* = X_2/500$.

RBIAS, $1000 \times$ absolute relative bias; RSE, empirical standard error, relative to parameter; RSEE1, standard error relative to parameter, with standard error estimate obtained using iterative method of §2.2.2 with $\hat{\Sigma}_{(0)} = I_2$; RSEE2A, standard error relative to parameter, with standard error estimate obtained using the resampling procedure of §2.2.3 with 500 random reweightings and regression parameters estimated using the induced smoothing procedure of §2.1.3 with the fixed choice $\Sigma = I_2$; RSEE2B, standard error relative to parameter, with standard error estimate based on (2.6) and regression parameters estimated as described for RSEE2A.

addition, estimates obtained using Method 1 frequently exhibit greater bias than those obtained using Method 2, with no apparent reduction in standard error. The standard error estimates for β_1 and β_2 are accurate and similar across all estimation methods. Remarkably, the results reported here are also comparable to those summarized in the right panel of Table 1 in Jin et al. (2006a) for the Gehan weight function, where no smoothing is employed.

Turning to the comparison of results for β_2 and β_2^* , biases generally follow the patterns described above. In addition, all methods of standard error esti-

mation perform well, though some evidence of inflation in the relative standard error $RSEE_{2B}$ is now present. Overall, the results suggest that the smoothing parameter has a minimal impact on the bias or actual standard error of the regression parameter estimates. However, given the relative accuracy of both $RSEE_1$ and $RSEE_{2A}$, the discrepancy observed in $RSEE_{2B}$ suggests that the scaling of the problem, hence choice of smoothing parameter, can adversely impact the accuracy of (2.6).

On the basis of these results, we recommend using Method 1 as described in §2.2.2; in comparison with the simulation-based methodology of Jin et al. (2006a), it requires far less computational effort with no evidence penalty in bias or accuracy of standard error estimation.

2.4 Remarks

The attractive nature of the induced smoothing procedure, in both computational and theoretical terms, stems largely from the convexity of the Gehan-weighted objective function (2.4). The asymptotic results obtained in this chapter make significant use of this convexity. A minor extension of these results also can be used to justify an alternative smoothing methodology for the bounded influence estimator introduced in Heller (2007). Variations on this smoothing methodology may facilitate simpler and more stable estimation procedures for AFT frailty models; see, for example, Pan (2001), Strawderman (2006), and Zhang and Peng (2007).

The use of the Gehan weight function in (1.2) has frequently been criticized for the inefficiency of the resulting estimator. The selection of an alternative weight function may result in efficiency improvements at the expense of monotonicity, resulting in weaker asymptotic statements and increased compu-

tational challenges. To counteract these drawbacks, Jin et al. (2003) propose to use the Gehan estimator as a starting point for successively solving a sequence of convex optimization problems derived from (1.1). Jin et al. (2006a) extends these results to the setting of multivariate failure time data. The resulting class of estimation procedures is computationally stable and yields a consistent and asymptotically normal sequence of estimators with reasonably general weight functions. However, it does not lend itself to a simple method of variance estimation. Use of the resampling method described in §2.2.3 is recommended for this purpose but only amplifies the required computational effort. Jin et al. (2006b) propose a strongly related class of procedures for the Buckley–James estimator. Starting from the Gehan estimator, Strawderman (2005) demonstrates how one may instead use one-step estimation to achieve the same goal and introduces an alternative simulation-based method of variance computation that requires no additional optimization. The results of this chapter show that the induced smoothing methodology provides an asymptotically valid and computationally convenient starting point for each of these other methods of estimation. In addition, the methodology itself can be directly incorporated as part of the iterative methods developed in Jin et al. (2003, 2006b,a); the asymptotic results of this chapter guarantee that their results also remain valid for the corresponding smoothed version.

A direct extension of this smoothing methodology is available for general weight functions. However, it lacks the same computational convenience due to important structural differences between the Gehan-weighted estimating equation and those used for general weight functions.

3.1 Estimation for the AFT Model with General Frailty

3.1.1 Notation and assumptions

As in §2.1.1, we consider a random sample of n independent clusters with K_i members in the i th cluster, we let \bar{T}_{ik} and C_{ik} respectively denote the failure time and censoring time for the k th member of the i th cluster, and we let X_{ik} denote the corresponding $p \times 1$ vector of covariates. We assume that the covariates X_{ik} have bounded support, that the components of X_{ik} are linearly independent for all i and k , and that the cluster sizes are bounded (i.e., $K_i \leq K < \infty$ for all i). Fixing i and conditional on a positive continuous random variable W_i , we further assume that \bar{T}_{ik} follows the semiparametric accelerated failure time (AFT) model

$$\log(\bar{T}_{ik}) = X'_{ik}\beta_0 + \epsilon_{ik}, \quad (3.1)$$

where β_0 is an unknown $p \times 1$ vector of bounded regression parameters, $\epsilon_{i1}, \dots, \epsilon_{iK_i}$ are independent, and each ϵ_{ik} is a continuous random variable with finite variance and hazard function

$$\lambda(\epsilon_{ik}|W_i = w_i) = w_i\lambda_0(\epsilon_{ik}) \quad (3.2)$$

for a positive, bounded, and continuous baseline hazard function $\lambda_0(\cdot)$ that is otherwise unspecified. Let $\Lambda_0(s) = \int_{-\infty}^s \lambda_0(u)du$ denote the cumulative baseline hazard of the error terms. In the context of the failure time data described above, W_i represents a random frailty term shared by the K_i members of the i th cluster, where W_i has a density function $g(\cdot|\theta)$ and θ is a scalar parameter. Un-

less otherwise specified, $g(\cdot|\theta)$ is assumed to be parameterized such that W_i has mean one; typically, θ then characterizes the variance of the frailty distribution.

The stated assumptions imply $(\epsilon_{i1}, \dots, \epsilon_{iK_i})'$ ($i = 1, \dots, n$) are independent random error vectors that, within each cluster, are correlated but exchangeable. It is possible to relax the exchangeability assumption through stratification or further modeling; we do not consider such extensions here. We further assume independent and noninformative censoring in the same sense as Nielsen et al. (1992); hence, all model parameters can be identified from the observed failure time data $O_{ik} = (\log T_{ik}, \Delta_{ik}, X_{ik})'$, where $T_{ik} = \min(\bar{T}_{ik}, C_{ik})$ and $\Delta_{ik} = I(\bar{T}_{ik} \leq C_{ik})$ ($k = 1, \dots, K_i, i = 1, \dots, n$). The setting of independent failure time data arises as a special case upon setting $W_i = 1$ (i.e., with probability one) for all i .

3.1.2 SES algorithm for AFT frailty models

EM and ES algorithms

To motivate our approach, we first consider how the EM algorithm might be implemented for the joint estimation of the unknown parameters $\psi = (\theta, \beta, \Lambda_0)$ of the model specified by (3.1) and (3.2), for a general frailty distribution $g(\cdot|\theta)$. Given W_i ($i = 1, \dots, n$), the relevant complete data log-likelihood is $L_c(\theta, \beta, \Lambda_0) = L_1(\theta) + L_2(\beta, \Lambda_0)$, where

$$L_1(\theta) = \sum_{i=1}^n \{D_i \log W_i + K_i \log g(W_i|\theta)\},$$

$$L_2(\beta, \Lambda_0) = \sum_{i=1}^n \sum_{k=1}^{K_i} \{\Delta_{ik} \log \lambda_0(e_{ik}(\beta)) - W_i \Lambda_0(e_{ik}(\beta))\},$$

$$e_{ik}(\beta) = \log(T_{ik}) - X'_{ik}\beta \text{ and } D_i = \sum_{k=1}^{K_i} \Delta_{ik}.$$

Let $\mathbf{O} = \{O_{ik}; k = 1, \dots, K_i, i = 1, \dots, n\}$ denote the observed data for all n clusters, and let $\hat{\psi}^{(s)} = (\hat{\beta}^{(s)}, \hat{\theta}^{(s)}, \hat{\Lambda}_0^{(s)})$ denote the current estimates of the param-

eters. For the E-step of the EM algorithm, we calculate the conditional expectation of the complete data log-likelihood with respect to the frailty terms given the observed data and under the assumption that the true parameter is $\hat{\psi}^{(s)}$. In particular, $E\{L_c(\theta, \beta, \Lambda_0) | \mathbf{O}, \hat{\psi}^{(s)}\}$ is given by the sum of

$$E\{L_1(\theta) | \mathbf{O}, \hat{\psi}^{(s)}\} = \sum_{i=1}^n \left[D_i E(\log W_i | \mathbf{O}, \hat{\psi}^{(s)}) + K_i E \left\{ \log g(W_i | \theta) | \mathbf{O}, \hat{\psi}^{(s)} \right\} \right] \quad (3.3)$$

and

$$E\{L_2(\beta, \Lambda_0) | \mathbf{O}, \hat{\psi}^{(s)}\} = \sum_{i=1}^n \sum_{k=1}^{K_i} \left\{ \Delta_{ik} \log \lambda_0(e_{ik}(\beta)) - E(W_i | \mathbf{O}, \hat{\psi}^{(s)}) \Lambda_0(e_{ik}(\beta)) \right\} \quad (3.4)$$

For the M-step of the EM algorithm, we would like to maximize equations (3.3) and (3.4) with respect to θ , β , and $\Lambda_0(\cdot)$. In principle, the univariate function (3.3) can be maximized with respect to θ using standard numerical methods. For example, when all required expectations exist in closed form (e.g., W_i follows a gamma distribution), implementation is quite straightforward. However, maximization of (3.4) with respect to β and $\Lambda_0(\cdot)$ is impossible without further parametric assumptions on $\lambda_0(\cdot)$.

Using arguments similar to those given in Pan (2001) and Zhang and Peng (2007), or a modification of the argument given in Strawderman (2006), $\Lambda_0(\cdot)$ can be estimated nonparametrically by

$$\hat{\Lambda}_0(t) = \sum_{i=1}^n \sum_{k=1}^{K_i} \frac{\Delta_{ik}}{\sum_{j=1}^n \sum_{l=1}^{K_j} \hat{W}_j I\{e_{jl}(\hat{\beta}) \geq e_{ik}(\hat{\beta})\}} I\{e_{ik}(\hat{\beta}) \leq t\}, \quad (3.5)$$

where $\hat{\beta}$ and \hat{W}_j respectively estimate β and $E(W_j | \mathbf{O})$ ($j = 1, \dots, n$). Similarly, an estimate of β can be obtained via the estimating equation

$$S_n^{(w)}(\beta) = \sum_{i=1}^n \sum_{k=1}^{K_i} w_{ik}(\beta) \Delta_{ik} \left[X_{ik} - \frac{\sum_{j=1}^n \sum_{l=1}^{K_j} \hat{W}_j X_{jl} I\{e_{ik}(\beta) \leq e_{jl}(\beta)\}}{\sum_{j=1}^n \sum_{l=1}^{K_j} \hat{W}_j I\{e_{ik}(\beta) \leq e_{jl}(\beta)\}} \right], \quad (3.6)$$

where $w_{ik}(\cdot)$ are nonnegative weight functions ($k = 1, \dots, K_i, i = 1, \dots, n$). In general, (3.6) differs from the efficient score function for estimating β . Hence, if

one replaces the maximization of (3.4) at a given stage of the EM algorithm with estimates for $\Lambda_0(\cdot)$ and β derived from (3.5) and (3.6) using $\hat{W}_j = E(W_j|\mathbf{O}, \hat{\psi}^{(s)})$, the resulting procedure is no longer a true EM algorithm but rather an example of an Expectation and Substitution (ES) algorithm (Elashoff and Ryan, 2004).

Setting $\hat{W}_j = 1$ ($j = 1, \dots, n$), it can be seen that (3.6) reduces to the estimating equation for β under the marginal independence approach (Lee et al., 1993) and, with $K_j = 1$ ($j = 1, \dots, n$), to the estimating function of Tsiatis (1990) given in (1.1). The fact that β only appears in (3.6) as an argument to indicator functions means that $S_n^{(w)}(\beta)$ is not a continuous function of β ; hence, a solution to $S_n^{(w)}(\beta) = 0$ typically does not exist. Parameter estimates may instead be obtained by minimizing $\|S_n^{(w)}(\beta)\|$, where $\|v\|$ denotes $(v'v)^{1/2}$ for a vector v . However, this minimization problem may admit several solutions. In addition, because $S_n^{(w)}(\beta)$ is not necessarily monotone in β , the resulting set of solutions may not be a convex set.

For the setting in which $\hat{W}_j = K_j = 1$ ($j = 1, \dots, n$), Fyngson and Ritov (1994) note that use of the Gehan weight function $w_i(\beta) = \sum_{j=1}^n I\{e_i(\beta) \leq e_j(\beta)\}$ leads to a discontinuous but monotone estimating equation. Following Strawderman (2006), substitution of the modified weights $w_{ik}(\beta) = \sum_{j=1}^n \sum_{l=1}^{K_j} \hat{W}_j I\{e_{ik}(\beta) \leq e_{jl}(\beta)\}$ into (3.6) leads to

$$S_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \hat{W}_j (X_{ik} - X_{jl}) I\{e_{ik}(\beta) - e_{jl}(\beta) \leq 0\}; \quad (3.7)$$

see Zhang and Peng (2007) and Xu and Zhang (2010) for related developments. The estimating equation (3.7) is monotone in each component of β and, importantly, equals the gradient of the convex objective function

$$L_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \hat{W}_j \{e_{jl}(\beta) - e_{ik}(\beta)\} I\{e_{ik}(\beta) - e_{jl}(\beta) \leq 0\}. \quad (3.8)$$

Hence, a regression parameter estimate can be found by minimizing $L_n(\beta)$ with

respect to β (e.g., Jin et al., 2006a). The resulting set of solutions forms a convex set; however, a unique minimizer may not exist. As above, upon setting $\hat{W}_j = 1$ ($j = 1, \dots, n$), (3.7) and (3.8) reduce to the Gehan estimating equation (2.1) and objective function (2.2) for clustered data under the marginal independence approach. This relationship between the sets of equations for the two approaches will be evident throughout the remainder of this chapter and will be reflected by the use of parallel notation.

Induced smoothing for estimation of β

In general, the ES algorithm based on (3.6) presents one solution to the inability to maximize (3.4). However, the well-known computational challenges summarized in the previous section continue to present barriers for implementation, even in the case where $L_n(\beta)$ is convex. Relevant examples of algorithms that attempt to cope with these challenges include those described in Pan (2001), Jin et al. (2003, 2006a), Strawderman (2006), Zhang and Peng (2007), and Xu and Zhang (2010). None make use of smoothing to ease the computational burden. In light of the connection of equations (3.7) and (3.8) to the problem of estimating β under a marginally specified semiparametric AFT regression model, we propose to incorporate a simple adaptation of the smoothing procedure introduced in §2.1.3 into the ES algorithm.

Define Z to be a $N(0, I_p)$ random vector independent of the data, where I_p denotes the $p \times p$ identity matrix. Let Γ be a $p \times p$ matrix such that $\|\Gamma\| = O(1)$ and $\Gamma^2 = \Sigma$, where Σ is some symmetric, positive definite matrix. Then, a smoothed estimating equation may be constructed by adding the random perturbation $n^{-1/2}\Gamma Z$ to the argument of $S_n(\beta)$ in (3.7) and taking its expectation with respect

to Z . Specifically, setting $\tilde{S}_n(\beta) = E_Z \{S_n(\beta + n^{-1/2}\Gamma Z)\}$, we obtain

$$\tilde{S}_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \hat{W}_j (X_{ik} - X_{jl}) \Phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \right\} \right], \quad (3.9)$$

where $r_{ikjl}^2 = (X_{ik} - X_{jl})' \Sigma (X_{ik} - X_{jl})$ and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Brown and Wang (2005) proposed this technique of smoothing as a way to compute standard errors of general rank-based estimating equations and briefly discussed its ‘‘pseudo-Bayesian’’ motivation. More obviously, the estimating equation (3.9) may be viewed as a kernel-smoothed version of (3.7) in which the indicator function is replaced by a monotone kernel function (i.e., the standard normal CDF) that uses a pair-dependent bandwidth.

Instead of an estimating equation, one can instead work directly with the smoothed objective function $\tilde{L}_n(\beta) = E_Z \{L_n(\beta + n^{-1/2}\Gamma Z)\}$. Let $\phi(\cdot)$ denote the standard normal density function. Using well-known results for normal random variables and integration by parts, it can be shown that

$$\tilde{L}_n(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \hat{W}_j \left[\{e_{jl}(\beta) - e_{ik}(\beta)\} H_{ikjl}^{(n)}(\beta) + \frac{r_{ikjl}}{n^{1/2}} h_{ikjl}^{(n)}(\beta) \right], \quad (3.10)$$

where r_{ikjl} is defined above,

$$H_{ikjl}^{(n)}(\beta) = \Phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \right\} \right] \quad \text{and} \quad h_{ikjl}^{(n)}(\beta) = \phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \right\} \right].$$

A straightforward calculation shows that $\nabla \tilde{L}_n(\beta) = \tilde{S}_n(\beta)$. Under mild conditions on the covariates, the smoothed objective function $\tilde{L}_n(\beta)$ is strictly convex and infinitely continuously differentiable; hence, standard numerical methods can be used to efficiently compute its unique minimizer $\hat{\beta} = \operatorname{argmin}_{\beta} \tilde{L}_n(\beta)$.

To obtain marginal regression parameter estimates for the AFT model with clustered data, we considered several choices for the smoothing matrix Σ in §2.3, including both $\Sigma = I_p$ and a data-dependent smoothing matrix that is computed

iteratively and which reflects the relative scaling of the regression parameters. It is not clear how a similar data-dependent Σ might be constructed in the AFT frailty model. In addition, we found that the choice of Σ generally had minimal impact on the bias or standard error of the regression estimates in the marginal approach. We therefore propose to use $\Sigma = I_p$ in (3.10) in order to estimate β , exploring this choice and others in the simulation studies of §3.3.

SES algorithm

The joint estimation procedure for $\psi = (\theta, \beta, \Lambda_0)$ that results from incorporating the smoothed regression parameter estimator into the ES algorithm, hereafter referred to as the Smoothing Expectation and Maximization (SES) algorithm, can now be implemented as follows:

1. Select initial values:

Initialize $\hat{W}_i^{(0)} = 1$ ($i = 1, \dots, n$).

Initialize $\hat{\beta}^{(0)}$ by minimizing $\tilde{L}_n(\beta)$ in (3.10) with $\Sigma = I_p$.

Initialize $\hat{\Lambda}_0^{(0)}$ using (3.5).

Initialize $\hat{\theta}^{(0)}$.

Set $s = 1$.

2. E-step:

Compute the update $\hat{W}_i^{(s)} = E(W_i | \mathbf{O}, \hat{\psi}^{(s-1)})$ ($i = 1, \dots, n$).

3. S-step:

Update $\hat{\beta}^{(s)}$ by minimizing $\tilde{L}_n(\beta)$ in (3.10) with $\Sigma = I_p$.

Update $\hat{\Lambda}_0^{(s)}$ using (3.5).

Update $\hat{\theta}^{(s)}$ by maximizing $\ell(\theta) = E\{L_1(\theta) | \mathbf{O}, \hat{\psi}^{(s)}\}$ in (3.3).

Set $s = s + 1$.

4. Iterate between steps 2 and 3 until a specified convergence criterion is met and report $\hat{\psi} = (\hat{\beta}, \hat{\theta}, \hat{\Lambda}_0)$.

A remark on notation is needed. When calculating $\hat{W}_i^{(s)} = E(W_i|\mathbf{O}, \hat{\psi}^{(s-1)})$ in Step 2, the expectation is considered to be a function of ψ and is evaluated at $\psi = \hat{\psi}^{(s-1)}$, where $\hat{\psi}^{(s-1)} = (\hat{\theta}^{(s-1)}, \hat{\beta}^{(s-1)}, \hat{\Lambda}_0^{(s-1)})$ is the most recently computed iterate. However, when calculating $E\{L_1(\theta)|\mathbf{O}, \hat{\psi}^{(s)}\}$ in Step 3, all instances of θ in $L_1(\theta)$ are left to vary freely and all instances of β and $\Lambda_0(\cdot)$ are replaced by the most recently computed values $\hat{\beta}^{(s)}$ and $\hat{\Lambda}_0^{(s)}(\cdot)$. As can be seen from the sequence of steps, $\hat{\beta}^{(s)}$ and $\hat{\Lambda}_0^{(s)}(\cdot)$ also indirectly depend on $\hat{\theta}^{(s-1)}$ through $\hat{W}_1^{(s)}, \dots, \hat{W}_n^{(s)}$ calculated in Step 2.

In the SES algorithm, the choice of frailty distribution directly impacts the calculation of $\hat{W}_i^{(s)} = E(W_i|\mathbf{O}, \hat{\psi}^{(s)})$ in Step 2 and the calculation, hence maximization, of $\ell(\theta) = E\{L_1(\theta)|\mathbf{O}, \hat{\psi}^{(s)}\}$ in Step 3. It follows that EM algorithms previously proposed for the proportional hazards regression model with a specific frailty distribution can be adapted to the current setting; examples include Nielsen et al. (1992) and Klein (1992), who consider the gamma frailty distribution, and Wang et al. (1995), who develop an EM algorithm for the positive stable proportional hazards frailty model. Indeed, one may choose any frailty distribution, the primary limitation being a computationally feasible characterization of the conditional expectations needed in Steps 2 and 3 of the SES algorithm.

If $\mathcal{L}(s|\tilde{\theta})$ denotes the Laplace transform of W when the true parameter is $\tilde{\psi} = (\tilde{\theta}, \tilde{\beta}, \tilde{\Lambda}_0)$ and $\mathcal{L}^{(r)}(t|\tilde{\theta})$ denotes its r th derivative with respect to s , then it can be shown that

$$E(W_i|\mathbf{O}, \tilde{\psi}) = \frac{\mathcal{L}^{(D_i+1)}(\tilde{H}_i|\tilde{\theta})}{\mathcal{L}^{(D_i)}(\tilde{H}_i|\tilde{\theta})} \quad (3.11)$$

where $\tilde{H}_i = \sum_{k=1}^{K_i} \tilde{\Lambda}_0(e_{ik}(\tilde{\beta}))$ (Aalen et al., 2008, §7.2.3). For this and other reasons, Aalen et al. (2008, §6.2.3) state that useful choices of frailty distributions should minimally have Laplace transforms that exist in closed form. One example is the class of power variance function distributions which includes the gamma, inverse Gaussian, and positive stable distributions as special cases (Hougaard, 2000). Another is the class of generalized inverse Gaussian distributions which also includes both the gamma and inverse Gaussian distributions (Jørgensen, 1982; Aalen et al., 2008). The lognormal distribution, though a perfectly valid choice, does not have a Laplace transform that exists in closed form. As a result, $\mathcal{L}(\cdot|\tilde{\theta})$, and various related quantities (e.g., (3.11)), must be approximated numerically.

Unfortunately, the existence of a Laplace transform in closed form is insufficient to ensure the availability of a useful algorithm, for this does not guarantee that $\ell(\theta)$ in Step 3 is easily computed. Aalen et al. (2008, §7.2.5) discuss the special nature of the gamma distribution in this regard and, abstracting that discussion, describe two other classes of distributions considered suitable for shared frailty models: the generalized inverse Gaussian and Kummer distribution families. We provide the necessary implementation details for the SES algorithm in the case of the gamma frailty distribution in §3.2.1; our algorithm may be compared with that of Pan (2001), Zhang and Peng (2007), and Xu and Zhang (2010). In §3.2.2, we discuss implementation in the case of the inverse Gaussian frailty distribution.

We close this subsection by noting that there exist several possibilities for initializing $\hat{\theta}^{(0)}$ in Step 1. For example, $\hat{\theta}^{(0)}$ may be set arbitrarily, or one may attempt to employ (3.3). In §3.1.4, we introduce several other possibilities for estimating θ derived from novel moment identities. The resulting estimators

can be used to estimate $\hat{\theta}^{(0)}$. In addition, one could also employ these estimators in lieu of the profile likelihood estimator in Step 3 of the SES algorithm.

Remark 3.1.1 *An approach advocated in both Nielsen et al. (1992) and Wang et al. (1995) in the case of the proportional hazards model is to fix $\hat{\theta}^{(0)}$ and then run a version of the above algorithm in which Step 3 is modified in a way that sets $\hat{\theta}^{(s)} = \hat{\theta}^{(0)}$ at every iteration and computes the MLE of both β and $\Lambda_0(\cdot)$. This procedure is then repeated for a grid of $\hat{\theta}^{(0)}$ values. The parameter set leading to the largest value of the profiled observed data likelihood function (i.e., in θ) is then selected as the approximate MLE. In principle, variance estimates can be obtained by calculating, or otherwise approximating, the Hessian matrix for the marginal log-likelihood function. Pan (2001) considers a related idea in the context of the AFT gamma frailty model. However, it is important to note that the proposed procedure does not yield an approximate MLE since β is not estimated using the efficient score function. The failure to use the efficient score also complicates variance estimation since one also cannot numerically differentiate the log-likelihood function. The SES algorithm as presented above, combined with the bootstrap methodology described in §3.1.3 below, provides a simple (if computationally intensive) method of variance estimation for all model parameters.*

3.1.3 Variance estimation

Variance estimation under the semiparametric AFT frailty model is a challenging problem. In the independent data setting, the induced smoothing procedure of §2.1.3 leads to a natural sandwich variance estimator for the regression parameters. This is not the case when the smoothing procedure is incorporated as part of the SES algorithm and standard errors for the regression parameters, frailty parameter and baseline cumulative hazard are all of interest.

