



A Two-Stage Estimator of the Dependence Parameter for the Clayton-Oakes Model

DAVID V. GLIDDEN

david@biostat.ucsf.edu

Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143

Received December 1, 1998; Revised September 22, 1999; Accepted September 23, 1999

Abstract. This paper describes the properties of a two-stage estimator of the dependence parameter in the Clayton-Oakes multivariate failure time model. The parameter is estimated from a likelihood function in which the marginal hazard functions are replaced by estimates. The method extends the approach of Shih and Louis (1995) and Genest, Ghoudi and Rivest (1995) to allow the marginal hazard for failure times to follow a stratified Cox (1972) model. The method is computationally simple and under mild regularity conditions produces a consistent, asymptotically normal estimator.

Keywords: clustered data, frailty model, multivariate failure time data

1. Introduction

Multivariate failure time (MVFT) data arise when a sample consists of clusters and each cluster contains several, possibly dependent failures. The analysis of MVFT data must reflect the possible non-independence of failures within clusters. Lin (1994) discusses biomedical examples of MVFT data. This paper is concerned with the joint parameter estimation from MVFT data when clustering follows a Clayton-Oakes model (Clayton, 1978; Oakes, 1982, 1986) and covariate effects follow a marginal proportional hazards model.

Let (T_1, \dots, T_K) be a vector of clustered failure times and let $\lambda_k(\cdot)$ be the marginal hazard function of T_k . Under the Clayton-Oakes model, the joint survivor function of (T_1, \dots, T_K) is

$$\text{pr}(T_1 > t_1, \dots, T_K > t_K; \theta_0) = \left[\sum_{k=1}^K \exp\{\theta_0 \int_0^{t_k} \lambda_k(u) du\} - K + 1 \right]^{-\theta_0^{-1}}, \quad (1)$$

which is parameterized by the marginal hazard functions $(\lambda_k(\cdot), k = 1, \dots, K)$ and a dependence parameter θ_0 .

The Clayton-Oakes model can also be derived as a random effects model with an alternative but equivalent parameterization. Frailty models (Vaupel et al., 1979) postulate that conditional on ξ , the T_k 's are independent with hazard functions:

$$\lim_{h \downarrow 0} h^{-1} \text{pr}(t \leq T_k < t + h \mid T_k \geq t, \xi) = \xi \lambda_k^c(t), \quad k = 1, \dots, K, \quad (2)$$

where $\lambda_k^c(\cdot)$ ($k = 1, \dots, K$) are termed basic hazard functions. If the random effects ξ follow a gamma distribution with mean one and variance θ_0 , then (T_1, \dots, T_K) follow the

model (1) with the marginal and basic hazard functions related by

$$\lambda_k^c(t) = \exp\left\{\theta_0 \int_0^t \lambda_k(s) ds\right\} \lambda_k(t). \quad (3)$$

Suppose that the risk of failure also depends on a set of covariates (Z_1, \dots, Z_k) . Within a Clayton-Oakes model, it is possible to adopt two different model strategies—a marginal proportional hazards frailty model or a proportional hazards frailty model. The choices are based on proportional hazards models for $\lambda_k(\cdot)$, $(k = 1, \dots, K)$ in (1) or $\lambda_k^c(\cdot)$, $(k = 1, \dots, K)$ in (2), respectively. This paper considers the former parameterization—marginal hazards for $T_k | Z_k$ which follow

$$\lambda_0(t) e^{\beta_0^T Z_k} \quad \text{or} \quad \lambda_{0k}(t) e^{\beta_0^T Z_k},$$

corresponding to the unstratified and stratified marginal Cox (1972) model. This paper will consider a model and notation developed by Spiekerman and Lin (1998) which includes the common and distinct hazards as special cases but also allows for intermediate specifications.

Suppose there are n independent failure time vectors \mathbf{T}_i with components T_{ikl} . The subscript i refers to cluster membership. Let failures also be grouped into $k = 1, \dots, K$ strata, each of which have their own reference hazard $\Lambda_{0k}(\cdot)$. For example, i could refer to family membership and k could refer to sex. This would allow for clustering of failures in families and allow for separate reference hazards by sex. The index l then refers to subjects of a given sex within a family.

Let T_{ikl} and C_{ikl} denote the failure and censoring times for the l th realization of the k th failure type within the i th cluster. The p -dimensional vector Z_{ikl} denotes a set of covariates. The marginal (cumulative) hazard for this model follows

$$\Lambda_{ikl}(t | Z_{ikl}) = \Lambda_{0k}(t) e^{\beta_0^T Z_{ikl}},$$

where $\Lambda_{0k}(\cdot)$ ($k = 1, \dots, K$) are continuous and bounded on $[0, \tau]$.