Pan (2001), Zhang and Peng (2007), and Xu and Zhang (2010) each propose estimating standard errors in the AFT gamma frailty model using the bootstrap. In this case, the random cluster-level weights are exchangeable but not independent. Ma and Kosorok (2005) propose the randomly weighted bootstrap, using independent weights, for computing variances of semiparametric M-estimators. Strawderman (2006) adapts this procedure to the problem of variance estimation in an accelerated gap times gamma frailty model proposed for recurrent event data. A similar procedure, to be described below, permits estimation of the variance of $\hat{\psi}$.

Let w_1, \dots, w_n be independent exponential random variables with mean one. Define $\bar{w}_i = w_i/\bar{w}$, so that $\sum_{i=1}^n \bar{w}_i = n$ and $\bar{w}_1, \dots, \bar{w}_n$ have a Dirichlet distribution, which Kosorok et al. (2004) suggest works well for semiparametric inference. Let \mathbb{P}_n denote the empirical distribution that assigns weight $1/n$ to each observation and let \mathbb{P}_n^* denote the corresponding weighted version that assigns the weight \bar{w}_i/n to each observation in cluster i ($i = 1, \dots, n$). Weighted versions of the equations $\tilde{L}_n(\cdot)$, $\hat{\Lambda}_0(\cdot)$, and $\ell(\cdot)$ used in Step 3 of the SES algorithm are obtained by writing (3.3), (3.5), and (3.10) in terms of \mathbb{P}_n , and then replacing \mathbb{P}_n by \mathbb{P}_n^* . In particular, for a given set of weights and with $\hat{\psi}_*^{(s)}$ denoting the current value of ψ , weighted versions of $\tilde{L}_n(\cdot)$, $\hat{\Lambda}_0(\cdot)$, and $\ell(\cdot)$ are respectively given by

$$\tilde{L}_n^*(\beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \bar{w}_i \sum_{k=1}^{K_i} \sum_{j=1}^n \bar{w}_j \sum_{l=1}^{K_j} \Delta_{ik} \hat{W}_j^{(s,*)} \left[\{e_{jl}(\beta) - e_{ik}(\beta)\} H_{ikjl}^{(n)}(\beta) + \frac{r_{ikjl}}{n^{1/2}} h_{ikjl}^{(n)}(\beta) \right], \quad (3.12)$$

$$\hat{\Lambda}_0^*(t) = \sum_{i=1}^n \bar{w}_i \sum_{k=1}^{K_i} \frac{\Delta_{ik}}{\sum_{j=1}^n \bar{w}_j \sum_{l=1}^{K_j} \hat{W}_j^{(s,*)} I\{e_{jl}(\beta) \geq e_{ik}(\beta)\}} I\{e_{ik}(\beta) \leq t\}, \quad (3.13)$$

and

$$\ell^*(\theta) = \sum_{i=1}^n \bar{w}_i \left[D_i E \{ \log W_i | \mathbf{O}, \hat{\psi}^{(s,*)} \} + K_i E \left\{ \log g(W_i | \theta) | \mathbf{O}, \hat{\psi}^{(s,*)} \right\} \right], \quad (3.14)$$

where r_{ikjl} , $H_{ikjl}^{(n)}(\cdot)$, and $h_{ikjl}^{(n)}(\cdot)$ are as defined in §3.1.2 and $\hat{W}_j^{(s,*)} = E(W_j | \mathbf{O}, \hat{\psi}^{(s,*)})$.

Each bootstrap replicate of $\hat{\psi}$, say $\hat{\psi}^*$, is obtained by generating a set of bootstrap weights and running the SES algorithm using $\hat{\psi}$ as the initial value and with $\tilde{L}_n(\cdot)$, $\hat{\Lambda}_0(\cdot)$, and $\ell(\cdot)$ replaced by the weighted versions given above. The estimated covariance matrix of $\hat{\theta}$, $\hat{\beta}$, and say, $\hat{\Lambda}_0(t_r)$ ($r = 1, \dots, d$), is now easily obtained using the empirical covariance matrix of the corresponding set of bootstrap replicates.

3.1.4 Moment-based estimation of θ

In this section, standard results from the martingale theory for counting processes are used to show how estimation of the frailty parameter θ may be carried out via the method of moments. For member k of cluster i , let $N_{ik}(u) = I(T_{ik} \leq u, \Delta_{ik} = 1)$ denote the event counting process and let $Y_{ik}(u) = I(T_{ik} \geq u)$ denote the at-risk process at time u . Let $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$ denote the full data filtration generated by these processes and frailties W_1, \dots, W_n , where independence is assumed across clusters; let the corresponding full data cumulative intensity be given by $W_i H_{ik}(t)$, where $H_{ik}(t) = \int_0^t Y_{ik}(u) d\Lambda_{ik}(u)$ is assumed to be continuous in t . Then, $M_{ik}(t) = N_{ik}(t) - W_i H_{ik}(t)$ is a mean zero martingale process in t under the filtration \mathcal{F} (Andersen et al., 1993, Ch. IX). The continuity of $H_{ik}(\cdot)$ and the independence of clusters further implies that the martingales $M_{ik}(\cdot)$ are orthogonal across k for each fixed i as well as across i . Also, for each i , $M_i(t) = N_i(t) - W_i H_i(t)$ is a mean zero martingale process in t under \mathcal{F} , where $M_i(t) = \sum_{k=1}^{K_i} M_{ik}(t)$, $N_i(t) = \sum_{k=1}^{K_i} N_{ik}(t)$, and $H_i(t) = \sum_{k=1}^{K_i} H_{ik}(t)$. These re-

sults are sufficient to establish the following moment identities.

Theorem 3.1.2 *Let $\tau > 0$ denote the largest observation time in the sample. Let $\mathcal{F}^{(O)}$ denote the corresponding observed data filtration (i.e., at time τ); that is, W_1, \dots, W_n are not assumed to be observed. Let $N_i = N_i(\tau)$, $H_i = H_i(\tau)$, and define $U_i = E(W_i|\mathcal{F}^{(O)})$, $V_i = E(W_i^2|\mathcal{F}^{(O)})$, and $Z_i := V_i - U_i^2 = \text{Var}(W_i|\mathcal{F}^{(O)})$. Then, for each i ($i = 1, \dots, n$), the following identities hold:*

$$E\{(N_i - U_i H_i)^2\} + E(Z_i H_i^2) = E(U_i H_i), \quad (3.15)$$

$$E\{(N_i - U_i H_i)^2\} + E(Z_i H_i^2) = E(N_i). \quad (3.16)$$

The proof of this result is provided in Appendix B, where it can be seen that (3.15) and (3.16) hold in great generality, relying only the fact that $M_{ik}(t) = N_{ik}(t) - W_i H_{ik}(t)$ are mean zero martingales under \mathcal{F} . In particular, these results are neither specific to the nature of the counting process nor the intensity model. Consequently, these identities provide a basis for deriving moment-based estimators of parameters in general univariate frailty models provided that a specific frailty distribution, hence $U_i = E(W_i|\mathcal{F}^{(O)})$ and $Z_i = \text{Var}(W_i|\mathcal{F}^{(O)})$, has been specified.

Let $\tilde{U}_i(\theta)$ and $\tilde{Z}_i(\theta)$ denote U_i and Z_i computed for a given frailty model, considered as a function of θ , assuming that H_{ik} is known for all i and k . Then, dropping the expectations in (3.15) and summing over clusters, θ can be estimated by solving the univariate nonlinear equation

$$\sum_{i=1}^n \left[\left\{ N_i - \tilde{U}_i(\theta) H_i \right\}^2 + \tilde{Z}_i(\theta) H_i^2 - \tilde{U}_i(\theta) H_i \right] = 0. \quad (3.17)$$

Proceeding similarly, but using relationship (3.16), we obtain the alternative nonlinear equation

$$\sum_{i=1}^n \left[\left\{ N_i - \tilde{U}_i(\theta) H_i \right\}^2 + \tilde{Z}_i(\theta) H_i^2 - N_i \right] = 0. \quad (3.18)$$

Equations (3.17) and (3.18) can be used to estimate θ in a general intensity model. For the model of §3.1.1, $H_{ik} = \Lambda_0(\epsilon_{ik}) = \Lambda_0(e_{ik}(\beta))$; hence, $H_i = \sum_{k=1}^{K_i} \Lambda_0(e_{ik}(\beta))$. Replacing N_i by $D_i = \sum_{k=1}^{K_i} \Delta_{ik}$ and H_i by a suitable estimate, the respective solutions to (3.17) and (3.18) provide two possible estimators for θ .

A third estimator of θ can be constructed using the generalized method of moments (GMM). Let $r_1(\theta)$ and $r_2(\theta)$ denote the left-hand side of the equations in (3.17) and (3.18), with N_i replaced by D_i and H_i replaced by an estimator, \hat{H}_i . Combining these two estimating equations into a vector, say $r(\theta) = (r_1(\theta), r_2(\theta))'$, a GMM estimator of θ can be defined as the minimizer of the quadratic form $r'(\theta)Q(\theta)r(\theta)$ where Q is a positive definite weighting matrix (Hansen, 1982). A suitable choice for Q is the inverse of the empirical covariance matrix constructed from the elements of $r(\theta)$, in which case one estimates θ by minimizing $r'(\theta)\hat{V}^{-1}(\theta)r(\theta)$ (Hansen et al., 1996).

3.2 Implementation

3.2.1 Gamma frailty distribution

As discussed in §3.1.2, a common and computationally convenient choice for the frailty distribution is the gamma distribution with shape parameter θ and scale parameter θ^{-1} (i.e., mean one and variance θ^{-1}). In this case, the joint survival function for ϵ_{ik} ($k = 1, \dots, K_i$) is (e.g., Klein, 1992)

$$P(\epsilon_{ik} > e_{ik}, k = 1, \dots, K_i) = \left\{ 1 + \theta^{-1} \sum_{k=1}^{K_i} \Lambda_0(e_{ik}) \right\}^{-\theta} \quad (3.19)$$

and the within-cluster association measured by Kendall's τ is $1/(1 + 2\theta)$; the strength of the association is monotone decreasing in θ with $\theta = \infty$ correspond-

ing to independent observations. In this section, we provide the necessary details for implementing the SES algorithm, the weighted bootstrap for standard error estimation and the moment-based estimators of θ in the gamma frailty setting.

The choice of frailty distribution impacts the implementation of the SES algorithm through the update calculation $\hat{W}_i^{(s)} = E(W_i|\mathbf{O}, \hat{\psi}^{(s-1)})$ in Step 2 and through the maximization of $\ell(\theta) = E\{L_1(\theta)|\mathbf{O}, \hat{\psi}^{(s)}\}$ in Step 3. Under the indicated parameterization, it is well known that the conditional distribution of W_i given \mathbf{O} is also a gamma distribution, but with respective shape and scale parameters A_i and C_i^{-1} , where

$$A_i = \theta + D_i \text{ and } C_i = \theta + \sum_{k=1}^{K_i} \Lambda_0(e_{ik}(\beta)), \quad (3.20)$$

assuming that the true parameter is $\psi = (\theta, \beta, \Lambda_0)$. Consequently, $E(W_i|\mathbf{O}, \hat{\psi}^{(s)}) = \hat{A}_i^{(s)}/\hat{C}_i^{(s)}$ and $E(\log W_i|\mathbf{O}, \hat{\psi}^{(s)}) = \Psi(\hat{A}_i^{(s)}) - \log \hat{C}_i^{(s)}$ where $\Psi(\cdot)$ denotes the digamma function and $\hat{A}_i^{(s)}$ and $\hat{C}_i^{(s)}$ denote the quantities in (3.20) computed assuming that $\psi = \hat{\psi}^{(s)}$. As a result, the update calculation in Step 2 becomes $\hat{W}_i^{(s)} = \hat{A}_i^{(s)}/\hat{C}_i^{(s)}$ and (3.3) immediately simplifies to

$$\ell(\theta) = -n \{-\theta \log \theta + \log \Gamma(\theta)\} + \sum_{i=1}^n \left[\{\theta + D_i - 1\} \{\Psi(\hat{A}_i^{(s)}) - \log \hat{C}_i^{(s)}\} - \theta \hat{A}_i^{(s)} / \hat{C}_i^{(s)} \right], \quad (3.21)$$

a univariate function of $\theta > 0$ that can be easily maximized in Step 3.

In §3.1.3, we proposed a method for standard error estimation in which equations (3.12) – (3.14) are used in the proposed SES algorithm. We have already shown how to compute $\hat{W}_i^{(s)}$; the calculation of $\hat{W}_i^{(s,*)}$ in (3.12) and (3.13) is completely analogous. In view of (3.21), it further follows that (3.14) reduces to

$$\ell^*(\theta) = -n \{-\theta \log \theta + \log \Gamma(\theta)\} + \sum_{i=1}^n \bar{w}_i \left[\{\theta + D_i - 1\} \{\Psi(\hat{A}_i^{(s,*)}) - \log \hat{C}_i^{(s,*)}\} - \theta \hat{A}_i^{(s,*)} / \hat{C}_i^{(s,*)} \right]$$

in the gamma frailty setting.

Finally, we provide details on the moment-based estimators of θ introduced in §3.1.4 for the gamma frailty setting. The conditional distribution of W_i given the observed data is $\text{Gamma}(\theta + D_i, \theta + H_i)$, where $H_i = \sum_{j=1}^{K_i} \Lambda_0(e_{ik}(\beta))$. Hence,

$$\tilde{U}_i(\theta) = \frac{\theta + D_i}{\theta + H_i} \quad \text{and} \quad \tilde{Z}_i(\theta) = \frac{\theta + D_i}{(\theta + H_i)^2}$$

and (3.17) and (3.18) reduce to the univariate estimating equations

$$\sum_{i=1}^n \left\{ \left(D_i - \frac{\theta + D_i}{\theta + \hat{H}_i} \hat{H}_i \right)^2 + \frac{\theta + D_i}{(\theta + \hat{H}_i)^2} \hat{H}_i^2 - \frac{\theta + D_i}{\theta + \hat{H}_i} \hat{H}_i \right\} = 0 \quad (3.22)$$

and

$$\sum_{i=1}^n \left\{ \left(D_i - \frac{\theta + D_i}{\theta + \hat{H}_i} \hat{H}_i \right)^2 + \frac{\theta + D_i}{(\theta + \hat{H}_i)^2} \hat{H}_i^2 - D_i \right\} = 0, \quad (3.23)$$

where \hat{H}_i denotes a suitable estimate of H_i . Similarly, in the GMM estimator, we respectively take $r_1(\theta)$ and $r_2(\theta)$ to be the left-hand side of the equations in (3.22) and (3.23).

These moment-based estimating equations, as well as (3.21), require a suitable estimate of $H_i = \sum_{j=1}^{K_i} \Lambda_0(e_{ik}(\beta))$. In general, under the AFT frailty model, one may use $\hat{H}_i = \sum_{k=1}^{K_i} \hat{\Lambda}_0(e_{ik}(\hat{\beta}))$, with $\hat{\Lambda}_0(\cdot)$ given by (3.5). An alternative estimator of H_i in the case of the gamma frailty model can be constructed as follows. Recalling that the joint survival function of the error terms can be expressed as in (3.19), and setting $e_{i1} = t$ and $e_{ik} = 0$ for $k = 2, \dots, K_i$, the marginal survival function can be written $S(t) = \{1 + \frac{1}{\theta} \Lambda_0(t)\}^{-\theta}$. It follows that the marginal cumulative hazard is $\Lambda(t) = \theta \log \{ \frac{1}{\theta} \Lambda_0(t) + 1 \}$ and, solving for the baseline cumulative hazard,

$$\Lambda_0(t) = \theta \left[\exp \left\{ \frac{1}{\theta} \Lambda(t) \right\} - 1 \right].$$

Hence, one can instead estimate $\Lambda(\cdot)$ and use the corresponding plug-in estimator for $\Lambda_0(\cdot)$. For highly stratified censored data (i.e., clustered data in which the

cluster sizes are small relative to the number of clusters), and given a consistent estimate of β , the results of Ying and Wei (1994) suggest that the marginal cumulative hazard $\Lambda(\cdot)$ can be consistently estimated by equation (3.5) with $W_j = 1$ ($j = 1, \dots, n$), leading to the estimator

$$\hat{\Lambda}_0^\dagger(t) = \theta \left[\exp \left\{ \frac{1}{\theta} \hat{\Lambda}(t) \right\} - 1 \right]. \quad (3.24)$$

One can then estimate H_i using $\hat{H}_i^\dagger = \sum_{k=1}^{K_i} \hat{\Lambda}_0^\dagger(e_{ik}(\hat{\beta}))$.

3.2.2 Inverse Gaussian frailty distribution

In this section, we assume that the frailties follow an inverse Gaussian distribution with mean one and variance θ^{-1} and provide a development of the requirements for implementing the SES algorithm in this setting. As previously discussed, implementation of the SES algorithm for a particular frailty distribution affects the update calculation $\hat{W}_i^{(s)} = E(W_i | \mathbf{O}, \hat{\psi}^{(s-1)})$ in Step 2 and the calculation and maximization of $\ell(\theta) = E\{L_1(\theta) | \mathbf{O}, \hat{\psi}^{(s)}\}$ from equation (3.3) in Step 3; results needed to carry out these computations for the inverse Gaussian frailty model are reviewed below.

Following Jørgensen (1982), a random variable B is said to have a generalized inverse Gaussian distribution if its probability density function is given by

$$h(b | \varphi_1, \varphi_2, \varphi_3) = \frac{\left(\frac{\varphi_3}{\varphi_2}\right)^{\frac{\varphi_1}{2}}}{2\mathcal{K}_{\varphi_1}(\sqrt{\varphi_2\varphi_3})} b^{\varphi_1-1} \exp\{-(\varphi_2 b^{-1} + \varphi_3 b)/2\}, \quad b > 0,$$

where $\mathcal{K}_{\varphi_1}(\cdot)$ denotes the modified Bessel function of the second kind with index φ_1 . We say $B \sim \text{GIG}(\varphi_1, \varphi_2, \varphi_3)$, where the density is well-defined for $\varphi_1 \in \mathbb{R}$ and $\varphi_j > 0$ for $j = 2, 3$; if $\varphi_1 > 0$ (< 0), then it is possible for φ_2 (φ_3) to be equal to zero. This class of distributions contains the Gamma ($\varphi_1 > 0, \varphi_2 = 0$),

reciprocal Gamma ($\varphi_1 < 0, \varphi_3 = 0$), inverse Gaussian ($\varphi_1 = -\frac{1}{2}$) and reciprocal inverse Gaussian ($\varphi_1 = \frac{1}{2}$) distributions as special cases. The Laplace transform of the distribution of B exists in closed form and is given by

$$\mathcal{L}(t) = \frac{\mathcal{K}_{\varphi_1} \left(\sqrt{\varphi_2 \varphi_3 (1 + 2\varphi_3^{-1}t)} \right)}{\mathcal{K}_{\varphi_1} \left(\sqrt{\varphi_2 \varphi_3} \right) (1 + 2\varphi_3^{-1}t)^{\frac{\varphi_1}{2}}}.$$

The assumption that the frailties follow an inverse Gaussian distribution with mean one and variance θ^{-1} is equivalent to asserting that $W_i \sim \text{GIG}(-\frac{1}{2}, \theta, \theta)$, ($i = 1, \dots, n$) (see Jørgensen (1982, §2.1)). Importantly, using results in Aalen et al. (2008, §7.2.5), it can be shown that $W_i | \mathbf{O} \sim \text{GIG}(D_i - \frac{1}{2}, \theta, \theta + 2H_i)$, where, as indicated previously, $H_i = \sum_{k=1}^{K_i} \Lambda_0(e_{ik}(\beta))$ for the AFT frailty model considered in this chapter. For $\theta > 0$ and $H_i \geq 0$, results in Jørgensen (1982, §2.1) now imply that

$$E(W_i^r | \mathbf{O}, \psi) = \frac{\mathcal{K}_{D_i - \frac{1}{2} + r}(\zeta_i(\theta))}{\mathcal{K}_{D_i - \frac{1}{2}}(\zeta_i(\theta))} (1 + 2\theta^{-1}H_i)^{-\frac{r}{2}}, \quad r \in \mathbb{R} \quad (3.25)$$

where $\zeta_i(\theta) = \sqrt{\theta(\theta + 2H_i)}$. The required formula for the update calculation in Step 2 of the SES algorithm is obtained from (3.25) with $r = 1$. In particular, if $\hat{\psi}^{(s-1)}$ denotes the most recently computed estimate and $\hat{\zeta}_i^{(s-1)} = \sqrt{\hat{\theta}^{(s-1)}(\hat{\theta}^{(s-1)} + 2\hat{H}_i^{(s-1)})}$, then

$$E(W_i | \mathbf{O}, \hat{\psi}^{(s-1)}) = \frac{\mathcal{K}_{D_i + \frac{1}{2}}(\hat{\zeta}_i^{(s-1)})}{\mathcal{K}_{D_i - \frac{1}{2}}(\hat{\zeta}_i^{(s-1)})} \left(1 + 2\hat{H}_i^{(s-1)} / \hat{\theta}^{(s-1)} \right)^{-\frac{1}{2}}.$$

Turning to Step 3, the assumption that W_i has an inverse Gaussian distribution with mean one and variance θ^{-1} implies that the frailty density $g(w|\theta)$ equals $h(w | -\frac{1}{2}, \theta, \theta)$. Using the fact that $\mathcal{K}_{-\frac{1}{2}}(\theta) \propto \theta^{-1/2}e^{-\theta}$, the relevant portion of the complete data log-likelihood reduces to

$$L_1(\theta) = c + \sum_{i=1}^n \left(D_i - \frac{3}{2}K_i \right) \log W_i + \frac{K_i}{2} \log \theta + K_i \theta - \frac{\theta K_i}{2} (W_i^{-1} + W_i)$$

where c is a constant independent of θ and W_1, \dots, W_n . In the SES algorithm, one can update θ by maximizing $\ell(\theta) = E\{L_1(\theta)|\mathbf{O}, \hat{\psi}^{(s)}\}$, which requires a formula for $E(\log W_i|\mathbf{O}, \psi)$; see Jørgensen (1982, §3.1). Alternatively, one can instead update θ by solving the equation $E\{S_1(\theta)|\mathbf{O}, \psi\} = 0$ for θ , where $S_1(\theta) = \frac{d}{d\theta}L_1(\theta)$ is the complete data score function. In this case, and with $m = \sum_{i=1}^n K_i$ denoting the total sample size,

$$S_1(\theta) = \frac{m}{2\theta} + m - \frac{1}{2} \sum_{i=1}^n K_i (W_i^{-1} + W_i).$$

Consequently, the solution $\tilde{\theta}$ to $E\{S_1(\theta)|\mathbf{O}, \psi\} = 0$ exists in closed form and equals

$$\tilde{\theta} = \left[\frac{1}{m} \sum_{i=1}^n K_i \{E(W_i^{-1}|\mathbf{O}, \psi) + E(W_i|\mathbf{O}, \psi) - 2\} \right]^{-1},$$

an expression that can be evaluated directly using (3.25). Such an approach is advocated in Karlis (2001), who considers estimation in Poisson regression models in the presence of an inverse Gaussian mixing distribution. Equation (3.25) can also be used to derive moment estimators for θ in §3.1.4.

3.3 Simulation Study

Simulation studies were carried out to assess the performance of the SES algorithm in the case of a gamma frailty distribution. Also included is an evaluation of the corresponding moment estimators (§3.1.4 and §3.2.1) and the weighted bootstrap variance estimator (§3.1.3 and §3.2.1). Unless otherwise specified, simulation results are based on 500 replications of $n = 50$ clusters of size $K_i = 4$ ($i = 1, \dots, n$). Simulations with one covariate consider a single continuous covariate generated from the uniform distribution on $(0, 1)$; simulations with two covariates include further choices described in §3.3.4. Frailties are generated from a gamma distribution with shape θ and scale θ^{-1} . The baseline hazard for

the error terms is assumed to be that of an extreme value distribution, $\lambda_0(s) = e^s$, and the censoring times are generated from a uniform distribution on $(0, a)$ where a is selected to achieve approximately 25% censoring. The true regression parameter is $\beta = 2$. When the true gamma frailty parameter is $\theta = 0.5$, the univariate simulation study matches that described in Table 1 of Zhang and Peng (2007) and Tables 1 and 2 of Xu and Zhang (2010). Several other choices of θ are considered here.

The stopping criteria used for assessing convergence of the SES algorithm is $\max\{D_1^{(s)}, D_2^{(s)}, D_3^{(s)}\} \leq 0.001$, where $D_1^{(s)} = \|\hat{\beta}^{(s)} - \hat{\beta}^{(s-1)}\|$, $D_2^{(s)} = \|\hat{\theta}^{(s)} - \hat{\theta}^{(s-1)}\|$, and $D_3^{(s)} = |\frac{1}{n} \sum_{i=1}^n (\hat{W}_i^{(s)} - \hat{W}_i^{(s-1)})|$. Since $\frac{1}{n} \sum_{i=1}^n \hat{W}_i^{(s)}$ should be close to 1.0 at convergence, the use of the mean difference in frailty estimates across successive iterations is both an absolute and a relative error criterion. In the interest of limiting the total time required for the various simulations, the SES algorithm was allowed to run for a maximum of 350 iterations. All simulations were conducted in R (R Development Core Team, 2005); the `nlm` routine was used for minimizing (3.10), the `optimize` routine was used for maximizing (3.21) and minimizing the GMM quadratic form, and the `uniroot` routine was used for solving (3.22) and (3.23). The simulation code is available upon request.

3.3.1 Estimator performance and impact of $\hat{\theta}^{(0)}$

Table 3.1 summarizes the results of implementing the SES algorithm using the four methods introduced for estimation of the gamma frailty parameter to compute the initial estimate of θ in Step 1 of the SES algorithm. PL denotes the profile likelihood of θ given by equation (3.21), while MM1 and MM2 denote the method of moments estimators for θ given by equations (3.22) and (3.23), and GMM denotes the generalized method of moments estimator. The determi-

Table 3.1: Frailty model simulation results

θ	Method	iterations	$\hat{\beta}$			$\hat{\theta}^{-1}$		$\hat{\Lambda}_0(-2)$		$\hat{\Lambda}_0(0)$		$\hat{\Lambda}_0(1.5)$	
			$\hat{\theta}^{(0)}$	Bias	MSE	RBias	RMSE	Bias	MSE	Bias	MSE	Bias	MSE
0.1	PL	181.19	9.9999	-0.0277	0.6682	0.0895	0.0385	0.0035	0.0113	-0.0483	0.4802	-0.3691	11.6802
	MM1	177.90	0.2254	-0.0275	0.6683	0.0894	0.0384	0.0035	0.0113	-0.0476	0.4805	-0.3671	11.6643
	MM2	177.98	0.6556	-0.0277	0.6680	0.0895	0.0385	0.0035	0.0113	-0.0486	0.4803	-0.3710	11.6759
	GMM	177.82	0.1749	-0.0275	0.6683	0.0894	0.0385	0.0035	0.0113	-0.0479	0.4804	-0.3684	11.6657
0.5	PL	44.75	9.9999	0.0015	0.2479	0.0188	0.0533	0.0065	0.0031	0.0526	0.1568	0.1781	4.2830
	MM1	40.17	0.7052	0.0015	0.2479	0.0188	0.0533	0.0065	0.0031	0.0526	0.1568	0.1777	4.2812
	MM2	40.81	1.5101	0.0015	0.2479	0.0187	0.0534	0.0065	0.0031	0.0528	0.1570	0.1788	4.2833
	GMM	40.13	0.6199	0.0015	0.2479	0.0188	0.0533	0.0065	0.0031	0.0526	0.1568	0.1780	4.2822
1	PL	35.98	9.9999	0.0195	0.1819	0.0506	0.0745	0.0051	0.0025	0.0375	0.1042	0.0358	2.9226
	MM1	29.27	1.5049	0.0195	0.1820	0.0506	0.0745	0.0051	0.0025	0.0375	0.1042	0.0360	2.9262
	MM2	30.96	2.7399	0.0195	0.1819	0.0506	0.0745	0.0051	0.0025	0.0375	0.1042	0.0359	2.9239
	GMM	28.74	1.2492	0.0195	0.1820	0.0506	0.0745	0.0051	0.0025	0.0375	0.1043	0.0359	2.9251
4	PL	57.88	9.9999	0.0072	0.1398	0.0361	0.2474	0.0039	0.0016	0.0156	0.0502	-0.0864	2.2506
	MM1	46.67	6.5010	0.0072	0.1398	0.0360	0.2474	0.0040	0.0016	0.0156	0.0502	-0.0863	2.2505
	MM2	53.94	8.0143	0.0072	0.1398	0.0361	0.2474	0.0039	0.0016	0.0156	0.0502	-0.0865	2.2503
	GMM	40.22	5.3031	0.0071	0.1398	0.0354	0.2471	0.0040	0.0016	0.0156	0.0503	-0.0854	2.2505

True parameter values: $\beta = 2$, $\Lambda_0(-2) = 0.1353$, $\Lambda_0(0) = 1$, $\Lambda_0(1.5) = 4.4817$

RBias (absolute relative bias) and RMSE (relative MSE) are reported for $\hat{\theta}^{-1}$ since four different values of θ are considered.

nation of an initial PL estimate (i.e., using (3.21)) is carried out by replacing $\theta^{(s)}$ in (3.20) with θ and maximizing (3.21) accordingly.

Four different values for the true gamma frailty parameter are considered, $\theta = 0.1, 0.5, 1$, and 4 , representing a decreasing order of within-cluster dependence. For each value of θ , the performance of the four initial estimation methods is evaluated using the same 500 datasets. The estimates of the variance of the frailty distribution, $\hat{\theta}^{-1}$, are assessed using the absolute relative bias and the relative mean squared error to account for the different values of θ . The average number of iterations required to reach convergence and the average initial estimate of the frailty parameter, $\hat{\theta}^{(0)}$, are also provided to allow a thorough comparison of the initial estimation procedures.

We begin with $\hat{\theta}^{(0)}$, where considerable variability in the quality of initial estimates is observed across the four estimation methods in Table 3.1. Use of PL in Step 1 of the proposed algorithm is problematic when initializing $\hat{W}_i^{(0)} = A_i^{(0)} / C_i^{(0)} \equiv 1$ ($i = 1, \dots, n$); since this initialization is equivalent to within-cluster independence (i.e., $\theta = \infty$), the initial estimate of θ derived from (3.21) always lies at the upper bound of the user-specified search re-

gion (here, (0.001, 10)). The method of moments estimators provide better initial estimates of θ than PL, with GMM clearly providing the least biased estimate. Notably, the estimators MM1 and MM2 used here employ the estimate $\hat{H}_i = \sum_{k=1}^{K_i} \hat{\Lambda}_0(e_{ik}(\hat{\beta}))$ obtained directly from equation (3.5) (i.e., initialized under independence, or $\theta = \infty$). In contrast, the GMM estimator uses the alternative estimate, $\hat{H}_i^\dagger = \sum_{k=1}^{K_i} \hat{\Lambda}_0^\dagger(e_{ik}(\hat{\beta}))$, obtained from equation (3.24). For MM1 and MM2, the estimator \hat{H}_i^\dagger was prohibitively large in some instances for small values of θ ; the independence estimator \hat{H}_i performed better. For GMM, implementation using either \hat{H}_i and \hat{H}_i^\dagger proved feasible (results not shown); however, the GMM estimator was observed to perform better using \hat{H}_i^\dagger .

The results in Table 3.1 further show that the estimates obtained at convergence are essentially the same regardless of the method used for determining $\hat{\theta}^{(0)}$. The biases of $\hat{\beta}$ and $\hat{\theta}^{-1}$ are both small with reasonable mean squared errors. The biases observed in $\hat{\beta}$ when $\theta = 0.5$ are also considerably smaller than those reported in Zhang and Peng (2007) and Xu and Zhang (2010). The mean squared errors for $\hat{\beta}$ are similar to those reported in Zhang and Peng (2007). Xu and Zhang (2010) do not report mean squared error results; instead, empirical standard errors are provided that are seemingly inconsistent with the mean squared errors reported in Zhang and Peng (2007). The biases and mean squared errors of $\hat{\theta}^{-1}$ are observed to be significantly smaller than those reported in both Zhang and Peng (2007) and Xu and Zhang (2010). Neither Zhang and Peng (2007) nor Xu and Zhang (2010) report simulation results for values of the frailty parameter other than $\theta = 0.5$.

In Table 3.1, the estimated baseline cumulative hazard, $\hat{\Lambda}_0(s)$, was evaluated at “times” (i.e., errors) of $s = -2, 0$, and 1.5 , corresponding to risk sets of approximately 90%, 60%, and 30% of the total sample when $\theta = 0.5$. As expected,

the bias and mean squared error increase as the size of the risk set decreases, the method used for determining $\hat{\theta}^{(0)}$ again observed to have minimal impact. For smaller risk sets, $\hat{\Lambda}_0(s)$ also tends to perform worse as θ decreases (i.e., dependence increases). Figure 3.1 displays the mean squared errors for $\hat{\Lambda}_0(s)$ using GMM for $s \in (-3, 2)$ when $\theta = 0.1, 0.5, 1,$ and 4 . Figure 3.2 displays the corresponding proportion of the sample still at risk over the interval $(-3, 2)$ for each value of θ . Similarly to Table 3.1, $\hat{\Lambda}_0(\cdot)$ performs well for all values of θ when the size of the risk set is large, the mean squared errors increasing as the size of the risk set decreases. At each value of s , the mean squared error of $\hat{\Lambda}_0(s)$ also increases as θ decreases; however, as seen in Figure 3.2, the size of the risk set is increasing with decreasing θ at each s , showing that these increases do not occur as a result of a decreasing risk set. We conjecture that both phenomena occur as a result of the increased variability in the W_i 's that occurs when θ decreases. Indeed, limited simulation results (not shown) confirm that increasing the variance of W_i 's tends to inflate the magnitude of both the true error terms (i.e., ϵ_{ik} 's) and estimated residuals, the asymmetry in the error term distribution amplifying this effect for $s > 0$. We suspect the increased variability in the W_i 's also induces the greater bias and/or variance in the \hat{W}_i 's, hence $\hat{\Lambda}_0(\cdot)$; however, the exact mechanism by which this occurs is less clear, for the bias in $\hat{\Lambda}_0(\cdot)$ is not necessarily monotone with decreasing θ (e.g., see Table 3.1 for $s = 1.5$).

We now comment briefly on computational issues, focusing on the case $\theta = 0.1$ and $\theta = 4$. When $\theta = 0.1$, a comparatively large number of iterations were required for the SES algorithm to reach convergence for each method of estimation. The increased effort observed here can be traced to slow convergence in $D_3^{(s)}$ (i.e., convergence of the estimated frailties) and is a likely consequence of the close proximity of θ to zero, an “extreme” case of dependence.

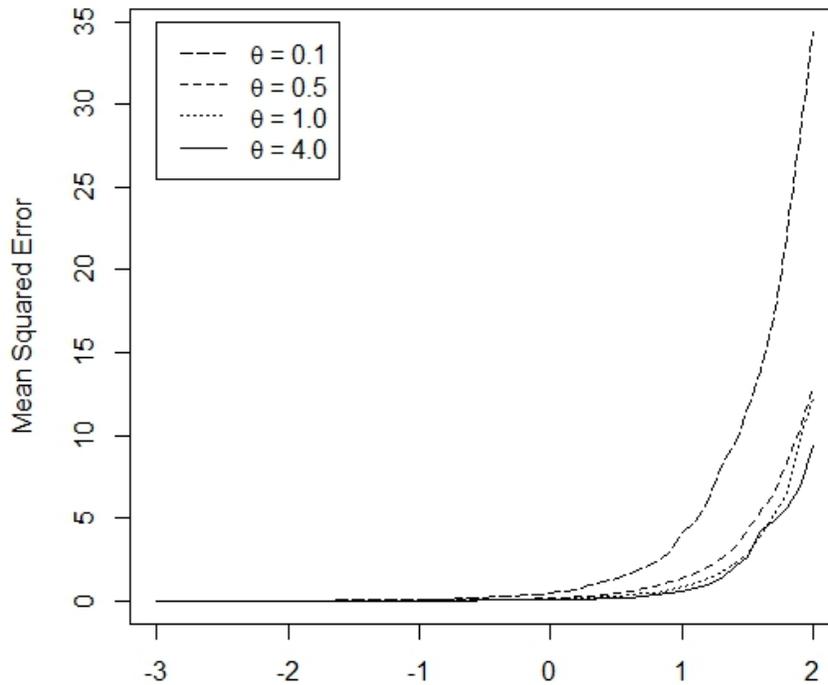


Figure 3.1: Mean squared error of the baseline cumulative hazard

Ultimately, convergence was obtained within 350 iterations in all cases and each method yields estimators that perform comparably at convergence. However, GMM clearly provides the best initial estimator of θ on average and (slightly) reduces the average total number of iterations required to reach convergence. When $\theta = 4$, the method of moments equations (3.22) and (3.23) often failed to cross zero within the specified search region (i.e., $(0.001, 10)$), yielding no solution. Increasing the search region did not alter this behavior (results not shown). In contrast, the GMM estimator yields the same solution as the PL estimator in such cases, namely the upper bound of the search region. Thus, in cases where MM1 or MM2 failed to provide a solution, we set the corresponding moment

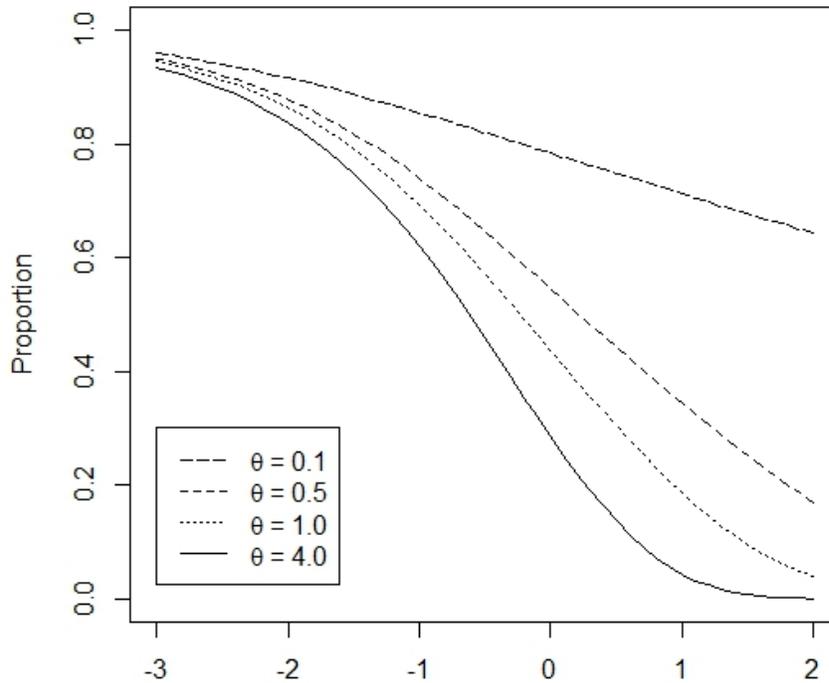


Figure 3.2: Proportion of sample at risk

estimator to the upper bound of the search region. Again, the GMM estimator provided the least biased initial estimator of θ and, in this case, requires substantially fewer iterations to reach convergence than PL, MM1, or MM2. Interestingly, and in contrast to $\theta = 0.1$, the rate of convergence here was primarily governed by β and/or θ , with the estimated frailties converging fairly quickly.

Overall, the results show that the use of smoothing leads to a simple estimator with excellent estimation performance, provided a reasonable method for initializing θ is used. The initialization of θ can, but doesn't always, have a moderate impact on computational efficiency. In general, the GMM estimator for $\hat{\theta}^{(0)}$ performs better than MM1, MM2, and PL for small and moderate values

of θ , and (at a minimum) no worse than MM1, MM2, and PL when θ is large, with the added advantage of reducing the total number of iterations needed for convergence without significantly increasing computational effort. Implementation of the MM1, MM2, and GMM estimators for the estimation of θ in the S-step of the SES algorithm (i.e., Step 3) was also considered. Not surprisingly, none of the moment based estimators outperformed the PL estimator (results not shown). On the basis of these results, we therefore recommend using the GMM estimator for deriving the initial estimate of θ and using the PL estimator for subsequent estimation of θ in the S-step. We employ this strategy in all subsequent simulation studies.

3.3.2 Impact of smoothing parameter choice

To obtain marginal regression parameter estimates for the AFT model with clustered data in §2.3, we considered several choices for the smoothing parameter Σ . They found that the choice of Σ generally had minimal impact on the bias or actual standard error of the regression estimates in the marginal approach. To assess the impact of the smoothing parameter on the SES algorithm in the AFT frailty model setting, we also explored two choices for selecting Σ . Method 1 refers to the SES algorithm exactly as stated in §3.1.2 with smoothing matrix $\Sigma = I_p$. Method 2 refers to applying the SES algorithm using a data-dependent smoothing matrix derived from the estimated variance of $\hat{\beta}$ that is obtained by fitting a marginal AFT model to the data using the iterative method of S 2.2.2.

Table 3.2 contains the results from applying Methods 1 and 2 to the AFT model with two covariates, X_{1ik} and X_{2ik} , generated from the uniform distribution on $(0, 1)$ and regression parameters $\beta_1 = 1$ and $\beta_2 = 0.5$. The simulations were repeated by fitting a model with covariates $X_{1ik}^\# = X_{1ik}$ and

Table 3.2: Impact of smoothing parameter

Parameters	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\theta}^{-1}$		$\hat{\Lambda}_0(-2)$		$\hat{\Lambda}_0(0)$		$\hat{\Lambda}_0(1.5)$	
	Bias	MSE	RBias	RMSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Method 1												
$\beta_1 = 1, \beta_2 = 0.5$	0.0047	0.2495	0.0440	0.8892	-0.0368	0.2020	0.0025	0.0037	0.0477	0.1960	0.0638	4.2923
$\beta_1 = 1, \beta_2 = 250$	0.0046	0.2497	0.0447	0.8863	-0.0373	0.2019	0.0022	0.0037	0.0463	0.1963	0.0635	4.3287
Method 2												
$\beta_1 = 1, \beta_2 = 0.5$	0.0070	0.2482	0.0416	0.8858	-0.0350	0.2035	0.0025	0.0037	0.0477	0.1947	0.0784	4.2879
$\beta_1 = 1, \beta_2 = 250$	0.0070	0.2482	0.0412	0.8824	-0.0351	0.2032	0.0024	0.0037	0.0479	0.1944	0.0803	4.2970

True parameter values: $\theta = 0.5$, $\Lambda_0(-2) = 0.1353$, $\Lambda_0(0) = 1$, $\Lambda_0(1.5) = 4.4817$

RBias (absolute relative bias) and RMSE (relative MSE) are reported for $\hat{\beta}_2$ since two different values of β_2 are considered.

Method 1 - fixed smoothing, $\Sigma = I_2$; Method 2 - data-dependent smoothing

$X_{2ik}^\# = X_{2ik}/500$; the underlying failure time model is identical, but the true regression parameters are now $\beta_1^\# = \beta_1 = 1$, $\beta_2^\# = 500$, and $\beta_2 = 250$. In particular, the magnitudes of $X_{2ik}^\#$ and $\beta_2^\#$ are quite different from those of $X_{1ik}^\#$ and $\beta_1^\#$. The intent here is to compare the impact of using a fixed smoothing parameter to a data-dependent choice that better reflects the different scaling of the regression parameters. The estimates $\hat{\beta}_2$ are assessed using the absolute relative bias and the relative mean squared error to account for the different values of β_2 . Evidently, Methods 1 and 2 perform equally well in both models. Consequently, we shall proceed from this point forward using the fixed smoothing matrix $\Sigma = I_p$.

3.3.3 Impact of censoring level, cluster size, and sample size

Table 3.3 explores the impact of censoring level, cluster size, and number of clusters on parameter estimation. The SES algorithm was run assuming 25% and 50% censoring. Clusters of size $K_i = 4$ and 8 were considered and the number of clusters was selected to be $n = 25, 50$, or 100 to result in total sample sizes of $N = 200$ and 400. As expected, when the censoring level increases, the mean squared errors of $\hat{\beta}$ and $\hat{\theta}^{-1}$ both increase and when the total sample size increases, the mean squared errors decrease. A similar pattern holds for the mean

Table 3.3: Impact of censoring level and cluster size

	$\hat{\beta}$			$\hat{\theta}^{-1}$			$\hat{\Lambda}_0(-2)$		$\hat{\Lambda}_0(0)$		$\hat{\Lambda}_0(1.5)$	
	Time	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	
25% Censoring												
$K_i = 4, n = 50$	8.6166	0.0015	0.2479	-0.0376	0.2131	0.0065	0.0031	0.0526	0.1568	0.1780	4.2822	
$K_i = 4, n = 100$	28.8497	-0.0018	0.1093	-0.0197	0.1024	-0.0013	0.0013	-0.0106	0.0594	-0.0317	1.5817	
$K_i = 8, n = 25$	10.4312	0.0015	0.1703	-0.0651	0.3406	0.0017	0.0029	0.0096	0.1346	-0.0284	3.9426	
$K_i = 8, n = 50$	37.2075	-0.0128	0.0929	-0.0030	0.1597	-0.0018	0.0016	-0.0048	0.0742	-0.0324	1.6869	
50% Censoring												
$K_i = 4, n = 50$	4.2127	0.0116	0.3031	-0.0082	0.2844	0.0085	0.0035	0.0652	0.1808	0.3098	9.4086	
$K_i = 4, n = 100$	13.7274	0.0003	0.1290	-0.0114	0.1357	-0.0002	0.0015	-0.0017	0.0694	-0.0074	2.5218	
$K_i = 8, n = 25$	5.1876	0.0053	0.2271	-0.0485	0.4301	0.0036	0.0032	0.0157	0.1444	-0.1376	4.7867	
$K_i = 8, n = 50$	17.5317	-0.0143	0.1180	0.0107	0.2045	-0.0009	0.0017	0.0007	0.0820	0.0148	2.4972	

True parameter values: $\beta = 2, \theta = 0.5, \Lambda_0(-2) = 0.1353, \Lambda_0(0) = 1, \Lambda_0(1.5) = 4.4817$
 Time = average time in seconds to convergence per simulation (s)

squared errors of $\hat{\Lambda}_0(\cdot)$, the greatest impact of increasing the censoring level being observed when the corresponding risk sets are smallest. Finally, increasing the cluster size while holding the number of clusters constant leads to a decrease in the mean squared error of $\hat{\theta}^{-1}$, reflecting the availability of greater information on within-cluster dependence. Notably, the biases and mean squared errors of $\hat{\beta}$ and $\hat{\theta}^{-1}$ observed when $K_i = 8, n = 50$ with 25% censoring are smaller than those reported in Zhang and Peng (2007) and Xu and Zhang (2010).

The average time required to reach convergence is also provided in Table 3.3. As expected, the time required increases as the total sample size increases. For the same total sample size, a larger cluster size and fewer clusters also requires longer to reach convergence. Although a larger cluster size reduces the number of \hat{W}_i computations necessary at each iteration, the reduced number of clusters results in a greater number of iterations required to reach convergence. It is important to note here that the proposed SES algorithm has not been optimized for speed. For example, the simulation code was entirely written in R and makes extensive use of R's built-in capabilities (e.g., instead of using optimization procedures designed specifically for convex differentiable objective functions). The proposed algorithm also uses a stringent but arbitrary criterion for assessing convergence and makes no attempt to leverage tools originally developed for

Table 3.4: Two covariate model

Second Covariate	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\theta}^{-1}$		$\hat{\Lambda}_0(-2)$		$\hat{\Lambda}_0(0)$		$\hat{\Lambda}_0(1.5)$	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Normal(0,1)	-0.0006	0.2457	-0.0016	0.0202	-0.0744	0.2231	-0.0021	0.0027	0.0169	0.1401	-0.0681	3.4609
Exponential(1)	0.0077	0.2366	-0.0082	0.0198	-0.0846	0.2207	0.0003	0.0033	0.0216	0.1740	0.0252	4.2876
Bernoulli(0.5)	0.0123	0.2590	-0.0100	0.0773	-0.0282	0.2133	0.0015	0.0031	0.0406	0.1540	0.0034	3.3066

True parameter values: $\beta_1 = 2, \beta_2 = -0.5, \theta = 0.5, \Lambda_0(-2) = 0.1353, \Lambda_0(0) = 1, \Lambda_0(1.5) = 4.4817$

the EM algorithm to accelerate convergence. As a result, the reported times, though useful for making relative comparisons within Table 3.3, should not be interpreted as an absolute measure of optimized performance.

3.3.4 Two covariate models

To evaluate the performance of the SES algorithm under different distributional assumptions for the covariates, a two covariate model was considered in which the first covariate has a uniform distribution on $(0, 1)$, as in the above simulations, and the second covariate has either a standard normal, exponential, or Bernoulli distribution. The true regression parameters are $\beta_1 = 2$ and $\beta_2 = -0.5$ and 25% censoring is assumed. Table 3.4 shows that the algorithm performs well for each of the two covariate models. The algorithm was also run for each of these models under the censoring, sample size, and cluster size configurations explored in Table 3.3 (results not shown). The algorithm performed well in these setting also, exhibiting patterns similar to those observed in Table 3.3. The average time required to reach convergence in each of these settings was approximately the same regardless of the distribution of the second covariate (results not shown); for each setting, there was a 1.5-2.0 fold increase in time required over the one covariate model on Table 3.3.

Table 3.5: Impact of true baseline and frailty distributions when fitting a gamma frailty model

Generating Distribution		$\hat{\beta}$		$\widehat{\text{Var}}(W_i)$		$\hat{\Lambda}_0(-2)$		$\hat{\Lambda}_0(0)$		$\hat{\Lambda}_0(1.5)$	
Baseline	True Frailty	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Extreme	Gamma	0.0015	0.2479	-0.0376	0.2131	0.0065	0.0031	0.0526	0.1568	0.1780	4.2822
Normal	Gamma	0.0007	0.1220	-0.0121	0.2166	0.0012	0.0003	0.0321	0.0618	0.0667	0.8228
Extreme	Lognormal	0.0142	0.1929	-1.2892	1.7127	-0.0012	0.0019	-0.0719	0.0709	-0.4457	2.0225
Normal	Lognormal	0.0092	0.1136	-1.2781	1.6878	0.0010	0.0002	-0.0509	0.0305	-0.2969	0.4472

True parameter values: $\beta = 2$

For gamma frailty distribution: $\widehat{\text{Var}}(W_i) = 2$

For lognormal frailty distribution: $\widehat{\text{Var}}(W_i) = 1.7$

For extreme value baseline distribution: $\Lambda_0(-2) = 0.1353$, $\Lambda_0(0) = 1$, $\Lambda_0(1.5) = 4.4817$

For normal baseline distribution: $\Lambda_0(-2) = 0.0230$, $\Lambda_0(0) = 0.6931$, $\Lambda_0(1.5) = 2.7059$

3.3.5 Robustness to distributional assumptions

Different distributional assumptions for the error terms and frailty terms were also explored. In addition to the extreme value distribution, data were also generated assuming a standard normal distribution for the error distribution. To assess the robustness of the SES algorithm to misspecification of the gamma frailty assumption, frailties were also generated from a lognormal distribution with mean 1 and a variance of 1.7; that is, a normal distribution with a mean of -0.5 and a variance of 1 for the natural logarithm of the frailty terms. These distributions were chosen to allow direct comparison of our results to those of Zhang and Peng (2007) and Xu and Zhang (2010).