Let \mathbf{T}_i be the vector $(T_{ikl}, k = 1, \dots, K; l = 1, \dots, L)$ with \mathbf{C}_i and \mathbf{Z}_i defined similarly. Assume that \mathbf{T}_i , \mathbf{C}_i , and \mathbf{Z}_i are independent and $(\mathbf{T}_i, \mathbf{C}_i)$ ($i = 1, \dots, n$) are independent and identically distributed with the components of \mathbf{T}_i and \mathbf{C}_i independent given \mathbf{Z}_i . Assume $|\mathbf{Z}_i|$ are bounded *a.s.*

Denote $X_{ikl} = T_{ikl} \wedge C_{ikl}$ and $\Delta_{ikl} = I(T_{ikl} \leq C_{ikl})$, where $I(\cdot)$ is the indicator function, and $a \wedge b = \min(a, b)$. It is convenient to regard K and L as fixed constants; however, clusters may have different sizes by setting C_{ikl} to zero wherever T_{ikl} is missing. The missingness probability for T_{ikl} may depend on covariates but not on T_{ihm} ($h \neq k, m \neq l$). In counting process notation, the data are represented by $Y_{ikl}(t) := I(X_{ikl} \geq t)$ and $N_{ikl}(t) := \Delta_{ikl} I(X_{ikl} \leq t)$. The maximum follow-up time is denoted by τ . Finally, we assume that data follow a Clayton-Oakes model with marginal proportional hazards and conditions (a)-(c) of Spiekerman and Lin (1998).

This paper will discuss estimation of the marginal parameters and θ_0 . Estimating equation and likelihood approaches have been attempted for joint estimation. Prentice and Hsu (1997) developed a general estimating equation framework. Separate estimating equations were developed for the marginal parameters and the dependence parameters. Glidden and Self

(1999) attempted to base estimation of the parameters on the construction and maximization of a semi-parametric likelihood. Their approach, in the spirit of the work of Nielsen et al. (1992), is computationally demanding and asymptotic theory for the estimators is largely unavailable.

This paper applies the “two-stage” approach to the problem. This approach first estimates the marginal parameters using working-independence estimators (Wei et al., 1989). In the second stage, these estimators are substituted into a likelihood for the dependence parameters, yielding a pseudo likelihood (in the terminology of Gong and Samaniego, 1981). Previous work on these approach (Shih and Louis 1995; Genest et al., 1995) has been limited to paired failure times without covariates. We present a treatment which considers multiple failures per cluster and covariates modeled by a marginal Cox model. The asymptotic theory for our two-stage estimators builds on the rigorous asymptotic theory developed by Spiekerman and Lin (1998) for working independence estimates for the marginal parameters β_0 and $\Lambda_k(\cdot) = \int \lambda_k$.

2. Parameter Estimation

Nielsen et al. (1992) present an approach to estimation in the Clayton-Oakes model using nonparametric maximum likelihood and the parameterization (2). Conditional on ξ , the likelihood for (2) has the form of the familiar counting process (partial) likelihood. Averaging over the distribution of ξ yields a (partial) likelihood for the Clayton-Oakes model. We assume the censoring mechanism is not indexed by parameters of the failure distribution. Thus, the partial likelihood is proportional to the full likelihood.

If the basic hazard functions $\lambda_{ikl}^c(\cdot)$ are known, estimation of θ may be based on Nielsen’s log-likelihood

$$\begin{aligned}
 l_n(\theta) = & n^{-1} \sum_{i=1}^n \int_0^\tau \log \{1 + \theta N_{i..}(u-)\} dN_{i..}(u) \\
 & + \sum_{k=1}^K \sum_{l=1}^L \int_0^\tau \log(Y_{ikl}(u)\lambda_{ikl}^c(u)) dN_{ikl}(u) \\
 & - \{\theta^{-1} + N_{i..}(\tau)\} \log \left\{ 1 + \theta \sum_{k=1}^K \sum_{l=1}^L \int_0^\tau Y_{ikl}(u) d\Lambda_{ikl}^c(u) \right\}
 \end{aligned}$$

where $\lambda_{ikl}^c(t)$ is the basic hazard function for the l th replicate of the k th failure type in the i th cluster and $\Lambda_{ikl}^c(t) = \int_0^t \lambda_{ikl}^c(s) ds$. Using equation (3) and applying integration by parts, the log-likelihood for θ has the form:

$$\begin{aligned}
 & n^{-1} \sum_{i=1}^n \int_0^\tau \log \{1 + \theta N_{i..}(u-)\} dN_{i..}(u) \\
 & + \sum_{k=1}^K \sum_{l=1}^L \theta N_{ikl}(\tau) H_{ikl} - \{\theta^{-1} + N_{i..}(\tau)\} \log\{R_i(\theta)\}
 \end{aligned} \tag{4}$$

where $H_{ikl} = \int_0^\tau Y_{ikl}(u) e^{\beta_0^T Z_{ikl}} d\Lambda_{0k}(u)$, $R_i(\theta) = \sum_{k=1}^K \sum_{l=1}^L e^{\theta H_{ikl}} - KL + 1$, and $N_{i..}(t)$ is the total number of failures in the i th cluster by time t . Estimation of θ_0 proceeds in two stages. First, “working independence” estimators of β_0 and $\Lambda_{0k}(\cdot)$ ($k = 1, \dots, K$) are calculated. The estimator $\hat{\beta}$ of β_0 solves $U(\hat{\beta}) = 0$, where

$$U(\beta) = \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^L \int_0^\tau \{Z_{ikl} - E_k(\beta, u)\} dN_{ikl}(u).$$

The estimator of $\hat{\Lambda}_k(\cdot)$ has the form

$$\hat{\Lambda}_k(t) = \int_0^t \frac{dN_{.k.}(u)}{nS_k^{(0)}(\hat{\beta}, u)} \quad (k = 1, \dots, K)$$

with $N_{.k.}(t) = \sum_{i=1}^n \sum_{l=1}^L N_{ikl}(t)$ and

$$E_k(\beta, t) = \frac{S_k^{(1)}(\beta, t)}{S_k^{(0)}(\beta, t)}, \quad S_k^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^n \sum_{l=1}^L Y_{ikl}(t) e^{\beta^T Z_{ikl}} Z_{ikl}^{\otimes r}, \quad r = 0, 1, 2,$$

where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$. Second, $\hat{\beta}$ and $\hat{\Lambda}_k(\cdot)$, replace β_0 and $\Lambda_{0k}(t)$, in the likelihood (4). This pseudo (i.e. second stage) log-likelihood is the basis for estimation of θ_0 and has the form,

$$\begin{aligned} \hat{l}_n(\theta) = & n^{-1} \sum_{i=1}^n \int_0^\tau \log(1 + \theta N_{i..}(u-)) dN_{i..}(u) \\ & + \sum_{k=1}^K \sum_{l=1}^L \theta N_{ikl}(\tau) \hat{H}_{ikl} - (\theta^{-1} + N_{i..}(\tau)) \log(\hat{R}_i(\theta)) \end{aligned} \quad (5)$$

where

$$\hat{H}_{ikl} = \int_0^\tau Y_{ikl}(s) e^{\hat{\beta}^T Z_{ikl}} d\hat{\Lambda}_k(s) \quad \text{and} \quad \hat{R}_i(\theta) = \sum_{k=1}^K \sum_{l=1}^L e^{\theta \hat{H}_{ikl}} - KL + 1.$$

When θ equals 0, the log-likelihoods (4) and (5) are defined by their limit as $\theta \rightarrow 0$. The pseudo log-likelihood (4) has the value $-n^{-1} \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^L \hat{H}_{ikl}$ when $\theta = 0$.

The pseudo score function $\hat{S}_n(\theta)$ for $\hat{l}_n(\theta)$ in (5) is

$$\begin{aligned} \hat{S}_n(\theta) = & n^{-1} \sum_{i=1}^n \int_0^\tau \frac{N_{i..}(u-)}{1 + \theta N_{i..}(u-)} dN_{i..}(u) + \theta^{-2} \log\{\hat{R}_i(\theta)\} \\ & - \{\theta^{-1} + N_{i..}(\tau)\} \hat{R}_i^{-1}(\theta) \sum_{k=1}^K \sum_{l=1}^L \hat{H}_{ikl} e^{\theta \hat{H}_{ikl}} + \sum_{k=1}^K \sum_{l=1}^L N_{ikl}(\tau) \hat{H}_{ikl}. \end{aligned} \quad (6)$$