As expected, Table 3.5 shows that the distribution of the error terms has minimal impact on the parameter estimates. In addition, we observe that misspecification of the frailty distribution impacts the bias and mean squared error of θ^{-1} but has little impact on the estimated regression parameters. These results are comparable to those reported in Zhang and Peng (2007) and Xu and Zhang (2010). While results for $\hat{\Lambda}_0(\cdot)$ are included for completeness, they are difficult to interpret since the size of the risk sets at $s = -2, 0$, and 1.5 differ depending on the generating distributions.

Table 3.6: Weighted bootstrap standard error estimates

	M = 500		M = 1,000		M = 2,000	
	Emp SE	Boot SE	Emp SE	Boot SE	Emp SE	Boot SE
$\hat{\beta}$	0.4983	0.4771	0.4861	0.4627	0.4730	0.4709
$\hat{\theta}^{-1}$	0.4607	0.4355	0.4438	0.4291	0.4548	0.4254
$\hat{\Lambda}_0(-2)$	0.0550	0.0619	0.0549	0.0613	0.0530	0.0648
$\hat{\Lambda}_0(0)$	0.3927	0.4119	0.3779	0.4200	0.3698	0.4238
$\hat{\Lambda}_0(1.5)$	2.0629	2.5335	1.9217	2.1730	1.8487	2.0117

True parameter values: $\beta = 2$, $\theta = 0.5$, $\Lambda_0(-2) = 0.1353$, $\Lambda_0(0) = 1$, $\Lambda_0(1.5) = 4.4817$
 M = number of simulated datasets, Emp SE = $SD(\hat{\psi} - \psi)$, Boot SE = $SD(\hat{\psi}^* - \hat{\psi})$

3.3.6 Variance estimation

Table 3.6 evaluates the performance of the weighted bootstrap standard error estimates. The centered weighted bootstrap estimates, $\hat{\psi}^* - \hat{\psi}$, should have the same unconditional distribution as the centered unweighted estimates, $\hat{\psi} - \psi$ (Ma and Kosorok, 2005). It follows that the performance of the bootstrap standard error estimates can be evaluated by computing a single bootstrap estimate for each simulated dataset, $\hat{\psi}^*$, and then computing the Monte Carlo estimate of the standard error of $\hat{\psi}^* - \hat{\psi}$. The result should be comparable to the empirical standard error of $\hat{\psi} - \psi$. Simulations were run using 500, 1,000, and 2,000 datasets. The centered weighted bootstrap performed well for $\hat{\beta}$, $\hat{\theta}^{-1}$, and $\hat{\Lambda}_0(\cdot)$, being within simulation error in each case and generally getting closer to the empirical standard error of $\hat{\psi} - \psi$ as the number of simulated datasets increases.

3.4 Illustrations

In this section, three datasets are re-analyzed for illustrative purposes, and with the intent of comparing the results obtained using the proposed SES algorithm and weighted bootstrap procedure to previously published results in other papers that have introduced methods for the AFT model with correlated data.

3.4.1 Litter-matched tumorigenesis

In this section, we revisit the litter-matched tumorigenesis data originally reported in Mantel et al. (1977) and subsequently analyzed in both Lee et al. (1993) and Jin et al. (2006a, §6). The study used 50 litters of female rats, with each litter containing one treated rat and two control rats; investigators were interested in determining whether a difference existed in the time until tumor appearance by treatment arm. The complete dataset may be found in Lee et al. (1993, Table 1).

Let \bar{T}_{ik} denote the time of tumor appearance for the k th rat of the i th litter and let X_{ik} indicate whether the rat was drug-treated or not by the values 0 and 1, respectively. The model specified by equations (3.1) and (3.2) of §3.1.1 was fit to the data using the SES algorithm with the GMM estimator for the initial estimate of θ , as recommended in §3.3. The weighted bootstrap standard errors were calculated based on 500 bootstrap replicates.

Using the proposed SES algorithm, the estimated regression parameter is 0.179 with a standard error of 0.096, the latter estimated with the weighted bootstrap procedures. Similarly, we estimate the variance of the frailty term is to be 0.458 with a standard error of 0.267. Increasing the number of weighted bootstrap replicates from 500 to 1000 does not measurably change our standard error estimates. The regression parameter results are comparable to those of Jin et al. (2006a) (§6), who employ marginal methods and report a regression parameter estimate of 0.156 with a resampling-based standard error of 0.093 based on 10,000 replications. Using the marginal methods of §2.1.3, the estimated regression parameter estimate is 0.166 with a standard error of 0.084.

3.4.2 Isosorbide dinitrate in angina

Danahy et al. (1977) report the results of a study on the oral administration of

isosorbide dinitrate to 21 coronary heart disease patients. The data has subsequently been analyzed by Zhang and Peng (2007) and Xu and Zhang (2010), among others. In the study, each patient performed an exercise test under 10 different study treatments during the course of a five day hospital stay. For each test, the patient was followed until the onset time of angina pectoris with censoring occurring when patients became too exhausted to continue. The first two days of the hospital stay involved practice exercise tests only. On the third day, patients were tested after receiving sublingual placebo (SLP) and again after receiving sublingual nitroglycerin (SLT). On the fourth day, patients were tested at baseline (P0) and at 1, 3, and 5 hours after receiving an oral placebo (P1, P3, and P5). On the fifth day, patients were tested at baseline (ID0) and at 1, 3, and 5 hours after receiving oral isosorbide dinitrate (ID1, ID3, and ID5). Thus a total of 10 measurements were taken on each of the 21 patients; repeated observations on a patient are likely correlated. The complete dataset may be found in Danahy et al. (1977, Table 2).

As in Zhang and Peng (2007) and Xu and Zhang (2010), we will compare the time until angina on SLP to the time until angina under the nine other treatments. Let \bar{T}_{ik} denote the time until angina for the i th patient on the k th treatment. The model will contain nine indicator covariates corresponding to the nine treatments (SLT, P0, P1, P3, P5, ID0, ID1, ID3, and ID5) to be compared to SLP. The model specified by equations (3.1) and (3.2) of §3.1.1 was fit to the data using the SES algorithm with the GMM initial estimate and the weighted bootstrap standard errors were calculated based on 500 bootstrap replicates.

Table 3.7 contains the parameter estimates and standard errors. The regression parameter estimates are similar to those reported by Zhang and Peng (2007) and Xu and Zhang (2010), generally lying in between the estimates reported in

Table 3.7: Illustration 3.4.2 results

	Parameter Estimate	Standard Error
SLT	0.3174	0.1394
P0	-0.0659	0.0804
P1	0.0719	0.0501
P3	-0.0208	0.0871
P5	-0.0682	0.0587
ID0	0.0202	0.0525
ID1	0.2033	0.1478
ID3	0.1513	0.1293
ID5	-0.0138	0.0939
$\hat{\theta}^{-1}$	2.5441	0.4145

Table 3 of Xu and Zhang (2010). In particular, the effect of SLT compared to SLP is significant with a parameter estimate and standard error of 0.3174 and 0.1394, similar to the corresponding estimates 0.352 and 0.118 reported by Zhang and Peng (2007). Our reported standard errors tend to be somewhat larger than those reported in Table 3 of Xu and Zhang (2010). Increasing the number of bootstrap replicates from 500 to 1000 did not appreciably change the standard error estimates. Our estimate of θ^{-1} (i.e., the variance of the frailty distribution) is 2.5411, approximately 21.5% larger than the estimates reported by Zhang and Peng (2007) and Xu and Zhang (2010), 2.096 and 2.091, respectively. Neither Zhang and Peng (2007) nor Xu and Zhang (2010) report standard error estimates for $\hat{\theta}^{-1}$.

3.4.3 Recurrent time to infection following catheterization

Finally, we revisit the recurrent kidney infection data originally reported and analyzed in McGilchrist and Aisbett (1991) and subsequently analyzed by Therneau and Grambsch (2000) and Pan (2001). The study followed 38 kidney disease patients from the time of catheterization until the time of infection at the catheterization site. Each patient was followed for 2 catheterization periods with censoring occurring when catheters were removed for reasons other than

infection. The data may be found in McGilchrist and Aisbett (1991, Table 1).

Let \bar{T}_{ik} denote the time until infection for the i th patient following their k th catheterization. As in Pan (2001), we consider a single covariate, X_{ik} , taking the value 0 if the patient was male and 1 if the patient was female. The model specified by equations (3.1) and (3.2) of §3.1.1 was fit to the data using the SES algorithm with the GMM initial estimate and the weighted bootstrap standard errors were calculated based on 500 bootstrap replicates.

The regression parameter estimate is 1.544 with a standard error of 0.317 and the estimate of the variance of the frailty term is 0.216 with a standard error of 0.068. As in the previous illustrations, increasing the number of bootstrap replicates does not impact the standard error estimates. The results are similar to those of Pan (2001) who also considers the gamma frailty AFT model and reports a regression parameter estimate of 1.42 with an unweighted bootstrap standard error of 0.44. Pan's estimate of the frailty variance θ^{-1} is 0.20, obtained using a grid search of the profile likelihood; a corresponding estimate of standard error is not reported.

3.5 Remarks

This chapter utilizes the induced smoothing procedure to develop a new estimation algorithm for the semiparametric AFT frailty model. Our results demonstrate that the estimates obtained using the proposed SES algorithm are comparable to those obtained using existing methods for the gamma frailty model. The proposed algorithm has the added advantages of being numerically stable and easy to implement using widely available software, providing estimates of all parameters and, combined with the weighted bootstrap, corresponding standard errors.

The attractive nature of the induced smoothing procedure used in the S-step of the SES algorithm stems from the strict convexity of the Gehan-weighted objective function (3.8). However, use of the Gehan weight is often criticized for the inefficiency of the resulting estimator. The use of an alternative weight function may result in efficiency improvements at the expense of losing the monotonicity of the corresponding estimating equation for β . Starting from the Gehan estimator, Strawderman (2005) demonstrates how one may instead use one-step estimation to obtain estimates for alternate weight functions. One-step estimation could conceivably be incorporated into the S-step of the algorithm, or be used following convergence of the algorithm to obtain parameter estimates for alternate weight functions. Alternatively, but in related fashion, one could implement the S-step for general weight functions by combining the induced smoothing procedure with the iteratively reweighted Gehan estimator of Jin et al. (2003, 2006a).

Underestimation of the frailty variance is an acknowledged problem in semi-parametric estimation methods for both the Cox model (e.g., Barker and Henderson, 2005; Nielsen et al., 1992) and the AFT model (e.g., Pan, 2001; Zhang and Peng, 2007). The simulation results show that this problem persists with the proposed SES algorithm; however, the negative bias observed in the simulation results of §3.3 is substantially smaller than the negative biases reported in Zhang and Peng (2007) and Xu and Zhang (2010). Our estimates of the variance of the frailty terms in the illustrations of §3.4 are larger than what has previously been reported. It is therefore possible that smoothing the objective function for the regression parameter helps to offset the tendency to underestimate the variance of the frailty term. Developing a smoothing method for the baseline hazard may further reduce the bias of the estimated frailty variance, as suggested

by the findings of Barker and Henderson (2005) in the case of the Cox regression model.

The simulation results, and also the data examples, assume that the frailty has a gamma distribution. While other frailty distributions may be considered (see §3.2.2), the regression parameter estimates obtained via the SES algorithm were found to be robust to misspecifications of the frailty distribution and are similar to results obtained by fitting marginal models. Both are consequences of the model (3.1), as (3.2) indicates that one may interpret the frailty as capturing the correlation structure of the residuals within each cluster. That is, unlike the Cox regression model, the regression coefficient β measures both the conditional and marginal impact of covariates, insofar as they impact the mean of the log-failure time; see Strawderman (2006) for related comments in the context of recurrent event data. Provided that one is only interested in regression coefficients, the results of §2.1.3 may instead be used to estimate β in the marginal AFT model. However, greater efficiency may be possible in settings where the exchangeability assumption inherent in (3.1) and (3.2) is violated. In particular, Wang and Zhu (2006) propose to use the fact that the dependence structure differs between and within clusters to derive a more efficient estimator of β in the uncensored rank regression problem. In Fu et al. (2010), asymptotic results for induced smoothing are established in this setting; an extension of these arguments may be used to prove comparable results in the case of censored response data under the marginal AFT model. Finally, we remark that the induced smoothing method and SES algorithm can be easily adapted to the “accelerated gap time” models for recurrent event data (see Strawderman (2005, 2006)).

CHAPTER 4

A TWO-STAGE RESIDUAL-DEPENDENT SAMPLING DESIGN

4.1 Two-Stage Sampling Designs

Suppose that a stage one sample of N subjects is selected from the target population. The stage one sample will be referred to as the parent population. A continuous outcome variable, Y_i , and a vector of stage one covariates, \mathbf{X}_i , is measured on the parent population. However, a univariate stage two risk factor, Z_i , will only be measured on a subsample of the parent population. The goal of the two-stage sampling designs presented in this dissertation is to use the available stage one data to identify subjects who may be informative about the relationship between Y_i and Z_i and, hence, should be oversampled for the stage two sample. Ideally, the sampling designs should lead to more efficient estimators than those that would have been obtained had the stage two sample been selected using simple random sampling (SRS).

We will assume throughout this section that interest lies in estimating the normal linear regression model

$$Y_i = \beta_0 + \mathbf{X}_i' \boldsymbol{\beta}_x + Z_i \beta_z + \epsilon_i, \quad (4.1)$$

where \mathbf{X}_i is a $p \times 1$ vector of stage one covariates, Z_i is the stage two covariate, β_0 is the intercept, $\boldsymbol{\beta}_x$ is a $p \times 1$ vector of regression parameters corresponding to \mathbf{X}_i , and β_z is the regression parameter for Z_i ($i = 1, \dots, n$). We assume that $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ has mean zero and covariance matrix $\sigma^2 I_n$. Alternatively, the model can be written

$$Y_i = \mathbf{W}_i' \boldsymbol{\beta} + \epsilon_i, \quad (4.2)$$

where $\mathbf{W}_i = (1 \ \mathbf{X}_i' \ Z_i)'$ and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_x', \beta_z)'$ are $(p+2) \times 1$ vectors. For simplicity, we further assume throughout that (Y_i, \mathbf{W}_i) ($i = 1, \dots, n$) are independent and

identically distributed with all components of \mathbf{W}_i being uniformly bounded, that ϵ_i is independent of \mathbf{W}_i ($i = 1, \dots, n$), and that $E[\mathbf{W}_1^{\otimes 2}]$ is positive definite, where $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}'$ for any vector \mathbf{v} .

We will first present a family of two-stage outcome-dependent sampling designs for continuous response variables which has been previously described; see, for example, Lawless et al. (1999). We will then introduce a new family of two-stage sampling designs which depend on the observed stage one data through the residuals of a stage one linear regression model.

4.1.1 Outcome-Dependent Sampling

One family of two-stage outcome-dependent sampling (ODS) designs for continuous outcome variables can be implemented as follows.

Stage 1:

Observe Y_i, \mathbf{X}_i for $i = 1, \dots, N$.

Stage 2:

- (i) Assume that the domain of the response variable, \mathcal{Y} , can be partitioned into K strata defined by the specified cutpoints $-\infty = a_0 < a_1 < \dots < a_K = \infty$. That is, the k th stratum is defined as $\mathcal{V}_k = (a_{k-1}, a_k]$. Stratify the N subjects in the parent population into $\mathcal{V}_1, \dots, \mathcal{V}_K$ using their observed Y_i values. Let N_k denote the number of subjects in \mathcal{V}_k ($k = 1, \dots, K$).
- (ii) Take a simple random sample of size n_k from \mathcal{V}_k ($k = 1, \dots, K$) and collect information on Z_i for the $n = \sum_{k=1}^K n_k$ selected subjects.

In order for the ODS design to be more efficient than SRS, the strata and strata-specific sample sizes must be selected to identify and oversample subjects that are informative about the relationship between Y_i and Z_i . With no prior information about the Z_i values, this will typically involve oversampling subjects with large or small values of Y_i .

If covariate information is collected at stage one, this information can easily be incorporated into the ODS design described above by allowing stratification to depend on both the response variable and the stage one covariates. For example, if data on a univariate covariate, X_i , was collected at stage one, the domain of the response variable and the covariate, $\mathcal{Y} \times \mathcal{X}$, could be partitioned to define the strata. The usefulness of this type of ODS plus covariates design in practice is not entirely clear. Subjects that are informative about the relationship between Y_i and X_i can easily be identified from the parent population but, without prior knowledge regarding the relationship between X_i and Z_i or Y_i and Z_i , stratification based on Y_i and X_i cannot easily be used to identify subjects that are informative about the relationship between Y_i and Z_i . Even in the simple case of a univariate stage one covariate, determining how to define strata on $\mathcal{Y} \times \mathcal{X}$ is a difficult problem. Consideration of additional stage one covariates increases the dimension of the stratification problem making these determinations decidedly more complex.

4.1.2 Residual-Dependent Sampling

As an alternative to defining the stage one strata using the response variable and the stage one covariates, or not using the stage one covariates at all, a linear regression model could be fit to the stage one data and the stage two sample could be selected based on the residual values from that model. The principal

rationale behind this type of residual-dependent sampling design is to try to identify those subjects whose outcomes are not predicted well by knowledge of \mathbf{X}_i alone and, hence, may be informative about the relationship between the outcome and the stage two covariate. An important advantage of using residuals in a two-stage sampling scheme is that the always-observed stage one data for a subject, (Y_i, \mathbf{X}_i) , is collapsed into a univariate summary score, thereby easing the burden of stratification through significant dimension reduction.

A linear model fit to the stage one covariates can be written

$$Y_i = \mathbf{V}_i' \boldsymbol{\alpha}_0 + \epsilon_i^*,$$

where $\mathbf{V}_i' = (1 \ \mathbf{X}_i')$, $\boldsymbol{\alpha}_0$ is a $(p + 1) \times 1$ vector of regression parameters to be discussed in detail later, and ϵ_i^* is assumed (possibly incorrectly) to have mean zero and constant variance ($i = 1, \dots, N$). A two-stage residual-dependent sampling (RDS) design which differentially selects subjects based on their stage one residual values may be implemented as follows.

Stage 1:

- (i) Observe Y_i, \mathbf{X}_i for $i = 1, \dots, N$.
- (ii) Compute the least squares estimate $\hat{\boldsymbol{\alpha}}$ for the first stage sample and, for $i = 1, \dots, N$, the residuals $e_i^* = Y_i - \hat{Y}_i$, where $\hat{Y}_i = \mathbf{V}_i' \hat{\boldsymbol{\alpha}}$.

Stage 2:

- (i) Assume that the domain of the stage one model residuals, \mathcal{E} , can be partitioned into K strata defined by the cutpoints $-\infty = b_0 < b_1 < \dots < b_K = \infty$. That is, the k th stratum is defined as $\mathcal{V}_k^* = (b_{k-1}, b_k]$. Stratify the N subjects in the parent population into $\mathcal{V}_1^*, \dots, \mathcal{V}_K^*$ using their stage one residual values, e_i^* . Let \hat{N}_k^* denote the number of subjects in \mathcal{V}_k^* ($k = 1, \dots, K$).

- (ii) Take a simple random sample of size n_k^* from \mathcal{V}_k^* ($k = 1, \dots, K$) and collect information on Z_i for the $n = \sum_{k=1}^K n_k^*$ selected subjects.

Under the population model (4.1), or equivalently (4.2), the observed stage one residual e_i^* may be written

$$e_i^* = Y_i - \mathbf{V}_i' \hat{\boldsymbol{\alpha}} = \mathbf{V}_i' \left(\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_x \end{pmatrix} - \hat{\boldsymbol{\alpha}} \right) + Z_i \beta_z + \epsilon_i. \quad (4.3)$$

Therefore, while calculation of e_i^* only requires stage one information, its magnitude clearly contains useful information about the stage two covariate Z_i , provided that β_z is non-zero. Ideally, we would like to be able to identify subjects for whom $Z_i \beta_z$ makes a comparatively significant contribution to e_i^* , such as when $|Z_i|$ is large (e.g., high leverage), for inclusion in the stage two sample. However, using e_i^* to identify such subjects may also inadvertently identify subjects with large values of ϵ_i , as well as subjects for which

$$\mathbf{V}_i' \left(\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_x \end{pmatrix} - \hat{\boldsymbol{\alpha}} \right)$$

is large in magnitude. Whether or not this is a true impediment to the RDS design appears to largely depend on the relationship between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta}$ and, in turn, the relationship between \mathbf{X}_i and Z_i . We explore these issues in greater detail below.

Let \mathbf{V} denote the $N \times (p + 1)$ design matrix containing the stage one covariates, \mathbf{W} denote the $N \times (p + 2)$ design matrix containing the stage one and two covariates, \mathbf{X} and \mathbf{Z} respectively denote the $N \times p$ matrix and $N \times 1$ vector formed from the stage one and two covariates, and \mathbf{Y} denote the $N \times 1$ vector of response. Then, under (4.2), the least squares estimate $\hat{\boldsymbol{\alpha}}$ based on the stage

one sample can be expressed

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{Y} \\ &= (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'(\mathbf{W}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{W}\boldsymbol{\beta} + (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\boldsymbol{\epsilon}\end{aligned}$$

where $\boldsymbol{\epsilon}$ is an $N \times 1$ vector of mean zero error terms. As \mathbf{V} and $\boldsymbol{\epsilon}$ are independent under assumptions stated earlier, $E(\hat{\boldsymbol{\alpha}}|\mathbf{W}) = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{W}\boldsymbol{\beta}$ and $E(\hat{\boldsymbol{\alpha}}|\mathbf{V}) = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'E(\mathbf{W}|\mathbf{V})\boldsymbol{\beta}$. Easy calculations further demonstrate that

$$\mathbf{e}^* = (\mathbf{I}_N - \mathbf{H})\mathbf{Z}\boldsymbol{\beta}_z + (\mathbf{I}_N - \mathbf{H})\boldsymbol{\epsilon},$$

where $\mathbf{e}^* = (e_1^*, \dots, e_N^*)'$ and $\mathbf{H} = \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'$ is the so-called hat matrix. It is easily seen that $E(\mathbf{e}^*|\mathbf{V}) = (\mathbf{I}_N - \mathbf{H})E(\mathbf{Z}|\mathbf{X})\boldsymbol{\beta}_z$. Moreover, a simple calculation shows that $E\{(\mathbf{e}^*)^{\otimes 2}|\mathbf{W}\} - E(\mathbf{e}^*|\mathbf{W})^{\otimes 2} = \sigma^2(\mathbf{I}_N - \mathbf{H})$; see, for example, Ramsey (1969). However, conditioning on \mathbf{V} only, this last result is no longer true and thus does not present a useful route towards estimation of σ^2 based only on the first stage sample information.

Considered alone, the first stage model parameter $\boldsymbol{\alpha}_0$ is neither especially meaningful nor of direct interest. This can be seen, for example, by noting that $\mathbf{V}'_i\boldsymbol{\alpha}_0$ fails to describe the mean of Y_i if the stage one model is incorrectly specified. In contrast, the corresponding residual vector \mathbf{e}^* constructed from $\hat{\boldsymbol{\alpha}}$ remains meaningful as a measure of prediction error. While it is possible to study the behavior of \mathbf{e}^* directly, helpful insights are obtained through an asymptotic analysis of $\hat{\boldsymbol{\alpha}}$ (i.e., as $N \rightarrow \infty$) and the associated impact of this behavior on \mathbf{e}^* .

Using the strong law of large numbers, our assumptions imply that $N^{-1}(\mathbf{V}'\mathbf{V})$ converges almost surely to the $(p+1) \times (p+1)$ matrix

$$E(\mathbf{V}_1^{\otimes 2}) = \begin{pmatrix} 1 & E(\mathbf{X}'_1) \\ E(\mathbf{X}_1) & E(\mathbf{X}_1^{\otimes 2}) \end{pmatrix}.$$

Since $\mathbf{W} = (\mathbf{V} \mathbf{Z})$ and \mathbf{Z} is $N \times 1$, $N^{-1}(\mathbf{V}'\mathbf{W})$ converges almost surely to the $(p+1) \times (p+2)$ matrix

$$(E(\mathbf{V}_1^{\otimes 2}) \ E(\mathbf{V}_1 Z_1)).$$

It follows that $(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{W}$ converges almost surely to $(\mathbf{I}_{p+1} \ \mathbf{K})$, where the $(p+1) \times 1$ vector $\mathbf{K} = E(\mathbf{V}_1^{\otimes 2})^{-1}E(\mathbf{V}_1 Z_1)$ and the required inverse of $E(\mathbf{V}_1^{\otimes 2})$ exist as a consequence of earlier assumptions. In summary, the above calculations imply

$$\hat{\boldsymbol{\alpha}} \xrightarrow{a.s.} \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_x \end{pmatrix} + \mathbf{K}\beta_z.$$

Writing

$$E(\mathbf{V}_1^{\otimes 2}) = \begin{pmatrix} 1 & E(\mathbf{X}'_1) \\ E(\mathbf{X}_1) & E(\mathbf{X}_1^{\otimes 2}) \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{\mu}'_x \\ \boldsymbol{\mu}_x & \boldsymbol{\delta}_x \end{pmatrix}$$

and using well-known results for calculating the inverse of a partitioned symmetric matrix, we may write

$$E(\mathbf{V}_1^{\otimes 2})^{-1} = \begin{pmatrix} 1 + \boldsymbol{\mu}'_x \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\mu}_x & -\boldsymbol{\mu}'_x \boldsymbol{\Sigma}_x^{-1} \\ -\boldsymbol{\Sigma}_x^{-1} \boldsymbol{\mu}_x & \boldsymbol{\Sigma}_x^{-1} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_x = \boldsymbol{\delta}_x - \boldsymbol{\mu}_x^{\otimes 2}$. Writing

$$E(\mathbf{V}_1 Z_1) = \begin{pmatrix} E(Z_1) \\ E(\mathbf{X}_1 Z_1) \end{pmatrix} = \begin{pmatrix} \mu_z \\ \boldsymbol{\gamma}_{xz} \end{pmatrix},$$

we further find

$$\mathbf{K} = \begin{pmatrix} \mu_z(1 + \boldsymbol{\mu}'_x \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\mu}_x) - \boldsymbol{\mu}'_x \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\gamma}_{xz} \\ \boldsymbol{\Sigma}_x^{-1}(\boldsymbol{\gamma}_{xz} - \boldsymbol{\mu}_x \mu_z) \end{pmatrix} = \begin{pmatrix} \mu_z - \boldsymbol{\mu}'_x \boldsymbol{\Sigma}_x^{-1} \mathbf{C}_{xz} \\ \boldsymbol{\Sigma}_x^{-1} \mathbf{C}_{xz} \end{pmatrix}$$

where $\mathbf{C}_{xz} = \boldsymbol{\gamma}_{xz} - \boldsymbol{\mu}_x \mu_z$ is the $p \times 1$ vector containing the covariances $\text{cov}(X_{1j}, Z_1)$ for $j = 1, \dots, p$. Hence, we now see that $\hat{\boldsymbol{\alpha}} \xrightarrow{a.s.} \boldsymbol{\alpha}$, where

$$\boldsymbol{\alpha} = \begin{pmatrix} \beta_0 + \beta_z(\mu_z - \boldsymbol{\mu}'_x \boldsymbol{\Sigma}_x^{-1} \mathbf{C}_{xz}) \\ \boldsymbol{\beta}_x + \beta_z \boldsymbol{\Sigma}_x^{-1} \mathbf{C}_{xz} \end{pmatrix}. \quad (4.4)$$

Notably, these results extend in a straightforward manner to the setting where the stage two covariate Z_i is a vector.

The vector α describes the limiting behavior of the least squares estimator $\hat{\alpha}$ and, unlike α_0 , remains an interpretable quantity regardless of the validity of the first stage model, $Y_i = \mathbf{V}_i' \alpha_0 + \epsilon_i^*$ ($i = 1, \dots, N$), used earlier to motivate the RDS design. With the exception of the case where both \mathbf{C}_{xz} and μ_z are equal to zero, the form of α shows that $\hat{\alpha}$ is biased for the vector $(\beta_0, \beta_x)'$, the degree of bias depending on a combination of factors involving \mathbf{C}_{xz} , μ_x , Σ_x , μ_z and β_z . This is known as omitted variable bias (Ramsey, 1969), and can be anticipated to occur in the setting of interest in this paper; that is, when Z_i is predictive of Y_i . In this regard, the results just developed provide useful information regarding the magnitude of the first stage residuals in (4.3) in settings where the first stage model may be misspecified. In particular, defining $\tilde{\epsilon}_i^* = Y_i - \mathbf{V}_i' \alpha$ with α defined in (4.4) and assuming (4.2), we now see that e_i^* “estimates”

$$\tilde{\epsilon}_i^* = \mathbf{V}_i' \left(\begin{pmatrix} \beta_0 \\ \beta_x \end{pmatrix} - \alpha \right) + Z_i \beta_z + \epsilon_i. \quad (4.5)$$

Using (4.4) and simplifying, we find

$$\tilde{\epsilon}_i^* = \beta_z (Z_i - \mu_z) - \beta_z \{ (\mathbf{X}_i - \mu_x)' \Sigma_x^{-1} \mathbf{C}_{xz} \} + \epsilon_i.$$

Suppose now that $\mathbf{C}_{xz} = \mathbf{0}$. That is, the stage two covariate is uncorrelated with all stage one covariates. Then,

$$\tilde{\epsilon}_i^* = \beta_z (Z_i - \mu_z) + \epsilon_i,$$

demonstrating that the “true” observed residuals only reflect two major sources of variability. More generally, the magnitude of the e_i^* 's depends on three sources of variability, two of which directly or indirectly reflect the influence of the stage two covariate.

The determination of strata on the basis of this procedure depends on (i) the true β , here through (4.4); and, (ii) the additional variability created by the estimation of α through $\hat{\alpha}$ (i.e., from using e_i^* in place of $\tilde{e}_i^* = Y_i - \mathbf{V}_i' \alpha$ for stratification). Ideally, for the RDS design, we would use \tilde{e}_i^* to stratify subjects, a process ultimately requiring knowledge of β ; in practice, we instead use the observed stage one model residuals. While we assume that the strata boundaries, b_0, b_1, \dots, b_K , are fixed, the number of subjects that fall into stratum k may be affected when these estimates are used. Define

$$N_k^*(\mathbf{a}) = \sum_{i=1}^N I\{e_i(\mathbf{a}) \in \mathcal{V}_k^*\},$$

where $e_i(\mathbf{a}) = Y_i - \mathbf{V}_i' \mathbf{a}$ ($i = 1, \dots, N$). Then, if α were known and able to be used in the implementation of RDS, the observed number of subjects in stratum k would be $N_k^* = N_k^*(\alpha)$. In practice, however, when $\hat{\alpha}$ is used in place of α , the observed number of subjects in stratum k is instead $\hat{N}_k^* = N_k^*(\hat{\alpha})$. The number of subjects sampled from the k th stratum, n_k^* , is fixed and will not be affected by the stage one regression parameter estimates. Thus the impact of using the stage one regression parameter estimates in place of the true regression parameters on the estimation of the stage two model will manifest itself through the substitution of \hat{N}_k^* for N_k^* in the calculation of the stage two sampling probabilities. This will be discussed further in §4.2, where the failure of $N_k^*(\mathbf{a})$ to be a smooth function of \mathbf{a} creates serious challenges in proving useful large sample results.

4.2 Estimation

We will first establish some terminology and notation commonly used in the two-stage sampling literature. Let R_i be an indicator of inclusion in the stage

two sample, $R_i = I(Z_i \text{ is observed})$. We will let the *full data* refer to the data that would be collected if all N subjects were included in the stage two sample, $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$. The *observed data* will refer to the data that is actually observed on the subjects, $(\mathbf{Y}, \mathbf{X}, \mathbf{Z}_{obs})$ where $\mathbf{Z}_{obs} = \{Z_i : R_i = 1\}$ and the *complete data* will refer to the data only from the subjects included in the stage two sample, $(\mathbf{Y}_{obs}, \mathbf{X}_{obs}, \mathbf{Z}_{obs})$ where $\mathbf{Y}_{obs} = \{Y_i : R_i = 1\}$, $\mathbf{X}_{obs} = \{\mathbf{X}_i : R_i = 1\}$.

We would like to obtain parameter estimates for the stage two linear regression model (4.2) under the sampling designs presented in §4.1. If the full data was observed, regression parameter estimates could be defined as the solution to the maximum likelihood score function,

$$S_F(\boldsymbol{\beta}) = \sum_{i=1}^N \frac{\partial \log f(Y_i | \mathbf{W}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \sum_{i=1}^N (\mathbf{Y}_i - \mathbf{W}_i' \boldsymbol{\beta}) \mathbf{W}_i'.$$

For simplicity of presentation, we assume throughout that σ^2 is known, noting that extensions to cover the case of estimated σ^2 are straightforward. Thus σ^2 will no longer appear in the score functions as we proceed.

Since Z_i is observed only for subjects included in the stage two sample, estimation of $\boldsymbol{\beta}$ might naïvely be based on the complete data score function,

$$S_C(\boldsymbol{\beta}) = \sum_{i=1}^N R_i (\mathbf{Y}_i - \mathbf{W}_i' \boldsymbol{\beta}) \mathbf{W}_i', \quad (4.6)$$

an estimating equation that ignores the biased sampling design, as well as fails to utilize the stage one data collected on subjects not included in the stage two sample.

The inverse probability weighted (IPW) estimator incorporates the sampling design into the maximum likelihood score equation by inversely weighting completely observed subjects with their stage two sampling probabilities. The resulting weighted score function is

$$S_W(\boldsymbol{\beta}; \mathbf{w}) = \sum_{i=1}^N R_i w_i (\mathbf{Y}_i - \mathbf{W}_i' \boldsymbol{\beta}) \mathbf{W}_i' \quad (4.7)$$

where the weights, $\mathbf{w} = (w_1, \dots, w_N)'$, are the inverse sampling probabilities for subjects $i = 1, \dots, N$. The solution to $S_W(\hat{\boldsymbol{\beta}}; \mathbf{w}) = 0$ is commonly referred to as the IPW estimator or the Horvitz-Thompson estimator.

In the appendix, we show that under specified regularity conditions, the solution to an IPW score function of the form (4.7) is consistent and asymptotically normal for a general two-stage stratified sampling scheme with known weights and a general score function derived from a concave likelihood function. The IPW score functions for the ODS and RDS designs are special cases of this more general result and we will present the relevant results for these designs in §4.2.1 and §4.2.2.

4.2.1 Outcome-Dependent Sampling

For the traditional ODS design, the sampling probability for a subject depends only on their observed response; that is, $\pi_i = P(R_i = 1 | Y_i, \mathbf{X}_i, Z_i) = P(R_i = 1 | Y_i)$. In other words, the missing data for a subject not included in the stage two sample, $\{Z_i : R_i = 0\}$, is missing at random (MAR). Under the MAR assumption, the IPW score function (4.7) can be used for the ODS design with $w_i = 1/\pi_i$ and $\pi_i = P(R_i = 1 | Y_i) = n_k/N_k$ if $Y_i \in \mathcal{V}_k$ ($k = 1, \dots, K$). The resulting estimating equation can be expressed

$$S_W(\boldsymbol{\beta}; \mathbf{w}) = \sum_{i=1}^N \sum_{k=1}^K R_i I(Y_i \in \mathcal{V}_k) \frac{N_k}{n_k} (\mathbf{Y}_i - \mathbf{W}_i' \boldsymbol{\beta}) \mathbf{W}_i'. \quad (4.8)$$

Using the results in the appendix, and under the assumptions outlined earlier, the solution to $S_W(\hat{\boldsymbol{\beta}}; \mathbf{w}) = 0$ is strongly consistent and asymptotically normal. The asymptotic variance of $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ can be estimated as $\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{J}}^{-1} + \hat{\mathbf{J}}^{-1} \hat{\mathbf{G}} \hat{\mathbf{J}}^{-1}$ where

$$\hat{\mathbf{J}} = \frac{1}{N\sigma^2} \sum_{i=1}^N \sum_{k=1}^K \frac{N_k}{n_k} R_i I(Y_i \in \mathcal{V}_k) \mathbf{W}_i \mathbf{W}_i'$$

and

$$\hat{\mathbf{G}} = \frac{1}{N\sigma^4} \sum_{k=1}^K \frac{N_k(N_k - n_k)}{n_k^2} \sum_{i=1}^N R_i \left\{ (Y_i - \mathbf{W}_i' \hat{\boldsymbol{\beta}}) \mathbf{W}_i' \right\}^{\otimes 2} I(Y_i \in \mathcal{V}_k) - \frac{1}{N\sigma^4} \sum_{k=1}^K \frac{N_k(N_k - n_k)}{n_k^3} \left[\sum_{i=1}^N R_i \left\{ (Y_i - \mathbf{W}_i' \hat{\boldsymbol{\beta}}) \mathbf{W}_i' \right\} I(Y_i \in \mathcal{V}_k) \right]^{\otimes 2}.$$

As discussed earlier, σ^2 is assumed known here. In practice, a suitable estimate must be obtained based on the stage one and two data and substituted into the above estimates.

4.2.2 Residual-Dependent Sampling

For the RDS design, the sampling probability for a subject depends only on their observed response and stage one covariate information, and similarly to ODS the missing data for a subject not included in the stage two sample is also MAR. As discussed earlier, stratification here depends on (i) the true $\boldsymbol{\beta}$ through knowledge of $\boldsymbol{\alpha}$ in (4.4); and, (ii) all stage one data, and the associated additional variability, resulting from the need to estimate $\boldsymbol{\alpha}$ by $\hat{\boldsymbol{\alpha}}$. It follows that estimation of the stage one regression parameters has the potential to impact estimation of the stage two regression model. We will begin by discussing estimation for the hypothetical case where $\boldsymbol{\alpha}$ in (4.4) is known, as the asymptotics are comparatively easy but revealing. The need to estimate $\boldsymbol{\alpha}$ creates significant challenges, at least from the perspective of developing proper asymptotics, and these will be discussed in turn.

Assuming $\boldsymbol{\alpha}$ is known

Under the MAR assumption, an IPW estimating equation with known weights can also be used for stage two regression parameter estimation under the RDS

design. Assuming that α is known, the IPW score function is

$$S_W^*(\beta; \alpha, \mathbf{w}(\alpha)) = \sum_{i=1}^N R_i w_i(\alpha) (\mathbf{Y}_i - \mathbf{W}_i' \beta) \mathbf{W}_i'$$

where $w_i(\alpha) = 1/\pi_i(\alpha)$ for $\pi_i(\alpha) = P(R_i = 1 | Y_i, \mathbf{X}_i, \alpha) = n_k^*/N_k^*$ if $\epsilon_i^* \in \mathcal{V}_k$ ($k = 1, \dots, K$). The resulting estimating equation can be expressed

$$S_W^*(\beta; \alpha, \mathbf{w}(\alpha)) = \sum_{i=1}^N \sum_{k=1}^K R_i I(\epsilon_i^* \in \mathcal{V}_k) \frac{N_k^*}{n_k^*} (\mathbf{Y}_i - \mathbf{W}_i' \beta) \mathbf{W}_i', \quad (4.9)$$

where $N_k^* = N_k^*(\alpha)$ ($k = 1, \dots, K$).

Again, using the results in the appendix, the solution to $S_W^*(\hat{\beta}^*; \alpha, \mathbf{w}(\alpha)) = 0$ is strongly consistent and asymptotically normal. The asymptotic variance of $N^{1/2}(\hat{\beta}^* - \beta)$ can be estimated as $\hat{\Sigma}^* = \hat{\mathbf{J}}^{*-1} + \hat{\mathbf{J}}^{*-1} \hat{\mathbf{G}}^* \hat{\mathbf{J}}^{*-1}$ where

$$\hat{\mathbf{J}}^* = \frac{1}{N\sigma^2} \sum_{i=1}^N \sum_{k=1}^K \frac{N_k^*}{n_k^*} R_i I(\epsilon_i^* \in \mathcal{V}_k) \mathbf{W}_i \mathbf{W}_i',$$

and

$$\begin{aligned} \hat{\mathbf{G}}^* &= \frac{1}{N\sigma^4} \sum_{k=1}^K \frac{N_k^*(N_k^* - n_k^*)}{n_k^{*2}} \sum_{i=1}^N R_i \left\{ (Y_i - \mathbf{W}_i' \hat{\beta}^*) \mathbf{W}_i' \right\}^{\otimes 2} I(\epsilon_i^* \in \mathcal{V}_k) \\ &\quad - \frac{1}{N\sigma^4} \sum_{k=1}^K \frac{N_k^*(N_k^* - n_k^*)}{n_k^{*3}} \left[\sum_{i=1}^N R_i \left\{ (Y_i - \mathbf{W}_i' \hat{\beta}^*) \mathbf{W}_i' \right\} I(\epsilon_i^* \in \mathcal{V}_k) \right]^{\otimes 2}. \end{aligned}$$

As before, recall that σ^2 is assumed known and must be estimated in practice.

Assuming α is estimated

In practice, α cannot be known and must be replaced by $\hat{\alpha}$. This substitution affects how the subjects are stratified according to their stage one data and, in turn, the sampling weights used in the IPW score function. Using the estimated stage one regression parameters, the sampling weights are $w_i(\hat{\alpha}) = 1/\pi_i(\hat{\alpha})$ where $\pi_i(\hat{\alpha}) = P(R_i = 1 | Y_i, \mathbf{X}_i, \hat{\alpha}) = n_k^*/\hat{N}_k^*$ if $e_i^* \in \mathcal{V}_k$ ($k = 1, \dots, K$), and

$\hat{N}_k^* = N_k^*(\hat{\alpha})$ ($k = 1, \dots, K$). Analogously to (4.9), the estimating equation can be written

$$S_W^*(\beta; \hat{\alpha}, w(\hat{\alpha})) = \sum_{i=1}^N \sum_{k=1}^K R_i I(e_i^* \in V_k^*) \frac{\hat{N}_k^*}{n_k^*} (\mathbf{Y}_i - \mathbf{W}_i' \beta) \mathbf{W}_i'. \quad (4.10)$$

The large sample theory for the estimator of β obtained via (4.10) is considerably more challenging than the cases discussed above, as well as cases covered in the most recent literature (e.g., Saegusa and Wellner, 2013). The challenge, as alluded to earlier, stems from the fact that $N_k^*(\mathbf{a})$ is not a smooth function of its argument. The substitution of $\hat{\alpha}$ for α (i.e., using \hat{N}_k^* in place of N_k^*) is anticipated to alter the asymptotic analysis of the previous section, and in particular the form of the second term in the variance estimate. However, it is less clear that such changes will have a noticeable impact in most settings of practical interest. In particular, in the setting of two stage designs, it is typically the case that the stage one sample size will be many times larger than the stage two sample size; hence, the variability contributed by the estimation of α should in general be tiny in comparison to the variability generated by the much smaller size of the stage two sample. Therefore, rather than attempt to develop detailed and comparatively intricate asymptotics here for the case where $\hat{\alpha}$ is estimated, we intend to explore the impact of estimating α , and the validity of these conjectures, through simulation studies in the following section.

4.3 Simulation Study

Simulation studies were carried out to evaluate the performance of the residual-based sampling design and to compare the RDS design to the ODS design and SRS. The population model considered is $Y_i = \beta_0 + X_i \beta_1 + Z_i \beta_2 + \epsilon_i$, where $X_i \sim N(0,1)$, $Z_i \sim N(0,1)$, and $\epsilon_i \sim N(0,1)$ are mutually independent. The parent

population contains $N = 2,000$ observations and the stage one variables, Y_i and X_i , are observed for the entire parent population. All simulation results are based on 10,000 independent datasets.

For SRS, the stage two sample consists of a simple random sample of size n taken from the entire parent population. For ODS, the subjects are stratified into three strata, \mathcal{V}_1 , \mathcal{V}_2 , and \mathcal{V}_3 , defined using percentiles of the observed outcome variable, Y_i , and random samples of size n_k are selected from \mathcal{V}_k , $k = 1, 2, 3$. For RDS, we consider two variations of the design in order to assess the impact of the stage one model estimation. First, we consider the case in which the subjects are stratified into three strata, \mathcal{V}_1^* , \mathcal{V}_2^* , and \mathcal{V}_3^* , defined using percentiles of the “true” stage one residuals (given by (4.5), with α given by (4.4)) and random samples of size n_k^* are selected from \mathcal{V}_k^* , $k = 1, 2, 3$. In other words, the stratification is performed assuming that α is known. This sampling design is referred to as RDS1. Then, the same sampling procedure is implemented with the strata defined using percentiles of the estimated stage one model residuals (given by (4.3)); that is, using $\hat{\alpha}$ in place of α . We will refer to this design as RDS2. Comparing the results for RDS1 and RDS2 will allow us to directly evaluate the impact of the first stage estimation on the stage two regression parameter and covariance estimates. In all results, the total stage two sample size is the same for each sampling design.

In Tables 1–4, the cutpoints used to define the strata for the ODS design, a_1 and a_2 , correspond to the 25th and 75th percentiles of Y_i . For the RDS1 design, the strata cutpoints, b_1 and b_2 , correspond to the 25th and 75th percentiles of $\tilde{\epsilon}_i^*$, and for the RDS2 design, the cutpoints correspond to the same percentiles of e_i^* . In Table 5, we evaluate the effect of changing these cutpoints. Weaver and Zhou (2005) also use this method for choosing the cutpoints in their ODS simulation

study; see Zhou et al. (2011) for an example in outcome-auxiliary-dependent sampling.

Regression parameter estimates for the SRS design are obtained by solving the complete data score function (4.6) or, equivalently, by solving the IPW estimating equation (4.7) with $w_i = 1/\pi_i$ and $\pi_i = n/N$ ($i = 1, \dots, N$). The regression estimates for the ODS design are obtained by solving the IPW estimating equation (4.8). Regression parameter estimates for RDS1 are obtained by solving the IPW estimating equation (4.9), and the IPW estimating equation (4.10) is used for RDS2. All simulations were conducted in R (R Development Core Team, 2005). The simulation code is available upon request.

In Table 1, we assess the impact of an increasing effect of the stage one covariate by considering increasing values of β_1 . For SRS, a simple random sample of size 400 is selected. For ODS, RDS1, and RDS2, samples of size 150 are taken from the tail strata, $\mathcal{V}_1, \mathcal{V}_3$ and $\mathcal{V}_1^*, \mathcal{V}_3^*$, and samples of size 100 are taken from the middle stratum, \mathcal{V}_2 and \mathcal{V}_2^* , resulting in a total stage two sample of size 400.

The bias and empirical standard error of the regression parameter estimates obtained under ODS, RDS1 and RDS2 are provided. The relative efficiency of the three designs versus SRS, defined as $\text{var}(\hat{\beta}_{SRS})/\text{var}(\hat{\beta}_{\text{other}})$, is also provided. ODS, RDS1, and RDS2 all lead to unbiased estimates of the regression parameters. For ODS, the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_2$ increase as β_1 increases, while $\hat{\beta}_1$ is largely unaffected. For RDS1 and RDS2, the standard errors are not affected by the increase in β_1 . Notably, the biases and standard errors are nearly identical for RDS1 and RDS2 for all values of β_1 .

In terms of relative efficiency, ODS outperforms SRS when β_1 is small. For larger values of β_1 , there is increased efficiency in the estimation of β_1 but a loss of efficiency in the estimation of β_0 and β_2 . This is to be expected when the sam-

Table 4.1: Increasing stage one effect

	ODS		RDS1		RDS2		RE vs. SRS		
	Bias	Emp SE	Bias	Emp SE	Bias	Emp SE	ODS	RDS1	RDS2
$\beta_0 = 1$	0.000	0.038	0.000	0.038	0.000	0.038	1.702	1.703	1.696
$\beta_1 = 0$	0.000	0.047	0.000	0.047	0.000	0.048	1.135	1.130	1.109
$\beta_2 = 0.5$	0.002	0.047	0.004	0.047	0.002	0.047	1.132	1.125	1.145
$\beta_0 = 1$	0.001	0.043	-0.003	0.038	0.000	0.039	1.398	1.766	1.809
$\beta_1 = 0.5$	0.001	0.049	-0.001	0.048	0.000	0.048	1.072	1.107	1.124
$\beta_2 = 0.5$	0.002	0.049	0.004	0.047	0.003	0.046	1.062	1.151	1.160
$\beta_0 = 1$	0.001	0.048	0.001	0.038	0.000	0.038	1.085	1.714	1.699
$\beta_1 = 1$	0.003	0.049	0.001	0.047	0.000	0.048	1.087	1.141	1.134
$\beta_2 = 0.5$	0.002	0.051	0.003	0.047	0.003	0.047	0.942	1.140	1.107
$\beta_0 = 1$	0.001	0.055	0.007	0.038	0.000	0.038	0.858	1.783	1.735
$\beta_1 = 2$	0.002	0.047	-0.001	0.047	0.000	0.047	1.134	1.131	1.137
$\beta_2 = 0.5$	0.001	0.056	0.002	0.047	0.003	0.047	0.821	1.169	1.160
$\beta_0 = 1$	0.001	0.057	0.000	0.038	0.000	0.038	0.786	1.742	1.764
$\beta_1 = 4$	0.001	0.045	0.000	0.046	0.000	0.047	1.229	1.169	1.143
$\beta_2 = 0.5$	0.001	0.057	0.002	0.047	0.002	0.047	0.757	1.141	1.143
$\beta_0 = 1$	0.000	0.058	0.001	0.038	0.000	0.038	0.751	1.693	1.731
$\beta_1 = 6$	0.001	0.044	0.000	0.048	0.000	0.048	1.224	1.047	1.067
$\beta_2 = 0.5$	-0.001	0.057	0.002	0.046	0.003	0.046	0.759	1.145	1.146

ODS: $n_1 = n_3 = 150, n_2 = 100$

RDS1, RDS2: $n_1^* = n_3^* = 150, n_2^* = 100$

RE = relative efficiency = $\text{var}(\hat{\beta}_{SRS})/\text{var}(\hat{\beta}_{\text{other}})$

pling depends only on Y_i and the effect of the stage one covariate is greater than that of the stage two covariate. RDS1 and RDS2, on the other hand, are more efficient than SRS for all three regression parameters in all cases. This result is also not surprising since the RDS design accounts for the increasing effect of β_1 by stratifying the parent population using the stage one model error terms or residuals. Modifying the ODS design to allow stratification to be based on both Y_i and X_i (e.g., Scott and Wild, 2011) may, in some situations, lead to efficiency gains more similar to what is observed with the RDS design. This is explored in Table 6. The relative efficiencies of RDS1 and RDS2 are quite similar and neither sampling procedure consistently outperforms the other. For reference, the relative efficiencies of the regression parameter estimates computed assuming that the stage two covariate was observed for the entire parent population versus SRS are 2.2203, 2.1962, and 2.2219, respectively, when $\beta_0 = 1, \beta_1 = 2, \beta_2 = 0.5$.

Table 4.2: Increasing stage two effect

	ODS		RDS1		RDS2		RE vs. SRS		
	Bias	Emp SE	Bias	Emp SE	Bias	Emp SE	ODS	RDS1	RDS2
$\beta_0 = 1$	0.000	0.038	0.000	0.031	0.000	0.031	1.712	2.598	2.575
$\beta_1 = 0.5$	0.003	0.047	0.000	0.044	0.001	0.044	1.148	1.301	1.319
$\beta_2 = 0$	0.000	0.047	0.001	0.044	0.001	0.044	1.155	1.297	1.312
$\beta_0 = 1$	0.001	0.043	0.000	0.038	0.000	0.038	1.398	1.766	1.809
$\beta_1 = 0.5$	0.001	0.049	0.000	0.048	0.000	0.048	1.072	1.107	1.124
$\beta_2 = 0.5$	0.002	0.048	0.004	0.047	0.003	0.046	1.062	1.151	1.160
$\beta_0 = 1$	-0.001	0.048	0.000	0.046	0.000	0.047	1.058	1.136	1.117
$\beta_1 = 0.5$	0.001	0.051	-0.001	0.051	0.000	0.052	0.952	0.968	0.945
$\beta_2 = 1$	0.002	0.049	0.002	0.048	0.004	0.049	1.072	1.089	1.072
$\beta_0 = 1$	0.001	0.054	0.002	0.053	0.000	0.054	0.848	0.875	0.841
$\beta_1 = 0.5$	0.000	0.055	0.000	0.056	-0.001	0.055	0.825	0.815	0.818
$\beta_2 = 2$	0.002	0.047	0.003	0.047	0.002	0.046	1.111	1.115	1.183
$\beta_0 = 1$	0.000	0.056	0.001	0.056	0.000	0.056	0.779	0.792	0.784
$\beta_1 = 0.5$	0.000	0.057	0.001	0.058	0.000	0.058	0.768	0.742	0.748
$\beta_2 = 4$	0.002	0.044	0.001	0.044	0.001	0.045	1.295	1.295	1.247
$\beta_0 = 1$	-0.001	0.058	0.000	0.058	-0.001	0.058	0.749	0.746	0.736
$\beta_1 = 0.5$	0.000	0.057	0.001	0.058	0.000	0.058	0.786	0.767	0.764
$\beta_2 = 6$	0.000	0.045	0.000	0.044	-0.001	0.045	1.212	1.251	1.235

ODS: $n_1 = n_3 = 150, n_2 = 100$

RDS1, RDS2: $n_1^* = n_3^* = 150, n_2^* = 100$

RE = relative efficiency = $\text{var}(\hat{\beta}_{SRS})/\text{var}(\hat{\beta}_{\text{other}})$

In Table 2, we assess the impact of an increasing effect of the stage two covariate by considering increasing values of β_2 . As β_2 increases, we see a similar pattern in the standard errors for ODS and as we did in Table 1. Specifically, as β_2 increases, the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ increase, while $\hat{\beta}_2$ is largely unaffected. In terms of relative efficiency, ODS, RDS1 and RDS2 each outperform SRS when β_2 is small. For larger β_2 values, there is increased efficiency in the estimation of β_2 but a loss of efficiency in the estimation of β_0 and β_1 . Again, the results for RDS1 and RDS2 are remarkably similar.

In Table 3, we assess the effect of increasing both β_1 and β_2 , as well as the effect of increasing β_1 and β_2 in opposite directions. Few meaningful patterns emerge with the exception that RDS1 and RDS2 are consistently more efficient than SRS and ODS at estimating β_2 with, in some cases, a loss of efficiency in estimating β_0 and β_1 .

Table 4.3: Increasing stage one and stage two effects

	ODS		RDS1		RDS2		RE vs. SRS		
	Bias	Emp SE	Bias	Emp SE	Bias	Emp SE	ODS	RDS1	RDS2
$\beta_0 = 1$	0.000	0.051	-0.001	0.047	0.000	0.047	0.951	1.138	1.139
$\beta_1 = 1$	0.001	0.051	0.000	0.051	0.000	0.051	0.967	0.962	0.964
$\beta_2 = 1$	0.002	0.051	0.003	0.047	0.003	0.048	0.985	1.135	1.130
$\beta_0 = 1$	0.001	0.056	0.000	0.053	-0.001	0.054	0.828	0.910	0.885
$\beta_1 = 2$	0.001	0.051	-0.001	0.056	0.000	0.055	0.974	0.834	0.839
$\beta_2 = 2$	0.001	0.051	0.002	0.046	0.003	0.046	0.945	1.150	1.154
$\beta_0 = 1$	0.000	0.057	0.000	0.058	0.000	0.057	0.781	0.770	0.779
$\beta_1 = 6$	0.000	0.052	0.000	0.057	0.000	0.057	0.940	0.755	0.760
$\beta_2 = 6$	0.002	0.051	0.001	0.044	0.001	0.044	0.939	1.273	1.259
$\beta_0 = 1$	0.001	0.050	0.000	0.047	0.000	0.046	1.002	1.163	1.182
$\beta_1 = -1$	-0.001	0.051	0.001	0.051	0.000	0.051	0.987	1.006	1.000
$\beta_2 = 1$	0.003	0.050	0.003	0.047	0.003	0.048	0.988	1.112	1.073
$\beta_0 = 1$	0.001	0.055	0.001	0.053	0.000	0.053	0.821	0.892	0.884
$\beta_1 = -2$	-0.001	0.051	0.000	0.056	0.001	0.056	0.958	0.804	0.796
$\beta_2 = 2$	0.001	0.051	0.002	0.046	0.002	0.047	0.956	1.171	1.148
$\beta_0 = 1$	0.001	0.058	0.000	0.057	0.001	0.057	0.749	0.754	0.750
$\beta_1 = -6$	0.000	0.052	0.001	0.058	0.001	0.058	0.944	0.747	0.766
$\beta_2 = 6$	0.001	0.052	0.001	0.045	0.000	0.044	0.965	1.297	1.328

ODS: $n_1 = n_3 = 150, n_2 = 100$

RDS1, RDS2: $n_1^* = n_3^* = 150, n_2^* = 100$

RE = relative efficiency = $\text{var}(\hat{\beta}_{SRS})/\text{var}(\hat{\beta}_{\text{other}})$

In Table 4, we consider correlated stage one and stage two covariates. Specifically, we assume that $X_i \sim N(0,1)$, $Z_i \sim N(0,1)$, and $\text{Cov}(X_i, Z_i) = 0.6$. RDS1 and RDS2 clearly have the best performance in this setting. The estimators are unbiased with reasonable standard errors and large efficiency gains over SRS for all three parameters. Additionally, RDS1 and RDS2 are more efficient than ODS when the regression parameter for the stage one covariate is non-zero.

In Table 5, we explore the effect of changing the strata sizes for ODS, RDS1, and RDS2 by setting the cutpoints $(a_1, a_2$ and $b_1, b_2)$ to correspond to the $(40^{th}, 60^{th})$, $(35^{th}, 65^{th})$, $(25^{th}, 75^{th})$, $(15^{th}, 85^{th})$, or $(10^{th}, 90^{th})$ percentiles of Y_i , $\tilde{\epsilon}_i^*$, and e_i^* , respectively. As the lower cutpoint decreases and the upper cutpoint increases, the upper and lower strata $(\mathcal{V}_1, \mathcal{V}_3$ and $\mathcal{V}_1^*, \mathcal{V}_3^*)$ become smaller and the stratified sampling from these strata identifies more observations with extreme values of Y_i , $\tilde{\epsilon}_i^*$, and e_i^* .

Table 4.4: Correlated stage one and stage two covariates

	ODS		RDS1		RDS2		RE vs. SRS		
	Bias	Emp SE	Bias	Emp SE	Bias	Emp SE	ODS	RDS1	RDS2
$\beta_0 = 1$	0.000	0.038	0.000	0.036	0.000	0.036	1.726	1.891	1.893
$\beta_1 = 0$	0.001	0.060	-0.001	0.058	-0.001	0.058	1.133	1.206	1.213
$\beta_2 = 0.5$	0.003	0.059	0.003	0.058	0.002	0.058	1.146	1.209	1.211
$\beta_0 = 1$	0.001	0.045	0.000	0.036	0.000	0.036	1.254	1.973	1.970
$\beta_1 = 0.5$	0.002	0.063	-0.002	0.057	-0.002	0.058	0.993	1.209	1.186
$\beta_2 = 0.5$	0.002	0.063	0.003	0.057	0.003	0.057	0.970	1.177	1.192
$\beta_0 = 1$	0.000	0.051	0.000	0.036	0.000	0.036	1.003	1.984	1.979
$\beta_1 = 1$	0.002	0.065	-0.001	0.058	-0.001	0.058	0.919	1.146	1.148
$\beta_2 = 0.5$	0.001	0.066	0.003	0.058	0.003	0.058	0.918	1.193	1.194
$\beta_0 = 1$	-0.001	0.055	0.000	0.036	-0.001	0.036	0.812	1.887	1.841
$\beta_1 = 2$	0.001	0.064	-0.001	0.058	-0.002	0.058	0.966	1.198	1.185
$\beta_2 = 0.5$	0.001	0.069	0.003	0.057	0.004	0.058	0.837	1.247	1.199
$\beta_0 = 1$	0.000	0.057	0.000	0.036	0.000	0.036	0.756	1.903	1.917
$\beta_1 = 4$	0.001	0.064	-0.002	0.058	-0.002	0.058	0.956	1.161	1.151
$\beta_2 = 0.5$	0.000	0.072	0.002	0.058	0.002	0.058	0.771	1.193	1.183
$\beta_0 = 1$	-0.001	0.057	-0.001	0.036	-0.001	0.036	0.794	1.959	2.012
$\beta_1 = 6$	0.001	0.062	-0.002	0.058	-0.001	0.057	1.004	1.158	1.192
$\beta_2 = 0.5$	0.001	0.072	0.004	0.058	0.002	0.057	0.746	1.164	1.205

ODS: $n_1 = n_3 = 150, n_2 = 100$ RDS1, RDS2: $n_1^* = n_3^* = 150, n_2^* = 100$ RE = relative efficiency = $\text{var}(\hat{\beta}_{SRS})/\text{var}(\hat{\beta}_{\text{other}})$

For all three designs, as the lower cutpoint decreases and the upper cutpoint increases (i.e., as you move down the table), the standard errors appear to first decrease and then increase. Correspondingly, the relative efficiencies compared to SRS increase and then decrease as you move down the table. This suggests that while targeting observations with small and large values of Y_i , $\tilde{\epsilon}_i^*$, or e_i^* may improve efficiency, focusing too much on the extreme tail values can actually be detrimental. In fact, for this simulation configuration it appears that the optimal cutpoints for ODS may be close to the (35th, 65th) percentiles of Y_i , while the optimal cutpoints for RDS1 and RDS2 may be close to the (25th, 75th) percentiles of $\tilde{\epsilon}_i^*$ and e_i^* , respectively.

In Table 6, we compare RDS1 and RDS2 with an outcome-dependent sampling design (ODS+) which allows the strata to depend on both the outcome variable and the stage one covariate by constructing strata on $\mathcal{Y} \times \mathcal{X}$. Specifi-

Table 4.5: Impact of strata definitions

cutpoints	ODS		RDS1		RDS2		RE vs. SRS		
	Bias	Emp SE	Bias	Emp SE	Bias	Emp SE	ODS	RDS1	RDS2
(40 th , 60 th)	0.000	0.048	0.000	0.037	0.001	0.037	1.141	1.903	1.940
	0.000	0.052	0.000	0.052	0.000	0.052	0.956	0.933	0.935
	0.001	0.051	0.000	0.051	0.000	0.050	0.950	0.952	0.980
(35 th , 65 th)	0.000	0.048	0.000	0.036	0.000	0.035	1.102	1.929	2.057
	0.001	0.049	0.000	0.049	0.000	0.049	1.077	1.062	1.046
	0.000	0.050	0.001	0.048	0.001	0.048	0.991	1.073	1.076
(25 th , 75 th)	0.001	0.055	0.001	0.038	0.000	0.038	0.858	1.783	1.735
	0.002	0.047	-0.001	0.047	0.000	0.047	1.134	1.131	1.137
	0.001	0.056	0.002	0.047	0.003	0.047	0.821	1.169	1.160
(15 th , 85 th)	-0.001	0.067	0.001	0.051	0.001	0.051	0.568	0.996	0.988
	0.004	0.054	0.000	0.055	0.000	0.054	0.870	0.843	0.861
	-0.001	0.069	0.005	0.053	0.006	0.054	0.539	0.903	0.872
(10 th , 90 th)	0.000	0.076	0.001	0.062	0.001	0.062	0.432	0.655	0.655
	0.005	0.062	0.000	0.063	0.000	0.063	0.668	0.635	0.640
	0.002	0.076	0.006	0.061	0.005	0.061	0.435	0.679	0.678

$$\beta_0 = 1, \beta_1 = 2, \beta_2 = 0.5$$

$$\text{ODS: } n_1 = n_3 = 150, n_2 = 100$$

$$\text{RDS1, RDS2: } n_1^* = n_3^* = 150, n_2^* = 100$$

$$\text{RE} = \text{relative efficiency} = \text{var}(\hat{\beta}_{SRS}) / \text{var}(\hat{\beta}_{\text{other}})$$

cally, we define four strata using cutpoints corresponding to the 50th percentiles of Y_i and X_i so that the plane is divided into quadrants which, beginning at the upper left and moving clockwise, we will denote $\mathcal{V}_1^\#$, $\mathcal{V}_2^\#$, $\mathcal{V}_3^\#$, and $\mathcal{V}_4^\#$. Samples of size 175 are taken from strata $\mathcal{V}_2^\#$ and $\mathcal{V}_4^\#$ and samples of size 25 are taken from strata $\mathcal{V}_1^\#$ and $\mathcal{V}_3^\#$ for a total stage two sample of size 400. Regression parameter estimates are obtained by solving the estimating equation (4.7) with weights equal to the inverse sampling probabilities for each subject.

Larger samples are taken from $\mathcal{V}_2^\#$ and $\mathcal{V}_4^\#$ because we know that there is a positive linear relationship between Y_i and Z_i . While it is possible that this information may not be available a priori in practical settings, we suspect that in most applications the direction of the relationship between the response variable and the stage two covariate will be known. We chose to use four strata and to define them as described in part for simplicity, but also in order to provide a reasonable comparison to the RDS design, which uses only three strata. The use

Table 4.6: ODS design including stage one covariate

	ODS+		RDS1		RDS2		RE vs. SRS		
	Bias	Emp SE	Bias	Emp SE	Bias	Emp SE	ODS+	RDS1	RDS2
$\beta_0 = 1$	0.000	0.050	0.000	0.038	0.000	0.038	0.983	1.703	1.696
$\beta_1 = 0$	0.001	0.058	0.000	0.047	0.000	0.048	0.763	1.123	1.109
$\beta_2 = 0.5$	0.002	0.069	0.004	0.047	0.002	0.047	0.504	1.125	1.145
$\beta_0 = 1$	0.003	0.043	0.000	0.038	0.000	0.038	1.339	1.766	1.809
$\beta_1 = 0.5$	0.002	0.050	0.000	0.048	0.000	0.048	1.021	1.107	1.124
$\beta_2 = 0.5$	0.001	0.059	0.004	0.047	0.003	0.046	0.683	1.151	1.160
$\beta_0 = 1$	0.000	0.042	0.001	0.038	0.000	0.038	1.440	1.714	1.699
$\beta_1 = 1$	0.001	0.046	0.001	0.047	0.000	0.048	1.185	1.141	1.134
$\beta_2 = 0.5$	0.000	0.053	0.003	0.047	0.003	0.047	0.869	1.140	1.107
$\beta_0 = 1$	-0.001	0.043	0.001	0.038	0.000	0.038	1.341	1.783	1.735
$\beta_1 = 2$	0.001	0.047	-0.001	0.047	0.000	0.047	1.169	1.131	1.137
$\beta_2 = 0.5$	-0.001	0.048	0.002	0.047	0.003	0.047	1.043	1.169	1.160
$\beta_0 = 1$	0.000	0.046	0.000	0.038	0.000	0.038	1.172	1.742	1.764
$\beta_1 = 4$	0.001	0.049	0.000	0.046	0.000	0.047	1.058	1.169	1.143
$\beta_2 = 0.5$	0.000	0.048	0.002	0.047	0.002	0.047	1.059	1.141	1.143
$\beta_0 = 1$	0.001	0.047	0.001	0.038	0.000	0.038	1.125	1.693	1.731
$\beta_1 = 6$	0.000	0.050	0.000	0.048	0.000	0.048	1.018	1.047	1.067
$\beta_2 = 0.5$	-0.001	0.049	0.002	0.046	0.003	0.046	0.998	1.145	1.149

ODS: $n_1 = n_3 = 150, n_2 = 100$
 ODS+: $n_1^\# = n_3^\# = 50, n_2^\# = n_4^\# = 150$
 RDS1, RDS2: $n_1^* = n_3^* = 150, n_2^* = 100$
 RE = relative efficiency = $\text{var}(\hat{\beta}_{SRS})/\text{var}(\hat{\beta}_{\text{other}})$

of more strata could certainly be considered and the strata could be constructed differently, but the number of possible strata can quickly become unwieldy and determining the appropriate sample sizes for each of the strata can be quite difficult with, potentially, minimal gain in the efficiency of estimation for the stage two regression parameter. This is an inherent advantage of the RDS design. The results in Table 6 show that ODS+ does in fact outperform SRS in some instances, but RDS1 and RDS2 are consistently more efficient than ODS+.

In Table 7, we evaluate the performance of the covariance estimators $\hat{\Sigma} = \hat{J}^{-1} - \hat{J}^{-1}\hat{G}\hat{J}^{-1}$ and $\hat{\Sigma}^* = \hat{J}^{*-1} - \hat{J}^{*-1}\hat{G}\hat{J}^{*-1}$ introduced in §4.2.1 and §4.2.2. For each design, we provide three estimates of the regression parameter standard errors. The first estimate, SE Est1, is computed using the true value of the regression parameter, β , and the error variance, $\sigma^2 = 1$. The second estimate,

Table 4.7: Standard error estimates

	ODS				RDS1				RDS2			
	Emp SE	SE Est1	SE Est2	SE Est3	Emp SE	SE Est1	SE Est2	SE Est3	Emp SE	SE Est1	SE Est2	SE Est3
$\beta_0 = 1$	0.039	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038
$\beta_1 = 0$	0.047	0.047	0.047	0.047	0.048	0.047	0.047	0.047	0.047	0.047	0.047	0.047
$\beta_2 = 0.5$	0.047	0.046	0.046	0.046	0.047	0.046	0.046	0.046	0.046	0.046	0.046	0.046
$\beta_0 = 1$	0.042	0.042	0.042	0.042	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038
$\beta_1 = 0.5$	0.049	0.048	0.048	0.048	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047
$\beta_2 = 0.5$	0.047	0.048	0.048	0.048	0.047	0.046	0.046	0.046	0.047	0.046	0.046	0.046
$\beta_0 = 1$	0.048	0.048	0.048	0.048	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038
$\beta_1 = 1$	0.049	0.048	0.048	0.048	0.048	0.047	0.047	0.047	0.047	0.047	0.047	0.047
$\beta_2 = 0.5$	0.051	0.051	0.051	0.051	0.047	0.046	0.046	0.046	0.047	0.046	0.046	0.046
$\beta_0 = 1$	0.054	0.053	0.053	0.053	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038
$\beta_1 = 2$	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047
$\beta_2 = 0.5$	0.055	0.054	0.054	0.054	0.047	0.046	0.046	0.046	0.047	0.046	0.046	0.046
$\beta_0 = 1$	0.056	0.056	0.056	0.056	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038
$\beta_1 = 4$	0.046	0.045	0.045	0.045	0.048	0.047	0.047	0.047	0.047	0.047	0.047	0.047
$\beta_2 = 0.5$	0.057	0.056	0.056	0.056	0.047	0.046	0.046	0.046	0.047	0.046	0.046	0.046
$\beta_0 = 1$	0.057	0.057	0.057	0.057	0.038	0.038	0.038	0.038	0.038	0.038	0.038	0.038
$\beta_1 = 6$	0.044	0.044	0.044	0.044	0.048	0.047	0.047	0.047	0.047	0.047	0.047	0.047
$\beta_2 = 0.5$	0.057	0.057	0.057	0.057	0.046	0.046	0.046	0.046	0.047	0.046	0.046	0.046

ODS: $n_1 = n_3 = 150, n_2 = 100$

RDS1: $n_1^* = n_3^* = 150, n_2^* = 100$

SE Est1: $\hat{\Sigma}$ and $\hat{\Sigma}^*$ computed using true values of β and σ^2

SE Est2: $\hat{\Sigma}$ and $\hat{\Sigma}^*$ computed using $\hat{\beta}$ for ODS, $\hat{\beta}^*$ for RDS, and true σ^2

SE Est3: $\hat{\Sigma}$ and $\hat{\Sigma}^*$ computed using $\hat{\beta}$ for ODS, $\hat{\beta}^*$ for RDS, and $\hat{\sigma}^2$

SE Est2, is computed using the appropriate regression parameter estimate for each sampling design and the true error variance, $\sigma^2 = 1$. The final estimate, SE Est3, is computed using the estimated regression parameter and a weighted estimate of the error variance. Evaluating the covariance estimate in these three ways allows us to assess the effect of substituting the estimated quantities into $\hat{\Sigma}$ and $\hat{\Sigma}^*$.

For all three designs, the SE Est1, SE Est2, and SE Est3 values are identical which indicates that substitution of the regression parameter estimates and error variance estimates into $\hat{\Sigma}$ and $\hat{\Sigma}^*$ does not impact the standard error estimates. Moreover, the standard error estimates are also virtually identical to the empirical standard errors.

4.4 Remarks

For the linear models setting, the proposed two-stage residual-dependent sampling design is easy to implement and allows the stage two sample to depend on both the observed response variable values and the stage one covariates. By summarizing the observed stage one data for each subject through a model-based residual, the issues inherent to the multidimensional stratification required by traditional outcome-dependent sampling designs which account for covariates are alleviated. Moreover, the simulation studies showed that the RDS design was more efficient than the ODS design in virtually every situation considered. All results can also be easily extended to the setting where there is a vector of stage two covariates that are of interest.

Estimation of the stage one regression parameters, which is required to implement the stratification and sampling procedure for the RDS design in practical settings, presents significant challenges to establishing the large sample properties of the IPW stage two estimators. However, due in large part to the fact that in most studies the stage one sample size tends to be significantly larger than the stage two sample size, the stage one estimation appears to have little impact on the stage two model estimates, as was clearly evidenced by the simulation studies. However, it is important to note that IPW estimators are known to be inefficient and hence, considering more efficient methods of estimation such as augmented IPW estimators would be a natural next step in the development of the RDS design.

In §4.1.2, we characterized an important relationship between the stage one regression parameter estimates and the stage two population model parameters. This development provides insight into the conditions under which the RDS design will successfully identify informative subjects for the stage two

sample. Further development on how to best exploit this relationship may lead to even greater efficiency gains and is another avenue for future research. One possibility would be to take a “pilot” stage two sample on which some of the quantities in (4.4) could be estimated and used to target informative subjects.

Finally, we have yet to consider RDS from a design standpoint. Specifically, methods for determining how to define the strata and how to select optimal strata-specific sample sizes would improve the utility of the RDS design. We intend to explore the design question in future work.

APPENDIX A
PROOFS FOR CHAPTER 2

The proof of Theorem 2.1.1 relies on the following pair of lemmas, both of which hold under conditions A1-A3. The proof of Lemma 1 is a direct consequence of the strong law of large numbers for U-statistics and results in Andersen and Gill (1982, Theorem II.1). The proof of Lemma 2 relies on this result and properties of the normal cumulative distribution and density functions.

Lemma 1 $\sup_{\beta \in \mathbb{B}} |L_n(\beta) - L_0(\beta)| \rightarrow 0$ almost surely, where $L_0(\beta)$ is convex for $\beta \in \mathbb{B}$.

Proof Recalling the notation from §2.1.1, let the survival data for cluster i be denoted $\mathcal{W}_i = \{W_{ik}, k = 1, \dots, K_i\}$. For $i, j = 1, \dots, n$ and $\beta \in \mathbb{B}$, define

$$h_\beta(\mathbf{W}_i, \mathbf{W}_j) = \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} \Delta_{ik} \{e_{ik}(\beta) - e_{jl}(\beta)\}^-,$$

where $\{x\}^- = |x|I\{x < 0\}$. Using (6) and defining $c_n = 1 - n^{-1}$, it is easily shown that

$$L_n(\beta) = \frac{1}{c_n} \frac{1}{n^2} \sum_{i=1}^n h_\beta(\mathbf{W}_i, \mathbf{W}_i) + \frac{1}{\binom{n}{2}} \sum_{i < j} \psi_\beta(\mathbf{W}_i, \mathbf{W}_j),$$

where

$$\psi_\beta(\mathbf{W}_i, \mathbf{W}_j) = \frac{1}{2} \{h_\beta(\mathbf{W}_i, \mathbf{W}_j) + h_\beta(\mathbf{W}_j, \mathbf{W}_i)\}.$$

Under Conditions A1-A3, we find $E[\psi_\beta(\mathbf{W}_i, \mathbf{W}_j)^2] < \infty$ for $i, j = 1, \dots, n$.

Hence, by the strong law of large numbers (Serfling, 1980, §1.8),

$$\frac{1}{n^2} \sum_{i=1}^n h_\beta(\mathbf{W}_i, \mathbf{W}_i) \rightarrow 0$$

almost surely. Using the strong law of large numbers for U-statistics (Serfling, 1980, §5.4), we further obtain the almost sure convergence

$$\frac{1}{\binom{n}{2}} \sum_{i < j} \psi_\beta(\mathbf{W}_i, \mathbf{W}_j) \rightarrow E[\psi_\beta(\mathbf{W}_1, \mathbf{W}_2)] := L_0(\beta).$$

Since $c_n \rightarrow 1$, it now follows that $L_n(\beta) \rightarrow L_0(\beta)$ almost surely for each $\beta \in \mathbb{B}$. Since $L_n(\beta)$ is convex for $\beta \in \mathbb{B}$, $L_0(\beta)$ is convex for $\beta \in \mathbb{B}$ and the convergence is uniform on the compact set \mathbb{B} (Andersen and Gill, 1982, Theorem II.1). \square

Lemma 2 $\sup_{\beta \in \mathbb{B}} |\tilde{L}_n(\beta) - L_0(\beta)| \rightarrow 0$ almost surely, where $L_0(\cdot)$ is defined in Lemma 1.

Proof By the triangle inequality,

$$|\tilde{L}_n(\beta) - L_0(\beta)| \leq |\tilde{L}_n(\beta) - L_n(\beta)| + |L_n(\beta) - L_0(\beta)|$$

for each $\beta \in \mathbb{B}$. From Lemma 1, we have $\sup_{\beta \in \mathbb{B}} |L_n(\beta) - L_0(\beta)| \rightarrow 0$ almost surely; hence, it suffices to show that $\sup_{\beta \in \mathbb{B}} |\tilde{L}_n(\beta) - L_n(\beta)| \rightarrow 0$ almost surely. Using (6) and (9),

$$|\tilde{L}_n(\beta) - L_n(\beta)| = \frac{1}{n(n-1)} \left| \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \{e_{jl}(\beta) - e_{ik}(\beta)\} D_{ikjl}^{(n)}(\beta) \right|,$$

where

$$D_{ikjl}^{(n)}(\beta) = H_{ikjl}^{(n)}(\beta) + \frac{r_{ikjl}}{n^{1/2}\{e_{jl}(\beta) - e_{ik}(\beta)\}} h_{ikjl}^{(n)}(\beta) - I\{e_{ik}(\beta) - e_{jl}(\beta) \leq 0\}.$$

Then, $|\tilde{L}_n(\beta) - L_n(\beta)| \leq Q_1(\beta) + Q_2(\beta)$, where

$$\begin{aligned} Q_1(\beta) &= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \left| \{e_{jl}(\beta) - e_{ik}(\beta)\} \left[H_{ikjl}^{(n)}(\beta) \right. \right. \\ &\quad \left. \left. - I\{e_{ik}(\beta) - e_{jl}(\beta) \leq 0\} \right] \right| \\ &= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} r_{ikjl} \left| \frac{\{e_{jl}(\beta) - e_{ik}(\beta)\}}{r_{ikjl}} \left[H_{ikjl}^{(n)}(\beta) \right. \right. \\ &\quad \left. \left. - I\left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} \geq 0 \right\} \right] \right| \end{aligned}$$

and

$$\begin{aligned} Q_2(\beta) &= \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \left| \frac{r_{ikjl}}{n^{1/2}\{e_{jl}(\beta) - e_{ik}(\beta)\}} h_{ikjl}^{(n)}(\beta) \right| \\ &= \frac{1}{n^{3/2}(n-1)} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \left\{ \frac{r_{ikjl}}{e_{jl}(\beta) - e_{ik}(\beta)} \right\}^2 \left| \frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}} h_{ikjl}^{(n)}(\beta) \right|. \end{aligned}$$

For $u \in \mathbb{R}$, it is easily shown that

$$|u \{ \Phi(n^{1/2}u) - I(u \geq 0) \}| = \begin{cases} u \Phi(-n^{1/2}u) & u > 0 \\ -u \Phi(n^{1/2}u) & u < 0 \\ 0 & u = 0 \end{cases}$$

and hence

$$\lim_{n \rightarrow \infty} \sup_{u \in \mathbb{R}} |u \{ \Phi(n^{1/2}u) - I(u \geq 0) \}| = 0.$$

Using (10), it follows that $\sup_{\beta \in \mathbb{B}} Q_1(\beta) \rightarrow 0$ almost surely as $n \rightarrow \infty$. Furthermore, since

$$\sup_{u \in \mathbb{R}} |u \phi(n^{1/2}u)| = \frac{1}{n^{1/2}},$$

it follows that

$$\lim_{n \rightarrow \infty} \sup_{u \in \mathbb{R}} |u \phi(n^{1/2}u)| = 0$$

and, again using (10), $\sup_{\beta \in \mathbb{B}} Q_2(\beta) \rightarrow 0$ almost surely. Thus

$$\sup_{\beta \in \mathbb{B}} |\tilde{L}_n(\beta) - L_n(\beta)| \rightarrow 0$$

almost surely, as desired. We remark here that, as in Lemma 1, the convexity of $\tilde{L}_n(\beta)$ implies that demonstrating pointwise convergence would have been sufficient to guarantee uniform convergence. \square

Proof of Theorem 2.1.1 Lemmas 1 and 2 respectively establish the uniform almost sure convergence of $L_n(\beta)$ and $\tilde{L}_n(\beta)$ to the convex function $L_0(\beta)$ for $\beta \in \mathbb{B}$. By condition A4, $L_0(\beta)$ is strictly convex at β_0 and β_0 is a unique minimizer. The respective minimizers $\hat{\beta}_n$ and $\tilde{\beta}_n$ of $L_n(\beta)$ and $\tilde{L}_n(\beta)$ thus converge almost surely to β_0 (Andersen and Gill, 1982, Corollary II.2). \square

The next lemma is required in order to prove Theorem 2.1.2; an abbreviated proof of this result and also Theorem 2.1.2 are provided below, with expanded

versions of these arguments available in a technical report. A fact used in proving Lemma 3 is that condition A4, in conjunction with (A.1), implies that the probability that $X_{1k} \neq X_{2l}$ for at least one (k, l) pair must be positive.

Lemma 3 *Under A1–A6 and as $n \rightarrow \infty$, $\|\nabla \tilde{S}_n(\beta_0) - A\| \rightarrow 0$ almost surely, where $A = \nabla S_0(\beta_0)$,*

$$\begin{aligned} \nabla S_0(\beta_0) = \frac{1}{2} \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} \int_{-\infty}^{\infty} E \{ (X_{1k} - X_{2l})(X_{1k} - X_{2l})' \bar{G}_{1k}(u) \bar{G}_{2l}(u) \} \{ f_0^2(u) \\ + f_0'(u) \bar{F}_0(u) \} du, \end{aligned} \quad (\text{A.1})$$

$\bar{G}_{rs}(\cdot)$ denotes the survivor function of $\log C_{rs} - X'_{rs}\beta_0$ and $\bar{F}_0(s) = \int_s^{\infty} f_0(u)du$ for every s .

Proof Using similar notation and proceeding as in the proof of Lemma 1, we begin by noting that for $i, j = 1, \dots, n$ and $\beta \in \mathbb{B}$, we may write

$$S_n(\beta) = \frac{1}{c_n} \frac{1}{n^2} \sum_{i=1}^n h_{\beta}^*(\mathbf{W}_i, \mathbf{W}_i) + \frac{1}{\binom{n}{2}} \sum_{i < j} \psi_{\beta}^*(\mathbf{W}_i, \mathbf{W}_j),$$

where

$$h_{\beta}^*(\mathbf{W}_i, \mathbf{W}_j) = \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} \Delta_{ik}(X_{ik} - X_{jl}) I \{ e_{ik}(\beta_0) - e_{jl}(\beta_0) + (\beta_0 - \beta)'(X_{ik} - X_{jl}) \leq 0 \}$$

and

$$\psi_{\beta}^*(\mathbf{W}_i, \mathbf{W}_j) = \frac{1}{2} \{ h_{\beta}^*(\mathbf{W}_i, \mathbf{W}_j) + h_{\beta}^*(\mathbf{W}_j, \mathbf{W}_i) \}.$$

Under the indicated conditions, it is easily seen that

$$E \{ S_n(\beta) \} = E \{ \psi_{\beta}^*(\mathbf{W}_1, \mathbf{W}_2) \} + O(n^{-1}),$$

where the $O(\cdot)$ term holds uniformly on $\beta \in \mathbb{B}$. Denote

$$S_0(\beta) = E \{ \psi_{\beta}^*(\mathbf{W}_1, \mathbf{W}_2) \};$$

with $\bar{G}_{rs}(\cdot)$ denoting the survivor function of $\log C_{rs} - X'_{rs}\beta_0$ and $\bar{F}_0(s) = \int_s^\infty f_0(u)du$, straightforward computations show that

$$S_0(\beta) = E \left[\sum_{k=1}^{K_1} \sum_{l=1}^{K_2} \{M_{1k,2l}(\beta) - M_{2l,1k}(\beta)\} \right] \quad (\text{A.2})$$

where

$$M_{ab,cd}(\beta) = \int_{-\infty}^{\infty} \bar{G}_{ab}(u) \bar{G}_{cd}\{u+(\beta_0-\beta)'(X_{ab}-X_{cd})\} \bar{F}_0\{u+(\beta_0-\beta)'(X_{ab}-X_{cd})\} f_0(u)du$$

and the outer expectation in (A.2) is understood to be taken over the joint distribution of the covariates. Evidently, $S_0(\beta_0) = 0$.

Conditions A1-A7 permit us to differentiate (A.2) directly; doing so and evaluating the result at $\beta = \beta_0$, we obtain (Fyngenson and Ritov, 1994, p. 737)

$$\begin{aligned} \nabla S_0(\beta_0) = \frac{1}{2} \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} \int_{-\infty}^{\infty} E \{ (X_{1k} - X_{2l})(X_{1k} - X_{2l})' \bar{G}_{1k}(u) \bar{G}_{2l}(u) \} \{ f_0^2(u) \\ + f_0'(u) \bar{F}_0(u) \} du, \end{aligned}$$

establishing (A.1). Note that condition A4, in conjunction with (A.1), implies that the probability that $X_{1k} \neq X_{2l}$ for at least one (k, l) pair is positive.

We will now proceed similarly using the smoothed equation (8), first writing

$$\tilde{S}_n(\beta) = \frac{1}{c_n} \frac{1}{n^2} \sum_{i=1}^n \tilde{h}_\beta(\mathbf{W}_i, \mathbf{W}_i) + \frac{1}{\binom{n}{2}} \sum_{i < j} \tilde{\psi}_\beta(\mathbf{W}_i, \mathbf{W}_j),$$

where

$$\tilde{h}_\beta(\mathbf{W}_i, \mathbf{W}_j) = \sum_{k=1}^{K_i} \sum_{l=1}^{K_j} \Delta_{ik}(X_{ik} - X_{jl}) \Phi \left[n^{1/2} \left\{ \frac{e_{jl}(\beta) - e_{ik}(\beta)}{\sqrt{r_{ikjl}}} \right\} \right]$$

and

$$\tilde{\psi}_\beta(\mathbf{W}_i, \mathbf{W}_j) = \frac{1}{2} \left\{ \tilde{h}_\beta(\mathbf{W}_i, \mathbf{W}_j) + \tilde{h}_\beta(\mathbf{W}_j, \mathbf{W}_i) \right\}.$$

Differentiating this representation with respect to β , setting $\beta = \beta_0$, and then using the strong laws of large numbers in a manner similar to Lemma 1, we find

$$\nabla \tilde{S}_n(\beta_0) \rightarrow \frac{1}{2} \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} E \{ (X_{1k} - X_{2l})(X_{1k} - X_{2l})' (\mathcal{A}_{1k,2l} + \mathcal{A}_{2l,1k}) \} \quad (\text{A.3})$$

almost surely as $n \rightarrow \infty$, where the expectation is taken with respect to the covariate distribution. The random variable $\mathcal{A}_{ab,cd}$ is defined to be zero with probability one if $X_{ab} = X_{cd}$; otherwise,

$$\mathcal{A}_{ab,cd} = \lim_{n \rightarrow \infty} k_n \iint_{\mathbb{R}^2} \bar{G}_{ab}(u) f_0(u) \xi_{cd}(s) \phi\{k_n(u-s)\} duds, \quad (\text{A.4})$$

where $k_n = n^{1/2}/r_{abcd}$, $r_{abcd}^2 = (X_{ab} - X_{cd})' \Sigma (X_{ab} - X_{cd}) > 0$, and $\xi_{cd}(s) = f_0(s) \bar{G}_{cd}(s) + \bar{F}_0(s) g_{cd}(s)$. An easy calculation shows that (A.4) may be equivalently written as

$$\mathcal{A}_{ab,cd} = \int_{-\infty}^{\infty} \bar{G}_{ab}(u) f_0(u) \xi_{cd}(u) du + \lim_{n \rightarrow \infty} \left(\frac{m_n}{\pi} \right)^{1/2} \int_{-\infty}^{\infty} \tau(w) e^{-m_n w^2} dw,$$

where $m_n = k_n^2/2$ and

$$\tau(w) = \int_{-\infty}^{\infty} \bar{G}_{ab}(u) f_0(u) \{ \xi_{cd}(w+u) - \xi_{cd}(u) \} du.$$

Under conditions A1-A6, $\tau(\cdot)$ is integrable, continuous, and bounded on \mathbb{R} with $\tau(0) = 0$. Using results in Kanwal (1998, p. 11), it follows that

$$\lim_{n \rightarrow \infty} \left(\frac{m_n}{\pi} \right)^{1/2} \int_{-\infty}^{\infty} \tau(w) e^{-m_n w^2} dw = \tau(0)$$

and the second term on the right-hand side therefore vanishes. Using the resulting formula for $\mathcal{A}_{ab,cd}$ and integration by parts, it can be shown that

$$\mathcal{A}_{ab,cd} = \int_{-\infty}^{\infty} \bar{G}_{ab}(u) f_0(u) \{ f_0(u) \bar{G}_{cd}(u) + \bar{F}_0(u) g_{cd}(u) \} du.$$

Similarly to (A.1), it can now be shown that

$$\mathcal{A}_{1k,2l} + \mathcal{A}_{2l,1k} = \int_{-\infty}^{\infty} \bar{G}_{1k}(u) \bar{G}_{2l}(u) \{ f_0^2(u) + f_0'(u) \bar{F}_0(u) \} du;$$

substituting this result into (A.3), we observe agreement with (A.1), proving the result. \square

Proof of Theorem 2.1.2 Using notation introduced in §2.1.2, we have that $A^{-1}n^{1/2}S_n(\beta_0)$ is asymptotically normal with mean zero and variance $A^{-1}\Omega A^{-1}$ under assumptions A1-A5 (Jin et al., 2006a, Theorem 5). Suppose that

$$n^{1/2}(\tilde{\beta}_n - \beta_0) + A^{-1}n^{1/2}S_n(\beta_0) \xrightarrow{p} 0. \quad (\text{A.5})$$

Then, it follows that $n^{1/2}(\tilde{\beta}_n - \beta_0) \rightarrow N(0, A^{-1}\Omega A^{-1})$ in distribution, establishing the desired asymptotic result as well as the equality of the limiting distributions of the smoothed and unsmoothed estimators.

To prove that (A.5) holds, we make use of Theorem 3 of Arcones (1998). Using notation from Arcones (1998), define $G_n(\beta) = n\tilde{L}_n(\beta)$ for all β in \mathbb{B} . For $n \geq 1$, define the sequence of $p \times 1$ random vectors $\eta_n = n^{1/2}S_n(\beta_0)$ and the sequences of nonsingular, symmetric $p \times p$ matrices $M_n = n^{1/2}I_p$ and $V_n = (1/2)A$. The required result (A.5) becomes

$$M_n(\tilde{\beta}_n - \beta_0) + \frac{1}{2}V_n^{-1}\eta_n \rightarrow 0 \quad (\text{A.6})$$

in probability. The result (A.6) follows directly from Arcones (1998, Theorem 3) provided that conditions A1-A7 are sufficient to ensure that the following regularity conditions hold.

- B1. $G_n(\beta)$ is convex and $\tilde{\beta}_n$ is a sequence satisfying $G_n(\tilde{\beta}_n) \leq \inf_{\beta \in \mathbb{B}} G_n(\beta) + o_p(1)$.
- B2. $\eta_n = O_p(1)$, $\liminf_{n \rightarrow \infty} \inf_{|\beta|=1} \beta'V_n\beta > 0$, and $\limsup_{n \rightarrow \infty} \sup_{|\beta|=1} \beta'V_n\beta < \infty$.
- B3. For each $\beta \in \mathbb{R}^p$, $G_n(\beta_0 + M_n^{-1}\beta) - G_n(\beta_0) - \beta'\eta_n - \beta'V_n\beta = o_p(1)$.

To establish condition B1, we first note that the convexity of $G_n(\beta)$ follows from the convexity of $H(t) = G_n(\beta + td) = n\tilde{L}_n(\beta + td)$ for $t \in \mathbb{R}$, where d is any

$p \times 1$ vector. Notice that

$$H(t) = \frac{1}{n-1} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \left[\{e_{jl}(\beta + td) - e_{ik}(\beta + td)\} H_{ikjl}^{(n)}(\beta + td) + \frac{r_{ikjl}}{n^{1/2}} h_{ikjl}^{(n)}(\beta + td) \right].$$

Using properties of the standard normal distribution, it is easily shown that

$$\frac{d^2 H(t)}{dt^2} = \frac{1}{n-1} \sum_{i=1}^n \sum_{k=1}^{K_i} \sum_{j=1}^n \sum_{l=1}^{K_j} \Delta_{ik} \frac{n^{1/2}}{r_{ikjl}} \{(X_{ik} - X_{jl})' d\}^2 h_{ikjl}^{(n)}(\beta + td).$$

Since this derivative is positive for $t \in \mathbb{R}$ when $d \neq 0$, $H(t)$ is convex and hence $G_n(\beta)$ is convex. Condition B1 now follows since $\tilde{\beta}_n = \operatorname{argmin}_{\beta \in \mathbb{B}} \tilde{L}_n(\beta) = \operatorname{argmin}_{\beta \in \mathbb{B}} G_n(\beta)$, $G_n(\beta)$ is continuous, and \mathbb{B} is compact.

To establish condition B2, we recall first that Conditions A1-A5 are sufficient to ensure that $\eta_n = n^{1/2} S_n(\beta_0)$ converges in distribution, hence $O_p(1)$ as required. Moreover, since $S_0(\beta_0) = 0$ and $V_n = (1/2)A$ is a positive definite matrix for $n \geq 1$, it follows that

$$\liminf_{n \rightarrow \infty} \inf_{|\beta|=1} \beta' V_n \beta > 0 \text{ and } \limsup_{n \rightarrow \infty} \sup_{|\beta|=1} \beta' V_n \beta < \infty,$$

establishing condition B2.

It remains to establish condition B3. Consider the Taylor series expansion

$$G_n(\beta_0 + M_n^{-1} \beta) = G_n(\beta_0) + (M_n^{-1} \beta)' \nabla G_n(\beta_0) + \frac{1}{2} (M_n^{-1} \beta)' \nabla^2 G_n(\beta_n^*) (M_n^{-1} \beta) + o_p(1)$$

where $\|\beta_n^* - \beta_0\| \leq \|M_n^{-1} \beta\|$ and ∇^2 denotes the second derivative with respect to β . Using the definitions of M_n , $G_n(\cdot)$, $\tilde{L}_n(\cdot)$ and $\tilde{S}_n(\cdot)$, this equality can be rewritten

$$G_n(\beta_0 + M_n^{-1} \beta) - G_n(\beta_0) - n^{1/2} \beta' \nabla \tilde{L}_n(\beta_0) - \frac{1}{2} \beta' \{\nabla^2 \tilde{L}_n(\beta_n^*)\} \beta = o_p(1)$$

or, equivalently,

$$G_n(\beta_0 + M_n^{-1} \beta) - G_n(\beta_0) - \beta' \{n^{1/2} \tilde{S}_n(\beta_0)\} - \frac{1}{2} \beta' \{\nabla \tilde{S}_n(\beta_n^*)\} \beta = o_p(1). \quad (\text{A.7})$$

Therefore, if

$$\left\| \nabla \tilde{S}_n(\beta_n^*) - A \right\| \rightarrow 0 \quad (\text{A.8})$$

in probability and

$$n^{1/2} \left\| \tilde{S}_n(\beta_0) - S_n(\beta_0) \right\| \rightarrow 0 \quad (\text{A.9})$$

in probability, the definitions of V_n and η_n imply that

$$G_n(\beta_0 + M_n^{-1}\beta) - G_n(\beta_0) - \beta' \eta_n - \beta' V_n \beta \rightarrow 0$$

in probability for all $\beta \in \mathbb{R}^p$ and, hence, that condition B3 holds.

To establish (A.8), we first use the triangle inequality to obtain

$$\left\| \nabla \tilde{S}_n(\beta_n^*) - A \right\| \leq \left\| \nabla \tilde{S}_n(\beta_n^*) - \nabla \tilde{S}_n(\beta_0) \right\| + \left\| \nabla \tilde{S}_n(\beta_0) - A \right\|.$$

Since $\{\tilde{S}_n\}$ is a sequence of bounded, continuously differentiable functions and $\beta_n^* \rightarrow \beta_0$ almost surely, the first term on the right-hand side is $o_p(1)$. Furthermore, since Lemma 3 implies that the second term is also $o_p(1)$, the required convergence result (A.8) holds.

To see that (A.9) holds, write

$$\tilde{S}_n(\beta_0) - S_n(\beta_0) = \int_{\mathbb{R}^p} \{S_n(\beta_0 + n^{-1/2}u) - S_n(\beta_0)\} \psi(u) du,$$

where $\psi(\cdot)$ denotes the pdf of ΓZ . Let Θ be a fixed matrix such that $\|\Theta\| \leq M$ for some $M < \infty$. Noting that $\int_{\mathbb{R}^p} u \psi(u) du = 0$, the right-hand side of the previous expression is evidently equal to

$$\int_{\mathbb{R}^p} \{S_n(\beta_0 + n^{-1/2}u) - S_n(\beta_0) - n^{-1/2}\Theta u\} \psi(u) du;$$

it immediately follows that

$$n^{1/2} \left\| \tilde{S}_n(\beta_0) - S_n(\beta_0) \right\| = n^{1/2} \left\| \int_{\mathbb{R}^p} \{S_n(\beta_0 + n^{-1/2}u) - S_n(\beta_0) - n^{-1/2}\Theta u\} \psi(u) du \right\|.$$

For suitable u , define the function

$$K_n(u; \beta_0, \Theta) = \left\| S_n(\beta_0 + n^{-1/2}u) - S_n(\beta_0) - n^{-1/2}\Theta u \right\|.$$

Then, the triangle inequality implies

$$\begin{aligned} n^{1/2} \left\| \tilde{S}_n(\beta_0) - S_n(\beta_0) \right\| &\leq n^{1/2} \int_{\|u\| \leq \epsilon_n} K_n(u; \beta_0, \Theta) \psi(u) du \\ &+ n^{1/2} \int_{\|u\| > \epsilon_n} K_n(u; \beta_0, \Theta) \psi(u) du \end{aligned} \quad (\text{A.10})$$

for any $\epsilon_n > 0$. The result (A.9) therefore holds if we can find $\epsilon_n > 0$ such that both integrals on the right-hand side of (A.10) converge in probability to zero.

Following Ying (1993, Theorem 2) and Jin et al. (2006a, Theorem 5), the matrix A satisfies

$$\sup_{\|b - \beta_0\| \leq d_n} \frac{\|S_n(b) - S_n(\beta_0) - A(b - \beta_0)\|}{1 + n^{1/2} \|b - \beta_0\|} = o_p(n^{-1/2}). \quad (\text{A.11})$$

for any positive sequence $d_n \rightarrow 0$. Suppose $\epsilon_n = o(n^{1/2})$. Then, taking $b = \beta_0 + n^{-1/2}u$, $d_n = n^{-1/2}\epsilon_n$, and $\Theta = A$, (A.11) implies

$$\sup_{\|u\| \leq \epsilon_n} \frac{\|K_n(u; \beta_0, A)\|}{1 + \|u\|} = o_p(n^{-1/2}).$$

It follows that

$$\begin{aligned} n^{1/2} \int_{\|u\| \leq \epsilon_n} K_n(u; \beta_0, A) \psi(u) du &= n^{1/2} \int_{\|u\| \leq \epsilon_n} \frac{K_n(u; \beta_0, A)}{1 + \|u\|} (1 + \|u\|) \psi(u) du \\ &\leq n^{1/2} \left\{ \sup_{\|u\| \leq \epsilon_n} \left\| \frac{K_n(u; \beta_0, A)}{1 + \|u\|} \right\| \right\} \int_{\|u\| \leq \epsilon_n} (1 + \|u\|) \psi(u) du \\ &= o_p(1) \int_{\|u\| \leq \epsilon_n} (1 + \|u\|) \psi(u) du. \end{aligned}$$

By condition A6, $E(\|\Gamma Z\|) < \infty$; hence, $\int_{\|u\| \leq \epsilon_n} (1 + \|u\|) \psi(u) du$ remains bounded, even if $\epsilon_n \rightarrow \infty$, proving that

$$n^{1/2} \int_{\|u\| \leq \epsilon_n} K_n(u; \beta_0, A) \psi(u) du \rightarrow 0$$

in probability.

With regard to the second term on the right-hand side of (A.10), we may use the definition of $K_n(\cdot; \beta_0, A)$ and the triangle inequality to write

$$n^{1/2} \int_{\|u\| > \epsilon_n} K_n(u; \beta_0, A) \psi(u) du \leq Q_3 + Q_4,$$

where

$$Q_3 = \left\{ \sup_{\|u\| > \epsilon_n} \|S_n(\beta_0 + n^{-1/2}u) - S_n(\beta_0)\| \right\} n^{1/2} \int_{\|u\| > \epsilon_n} \psi(u) du,$$

$$\text{and } Q_4 = \|A\| \int_{\|u\| > \epsilon_n} \|u\| \psi(u) du.$$

For all $\beta \in \mathbb{B}$, $\|S_n(\beta)\| \leq Q$ for some constant $Q < \infty$ by condition A2; hence,

$$Q_3 \leq 2Qn^{1/2}P(\|\Gamma Z\| > \epsilon_n).$$

However, under condition A6 and assuming that $\epsilon_n \rightarrow \infty$, it follows that $n^{1/2}P(\|\Gamma Z\| > \epsilon_n) \rightarrow 0$ as $n \rightarrow \infty$ and hence $Q_3 \rightarrow 0$ in probability. Similarly, $\int_{\|u\| > \epsilon_n} \|u\| \psi(u) du \rightarrow 0$, and therefore, $Q_4 \rightarrow 0$ in probability. Thus, provided that $n, \epsilon_n \rightarrow \infty$, we have

$$n^{1/2} \int_{\|u\| > \epsilon_n} K_n(u; \beta_0, A) \psi(u) du \rightarrow 0$$

in probability. Since we can select a sequence $\epsilon_n = o(n^{1/2})$ such both $n, \epsilon_n \rightarrow \infty$, it follows that (A.10), and hence (A.9), converge in probability to zero as $n \rightarrow \infty$, establishing (A.7) and concluding the proof. \square

APPENDIX B
PROOFS FOR CHAPTER 3

Proof of Theorem 3.1.2 Given the definitions of §3.1.4 and using standard rules for iterated expectations, it follows immediately that $E(N_{ik}) = E(U_i H_{ik})$, where $N_{ik} = N_{ik}(\tau)$ and $H_{ik} = H_{ik}(\tau)$. This equality evidently holds when aggregating over observations within a cluster; that is, $E(N_i) = E(U_i H_i)$. Since the martingales $M_{ik}(\cdot)$ are assumed to be orthogonal across k given W_i , it further follows that

$$\begin{aligned} E(M_i^2) &= E \left\{ E(M_i^2 | W_i) \right\} = E \left\{ \sum_{k=1}^{K_i} E(\langle M_{ik} \rangle | W_i) \right\} \\ &= E \left(\sum_{k=1}^{K_i} U_i H_{ik} \right) = E(U_i H_i) = E(N_i). \end{aligned} \quad (\text{B.1})$$

Now, considering only $E(M_i^2)$, we instead expand $M_i^2 = (N_i - W_i H_i)^2$; standard rules for iterated expectations and simple algebra show

$$E(M_i^2) = E(N_i^2) - 2E(N_i U_i H_i) + E(U_i^2 H_i^2) + E\{(V_i - U_i^2)H_i^2\}.$$

Rewriting the term on the right-hand side of the above expression and recalling that $E(M_i^2) = E(U_i H_i)$ from equation (B.1), we obtain the identity

$$E\{(N_i - U_i H_i)^2\} + E\{(V_i - U_i^2)H_i^2\} = E(U_i H_i),$$

establishing (3.15). Since $E(U_i H_i) = E(N_i)$, equation (3.16) also follows immediately. \square

APPENDIX C

THEOREMS AND PROOFS FOR CHAPTER 4

We will consider the asymptotic properties of IPW estimators for a general two-stage sampling design with stratified sampling. This general setting includes normal linear regression models under the ODS and RDS with known α designs as special cases. We assume that, had the independent and identically distributed sample (Y_i, \mathbf{W}_i) ($i = 1 \dots N$) been observed, the parameter of interest β is estimated as the solution to the score function

$$S_F(\beta) = \sum_{i=1}^N S_\beta(Y_i, \mathbf{W}_i)$$

where $S_\beta(Y_i, \mathbf{W}_i)$ is the contribution to the score function for subject i under the given full data likelihood function.

We will assume that the two-stage stratified sampling design entails stratifying the parent population into K strata, $\mathcal{V}_1, \dots, \mathcal{V}_K$, at stage one and selecting a random sample of fixed size n_k from strata \mathcal{V}_k at stage two. We remark here that the notation \mathcal{V}_k is meant to be generic and not necessarily the same as that used in §4.1.1. Under this sampling design, the IPW estimating equation has the form

$$S_W(\beta; \mathbf{w}) = \sum_{i=1}^N R_i w_i S_\beta(Y_i, \mathbf{W}_i) = \sum_{i=1}^N \sum_{k=1}^K R_i \frac{N_k}{n_k} \Delta_{ik} S_\beta(Y_i, \mathbf{W}_i)$$

where Δ_{ik} is an indicator of inclusion for the i th subject in strata \mathcal{V}_k , $N_k = \sum_{i=1}^N \Delta_{ik}$, and w_i is the inverse sampling probability for subject i ; that is, $w_i = 1/\pi_i$, where $\pi_i = n_k/N_k$ for the unique k such that $\Delta_{ik} = 1$. We will assume that the sampling probabilities and strata inclusion are known given the stage one data (e.g. Breslow and Wellner, 2007); hence, the data missing for subjects not included in the stage two sample are necessarily missing at random.

Let $\hat{\beta}$ denote the solution to $S_W(\hat{\beta}; \mathbf{w}) = 0$ or, equivalently, a minimizer of the objective function

$$L_W(\beta; \mathbf{w}) = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K R_i \frac{N_k}{n_k} \Delta_{ik} L_\beta(Y_i, \mathbf{W}_i) \quad (\text{C.1})$$

where $L_\beta(Y_i, \mathbf{W}_i)$ is the contribution to the negative log-likelihood for the i th subject.

In order to establish consistency and asymptotically normality of $\hat{\beta}$, we will assume that the full data negative log-likelihood function $L_\beta(Y_i, \mathbf{W}_i)$ is convex in β for every i . For simplicity, we further assume that the model depends on no additional unknown nuisance parameters and impose the following regularity conditions.

- C1. The parameter β takes values in an open convex set Ω . For each i , and given \mathbf{W}_i , the unique parameter $\beta_0 \in \Omega$ generates Y_i ; in addition, the distribution of \mathbf{W} does not depend on β_0 and $E_{\beta_0}[S_{\beta_0}(Y_i, \mathbf{W}_i)] = \mathbf{0}$ for $i = 1, \dots, N$.
- C2. For fixed values (Y, \mathbf{W}) , $S_\beta(Y, \mathbf{W})$ is twice continuously differentiable with respect to β for $\beta \in \mathcal{B} \subset \Omega$, where \mathcal{B} is an open set containing β_0 .
- C3. Let $\tilde{\mathbf{J}}_0 = -\partial S'_{\beta_0}(Y_g, \mathbf{W}_g)/\partial \beta_0$ for a generic observation (Y_g, \mathbf{W}_g) . The matrix $\mathbf{J}_0 = E_{\beta_0}[\tilde{\mathbf{J}}_0]$ is positive definite. In addition, there exist scalar functions $M_{ijk}(Y_g, \mathbf{W}_g)$ such that $E_{\beta_0}[|M_{ijk}(Y_g, \mathbf{W}_g)|] < \infty$ and

$$\left| \frac{\partial^2}{\partial \beta_j \partial \beta_k} [S_\beta(Y_g, \mathbf{W}_g)]_i \right| \leq M_{ijk}(Y_g, \mathbf{W}_g)$$

for all $\beta \in \mathcal{B}$, where β_j denotes the j th component of β and $[S_\beta(Y_g, \mathbf{W}_g)]_i$ the i th component of the vector $S_\beta(Y_g, \mathbf{W}_g)$.

- C4. $\frac{n_k}{N} \rightarrow \rho_k > 0$, $\frac{N_k}{N} \rightarrow q_k > 0$ and $\frac{n_k}{N} \rightarrow \rho_k q_k > 0$ as $N \rightarrow \infty$, for $k = 1, \dots, K$.

Theorem 4.1.1 Under conditions C1-C4, $\hat{\beta}$ is a strongly consistent estimator of β_0 .

Proof The IPW objective function (C.1) can be rewritten

$$L_W(\beta; \mathbf{w}) = \sum_{k=1}^K \frac{N_k^2}{N n_k} \left\{ \frac{1}{N_k} \sum_{i=1}^N L_{ik}(\beta) \right\}$$

where $L_{ik}(\beta) = R_i \Delta_{ik} L_{\beta}(Y_i, \mathbf{W}_i)$. Note that the outer sum over the K strata is fixed as the sample size N increases. Letting $\hat{\rho}_k = n_k/N_k$, $\hat{q}_k = N_k/N$, and $D_k(\beta) = \frac{1}{N_k} \sum_{i=1}^N L_{ik}(\beta)$, the objective function can also be written

$$\begin{aligned} L_W(\beta; \mathbf{w}) &= \sum_{k=1}^K \frac{\hat{q}_k}{\hat{\rho}_k} D_k(\beta) \\ &= \sum_{k=1}^K \frac{q_k}{\rho_k} D_k(\beta) + \sum_{k=1}^K \left(\frac{\hat{q}_k}{\hat{\rho}_k} - \frac{q_k}{\rho_k} \right) D_k(\beta) \\ &= \sum_{k=1}^K \frac{q_k}{\rho_k} D_k(\beta) + \sum_{k=1}^K \frac{\hat{q}_k}{\rho_k} \left(\frac{\rho_k}{\hat{\rho}_k} - 1 \right) D_k(\beta) + \sum_{k=1}^K \frac{q_k}{\rho_k} \left(\frac{\hat{q}_k}{q_k} - 1 \right) D_k(\beta). \end{aligned} \quad (\text{C.2})$$

Under conditions C1-C3, $D_k(\beta)$ converges to $L_{k,0}(\beta) := E[L_{ik}(\beta)]$ for all $\beta \in \mathcal{B}$ by the strong law of large numbers. Thus the first term in (C.2) converges almost surely to $L_0(\beta) \equiv \sum_{k=1}^K \frac{q_k}{\rho_k} L_{k,0}(\beta)$ for all $\beta \in \mathcal{B}$. Under condition C4, $\hat{\rho}_k \rightarrow \rho_k$ and $\hat{q}_k \rightarrow q_k$ so the remaining terms in (C.2) converge almost surely to zero.

Since $L_W(\beta; \mathbf{w})$ is convex for $\beta \in \mathcal{B}$, $L_0(\beta)$ is convex for $\beta \in \mathcal{B}$ and the convergence is uniform on all compact sets $A \subset \Omega$ (Andersen and Gill, 1982, Theorem II.1). By condition C3, $L_0(\beta)$ is strictly convex at β_0 and β_0 is a unique minimizer. It follows that the minimizer of $L_W(\beta; \mathbf{w})$ converges almost surely to β_0 (Andersen and Gill, 1982, Corollary II.2). \square

Theorem 4.1.2 Under conditions C1-C4, $N^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}, \Sigma)$ where

$$\Sigma = \mathbf{J}_0^{-1} \left[\mathbf{J}_0 + \sum_{k=1}^K \nu_k \frac{1 - \rho_k}{\rho_k} \text{Var}_{\mathcal{V}_k} \{S_{\beta_0}(Y_1, \mathbf{W}_1)\} \right] \mathbf{J}_0^{-1}, \quad (\text{C.3})$$

$\nu_k = P(\Delta_{ik} = 1)$ ($k = 1, \dots, K$); ρ_k and \mathbf{J}_0 are defined in conditions A1-A4; and, $Var_{\mathcal{V}_k}(S_{\beta_0}(Y_1, \mathbf{W}_1)) = Var(S_{\beta_0}(Y_1, \mathbf{W}_1)|\Delta_{1k} = 1)$ denotes the conditional variance of $S_{\beta_0}(Y_1, \mathbf{W}_1)$ given that a subject with data (Y_1, \mathbf{W}_1) falls in stratum \mathcal{V}_k .

It is worth pointing out here that the variance in (C.3) can be rewritten as

$$\mathbf{J}_0^{-1} + \mathbf{J}_0^{-1} \left[\sum_{k=1}^K \nu_k \frac{1 - \rho_k}{\rho_k} Var_{\mathcal{V}_k} \{S_{\beta_0}(Y_1, \mathbf{W}_1)\} \right] \mathbf{J}_0^{-1},$$

the first term representing the inverse Fisher information for estimating β_0 in the case where full data is observed on all subjects at stage one, and the second essentially representing the additional cost incurred as a result of using a two-stage sampling scheme.

Proof It can be shown that conditions C1-C4, in conjunction with the strong consistency of $\hat{\beta}$, are sufficient to satisfy the requirements imposed in Breslow and Wellner (2007, §2) to establish asymptotic normality of IPW estimators for two-stage designs with semiparametric models. For parametric models with smooth estimating equations, the regularity conditions of Breslow and Wellner (2007) may be simplified as described in van der Vaart and Wellner (1996); see, in particular, Theorem 3.3.1 and Example 3.3.8. Hence, we may assert, per equations (19) and (22) of Breslow and Wellner (2007), or Theorem 3.3.1 of van der Vaart and Wellner (1996), that

$$N^{1/2}(\hat{\beta} - \beta_0) = N^{-1/2} \sum_{i=1}^N R_i w_i \{ \mathbf{J}_0^{-1} S_{\beta_0}(Y_i, \mathbf{W}_i) \} + o_P(1) \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where Σ is defined in (C.3).□

Let

$$\mathbf{G} = \sum_{k=1}^K \nu_k \frac{1 - \rho_k}{\rho_k} Var_{\mathcal{V}_k} \{S_{\beta_0}(Y_1, \mathbf{W}_1)\}$$

so that the desired asymptotic variance can be written $\Sigma = \mathbf{J}_0^{-1}(\mathbf{J}_0 + \mathbf{G})\mathbf{J}_0^{-1}$. To estimate Σ , we must find suitable estimators of \mathbf{J}_0 and \mathbf{G} . The probabilities ν_k and ρ_k can be estimated by $\hat{\nu}_k = N_k/N$ and $\hat{\rho}_k = n_k/N_k$, respectively, and \mathbf{J}_0 can be estimated as

$$\begin{aligned}\hat{\mathbf{J}} &= -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \frac{R_i}{\hat{\rho}_k} \Delta_{ik} \frac{\partial S'_{\hat{\beta}}(Y_i, \mathbf{W}_i)}{\partial \hat{\beta}} \\ &= -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \frac{N_k}{n_k} R_i \Delta_{ik} \frac{\partial S'_{\hat{\beta}}(Y_i, \mathbf{W}_i)}{\partial \hat{\beta}}\end{aligned}$$

Furthermore, we can rewrite the variance term in \mathbf{G} as

$$\begin{aligned}\text{Var}_{\nu_k} \{S_{\beta_0}(Y, \mathbf{W})\} &= E_{\nu_k} \{S_{\beta_0}(Y, \mathbf{W})^{\otimes 2}\} - E_{\nu_k} \{S_{\beta_0}(Y, \mathbf{W})\}^{\otimes 2} \\ &= \frac{1}{\nu_k} E \{S_{\beta_0}(Y, \mathbf{W})^{\otimes 2} \Delta_{ik}\} \\ &\quad - \frac{1}{\nu_k^2} E \{S_{\beta_0}(Y, \mathbf{W}) \Delta_{ik}\}^{\otimes 2}.\end{aligned}\tag{C.4}$$

The first term in (C.4) can be written

$$\begin{aligned}E_{\nu_k} \{S_{\beta_0}(Y, \mathbf{W})^{\otimes 2}\} &= \frac{1}{\nu_k} E \{S_{\beta_0}(Y, \mathbf{W})^{\otimes 2} \Delta_{ik}\} \\ &= \frac{1}{\nu_k} E \left\{ \frac{R}{\rho_k} S_{\beta_0}(Y, \mathbf{W})^{\otimes 2} \Delta_{ik} \right\}\end{aligned}$$

which can be estimated as

$$\frac{1}{\hat{\nu}_k} \frac{1}{N} \sum_{i=1}^N \frac{R_i}{\hat{\rho}_k} S_{\hat{\beta}}(Y_i, \mathbf{W}_i)^{\otimes 2} \Delta_{ik}.$$

The second term in (C.4) can be written

$$E_{\nu_k} \{S_{\beta_0}(Y, \mathbf{W})\}^{\otimes 2} = \frac{1}{\nu_k^2} E \{S_{\beta_0}(Y, \mathbf{W}) \Delta_{ik}\}^{\otimes 2} = \frac{1}{\nu_k^2} E \left\{ \frac{R}{\rho_k} S_{\beta_0}(Y, \mathbf{W}) \Delta_{ik} \right\}^{\otimes 2}$$

which can be estimated as

$$\frac{1}{\hat{\nu}_k^2} \left\{ \frac{1}{N} \sum_{i=1}^N \frac{R_i}{\hat{\rho}_k} S_{\hat{\beta}}(Y_i, \mathbf{W}_i) \Delta_{ik} \right\}^{\otimes 2}.$$

It follows that \mathbf{G} can be estimated

$$\begin{aligned}
\hat{\mathbf{G}} &= \sum_{k=1}^K \hat{\nu}_k \frac{1 - \hat{\rho}_k}{\hat{\rho}_k} \left[\frac{1}{\hat{\nu}_k} \frac{1}{N} \sum_{i=1}^N \frac{R_i}{\hat{\rho}_k} S_{\hat{\beta}}(Y_i, \mathbf{W}_i)^{\otimes 2} \Delta_{ik} \right. \\
&\quad \left. - \frac{1}{\hat{\nu}_k^2} \left\{ \frac{1}{N} \sum_{i=1}^N \frac{R_i}{\hat{\rho}_k} S_{\hat{\beta}}(Y_i, \mathbf{W}_i) \Delta_{ik} \right\}^{\otimes 2} \right] \\
&= \frac{1}{N} \sum_{k=1}^K \frac{N_k(N_k - n_k)}{n_k^2} \sum_{i=1}^N R_i S_{\hat{\beta}}(Y_i, \mathbf{W}_i)^{\otimes 2} \Delta_{ik} \\
&\quad - \frac{1}{N} \sum_{k=1}^K \frac{N_k(N_k - n_k)}{n_k^3} \left\{ \sum_{i=1}^N R_i S_{\hat{\beta}}(Y_i, \mathbf{W}_i) \Delta_{ik} \right\}^{\otimes 2}
\end{aligned}$$

and $\hat{\Sigma} = \hat{\mathbf{J}}^{-1} + \hat{\mathbf{J}}^{-1} \hat{\mathbf{G}} \hat{\mathbf{J}}^{-1}$ estimates the asymptotic variance of $N^{1/2}(\hat{\beta} - \beta_0)$.

BIBLIOGRAPHY

- Odd O. Aalen, Ørnulf Borgan, and Høakon K. Gjessing. *Survival and Event History Analysis*. Springer-Verlag, New York, 2008.
- P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, 10(4):1100–1120, 1982. ISSN 0090-5364.
- Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1993.
- Miguel A. Arcones. Asymptotic theory for M -estimators over a convex kernel. *Econometric Theory*, 14(4):387–422, 1998. ISSN 0266-4666.
- Peter Barker and Robin Henderson. Small sample bias in the gamma frailty model for univariate survival. *Lifetime Data Anal.*, 11(2):265–284, 2005.
- Norman E. Breslow and Jon A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34(1):86–102, 2007.
- B. M. Brown and You-Gan Wang. Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, 92(1):149–158, 2005. ISSN 0006-3444.
- B. M. Brown and You-Gan Wang. Induced smoothing for rank regression with censored survival times. *Statistics in Medicine*, 26(4):828–836, 2006.
- D. R. Cox. Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972. ISSN 0035-9246.
- D.T. Danahy, D.T. Burwell, W.S. Aronow, and R. Prakash. Sustained hemodynamic and antianginal effect of high doses oral isosorbide dinitrate. *Circulation*, 55(2):381–387, 1977.

- Michael Elashoff and Louise Ryan. An EM algorithm for estimating equations. *J. Comput. Graph. Statist.*, 13(1):48–65, 2004.
- Liya Fu, You-Gan Wang, and Zhidong Bai. Rank regression for analysis of clustered data: a natural induced smoothing approach. *Computational Statistics and Data Analysis*, 54(4):1036–1050, 2010.
- Mendel Fyngenson and Ya’acov Ritov. Monotone estimating equations for censored data. *Ann. Statist.*, 22(2):732–746, 1994. ISSN 0090-5364.
- Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.
- Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, 14(3):262–280, 1996.
- Glenn Heller. Smoothed rank regression with censored data. *J. Amer. Statist. Assoc.*, 102(478):552–559, 2007.
- Philip Hougaard. *Analysis of multivariate survival data*. Statistics for Biology and Health. Springer-Verlag, New York, 2000.
- Yijian Huang. Calibration regression of censored lifetime medical cost. *J. Amer. Statist. Assoc.*, 97(457):318–327, 2002.
- Z. Jin, D. Y. Lin, L. J. Wei, and Z. Ying. Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353, 2003. ISSN 0006-3444.
- Z. Jin, D. Y. Lin, and Z. Ying. Rank regression analysis of multivariate failure time data based on marginal linear models. *Scand. J. Statist.*, 33(1):1–23, 2006a.

- Z. Jin, D. Y. Lin, and Z. Ying. On least squares regression with censored data. *Biometrika*, 93(1):147–162, 2006b.
- Lynn M. Johnson and Robert L. Strawderman. Induced smoothing for the semi-parametric accelerated failure time model: asymptotics and extensions to clustered data. *Biometrika*, 96(3):577–590, 2009.
- Lynn M. Johnson and Robert L. Strawderman. A Smoothing Expectation and Substitution algorithm for the semiparametric accelerated failure time model. *Statistics in Medicine*, 31(21):2335–2358, 2012.
- Mark E. Johnson. *Multivariate Statistical Simulation*. Wiley Series in Probability and Mathematical Statistics, New York, 1987.
- Michael P. Jones. A class of semiparametric regressions for the accelerated failure time model. *Biometrika*, 84(1):73–84, 1997.
- Bent Jørgensen. *Statistical properties of the generalized inverse Gaussian distribution*, volume 9 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1982.
- John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, New Jersey, 2002.
- Ram P. Kanwal. *Generalized Functions: Theory and Technique*. Birkhauser, Boston, 1998.
- D. Karlis. A general EM approach for maximum likelihood estimation in mixed poisson regression models. *Statistical Modelling: An International Journal*, 1(4): 305 – 318, 2001. ISSN 1471082X.
- John P. Klein. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48:795–806, 1992.

- Michael R. Kosorok, Bee Leng Lee, and Jason P. Fine. Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.*, 32(4):1448–1491, 2004.
- J.F. Lawless, J.D. Kalbfleisch, and C.J. Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B*, 61(2):413–438, 1999.
- Eric W. Lee, L. J. Wei, and Z. Ying. Linear regression analysis for highly stratified failure time data. *J. Amer. Statist. Assoc.*, 88(422):557–565, 1993. ISSN 0162-1459.
- J. S. Lin and L. J. Wei. Linear regression analysis for multivariate failure time observations. *J. Amer. Statist. Assoc.*, 87(420):1091–1097, 1992.
- Shuangge Ma and Michael R. Kosorok. Robust semiparametric M-estimation and the weighted bootstrap. *J. Multivariate Anal.*, 96(1):190–217, 2005.
- N. Mantel, N.R. Bohidar, and J.L. Ciminera. Mantel-haenszel analysis of litter-matched time-to-response data. *Cancer Research*, 37(11):3863–3868, 1977.
- Torben Martinussen and Thomas H. Scheike. *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, New York, 2006. ISBN 978-0387-20274-7; 0-387-20274-9.
- C.A. McGilchrist and C.W. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47(2):461–466, 1991.
- Gert G. Nielsen, Richard D. Gill, Per Kragh Andersen, and Thorkild I.A. Sørensen. A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, 19(1):25–43, 1992.

- Wei Pan. Using frailties in the accelerated failure time model. *Lifetime Data Anal.*, 7(1):55–64, 2001. ISSN 1380-7870.
- R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL <http://www.R-project.org>. ISBN 3-900051-07-0.
- J.B. Ramsey. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B*, 31(2):350–371, 1969.
- Takumi Saegusa and Jon A. Wellner. Weighted likelihood estimation under two-phase sampling. *Annals of Statistics*, 41(1):269–295, 2013.
- Alastair J. Scott and Chris J. Wild. Fitting regression models with response-biased samples. *The Canadian Journal of Statistics*, 39(3):519–536, 2011.
- Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons Canada, Ltd, 1980.
- X. Song, S. Ma, J. Huang, and X. Zhou. A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics*, 6(2):197–211, 2007.
- Robert L. Strawderman. The accelerated gap times model. *Biometrika*, 92(3): 647–666, 2005.
- Robert L. Strawderman. A regression model for dependent gap times. *The International Journal of Biostatistics*, 2(1):Art. 1, 34 pp. (electronic), 2006.
- Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer, New York, 2000.

- Anastasios A. Tsiatis. Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.*, 18(1):354–372, 1990. ISSN 0090-5364.
- Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.
- S. T. Wang, John P. Klein, and M. L. Moeschberger. Semi-parametric estimation of covariate effects using the positive stable frailty model. *Appl. Stochastic Models Data Anal.*, 11(2):121–133, 1995.
- You-Gan Wang and Min Zhu. Rank-based regression for analysis of repeated measures. *Biometrika*, 93(2):459–464, 2006.
- Mark A. Weaver and Haibo Zhou. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, 100(470):459–469, 2005.
- Chris J. Wild. Fitting prospective regression models to case-control data. *Biometrika*, 78(4):705–717, 1991.
- Linzhi Xu and Jiajia Zhang. An EM-like algorithm for the semiparametric accelerated failure time gamma frailty model. *Comput. Statist. Data Anal.*, 54(6):1467–1474, 2010.
- Z. Ying and L.J. Wei. The kaplan-meier estimate for dependent failure time observations. *Journal of Multivariate Analysis*, 50(1):17–29, 1994.
- Zhiliang Ying. A large sample study of rank estimation for censored regression data. *Ann. Statist.*, 21(1):76–99, 1993. ISSN 0090-5364.
- J. Zhang and Y. Peng. An alternative estimation method for the accelerated failure time frailty model. *Comput. Statist. Dat. Anal.*, 51(9):4413–4423, 2007.

Haibo Zhou, M.A. Weaver, J. Qin, M.P. Longnecker, and M.C. Wang. A semi-parametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2):413–421, 2002.

Haibo Zhou, Yuanshan Wu, Yanyan Liu, and Jianwen Cai. Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. *Biostatistics*, 12(2):521–534, 2011.