The value of the pseudo score function at 0 is defined by its limit at 0,

$$\begin{aligned} \hat{S}_n(0) = & n^{-1} \sum_{i=1}^n \int_0^\tau N_{i..}(u-) dN_{i..}(u) - N_{i..}(\tau) \sum_{k=1}^K \sum_{l=1}^L \hat{H}_{ikl} \\ & + \frac{1}{2} \left\{ \left(\sum_{k=1}^K \sum_{l=1}^L \hat{H}_{ikl} \right)^2 - \sum_{k=1}^K \sum_{l=1}^L \hat{H}_{ikl}^2 \right\} + \sum_{k=1}^K \sum_{l=1}^L N_{ikl}(\tau) \hat{H}_{ikl}. \end{aligned} \quad (7)$$

Recent work by Spiekerman and Lin (1998) has developed rigorous and unified asymptotic theory for $\hat{\beta}$ and $\hat{\Lambda}_k(\cdot)$, ($k = 1, \dots, K$). This permits an extended treatment of the two-stage estimators considered by previous authors.

The pseudo log-likelihood is continuous in θ and is defined at zero and for negative values close to zero. The estimator of θ_0 , $\hat{\theta}$, is the root of the pseudo score equation $\hat{S}_n = 0$ which can be solved by Newton-Raphson. Newton-Raphson can sometimes diverge when θ_0 is small. For small values, the bisection or golden section methods will locate the root of (6) more consistently. For larger values, Newton-Raphson solves the pseudo score equation rapidly.

3. Asymptotic Theory

The proofs of the consistency and asymptotic normality for the estimator rely on the results of Spiekerman and Lin (1998). Their work proves the consistency and weak convergence of $\hat{\beta}$ and $\hat{\Lambda}_k(\cdot)$ ($k = 1, \dots, K$) for general dependence structures. These results ensure that the pseudo log-likelihood (5) and pseudo score (6) inherit key properties which yield the asymptotic properties of $\hat{\theta}$. Our proofs assume the conditions given by Spiekerman and Lin for the asymptotic properties for the marginal parameters. Standard likelihood conditions are assumed for the likelihood (4).

3.1. Consistency

The consistency proof is a modification of the classic proof in Lehmann (1983). The consistency of the marginal parameters implies that the pseudo log-likelihood $\hat{l}(\cdot)$ converges uniformly to the (log) likelihood $l(\cdot)$ for all θ in the parameter space Θ . Thus as n becomes large, the true value θ_0 tends to maximize the pseudo likelihood with probability approaching 1.

THEOREM 1 (CONSISTENCY OF $\hat{\theta}$) *With probability approaching one, there exists a root of the pseudo score function given in (6) and (7), denoted $\hat{\theta}$, which converges in probability to θ_0 , such that for any fixed $\epsilon > 0$, $P(|\hat{\theta} - \theta_0| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$.*

Proof: The log-likelihood $l_n(\theta)$ can trivially be written as the sum of i.i.d. random variables $n^{-1} \sum \phi_i(\theta)$. The proof for the maximum likelihood estimator based on (4) relies

on showing that for $\theta \neq \theta_0$, $\lim_{n \rightarrow \infty} l_n(\theta) - l_n(\theta_0) < 0$ *a.s.* which follows from the law of large numbers and Jensen's inequality. The consistency of $\hat{\Lambda}_k(\cdot)$ ($k = 1, \dots, K$) and $\hat{\beta}$ gives

$$\sup_{\theta \in \Theta} |\hat{l}_n(\theta) - l_n(\theta)| \rightarrow 0 \quad (8)$$

as $n \rightarrow \infty$. From (8) and the triangle inequality it follows that for $\theta \neq \theta_0$ $\lim_{n \rightarrow \infty} \hat{l}_n(\theta) - \hat{l}_n(\theta_0) < 0$. This means $P\{\hat{l}_n(\theta_0) > \hat{l}_n(\theta)\} \rightarrow 1$ as $n \rightarrow \infty$. The pseudo log-likelihood (5) is a continuous function of θ on Θ and by theorem 2.2 in Lehmann (1983) with probability approaching 1, the estimation equation (6) has a weakly consistent root $\hat{\theta}$. ■

3.2. Weak Convergence

The proof of the weak convergence of $\sqrt{n}(\hat{\theta} - \theta_0)$ also parallels classic weak convergence proofs. The proof relies on showing the asymptotic normal distribution of the pseudo score at θ_0 . This follows because the pseudo score is asymptotically equivalent to the sum of i.i.d. random variables. The asymptotic variance of the pseudo score is inflated by terms which reflect the fact that the marginal parameters are estimated in (6). The key result, given in Lemma 1, is showing the asymptotic distribution of the pseudo score evaluated at θ_0 .

LEMMA 1 (WEAK CONVERGENCE OF THE PSEUDO SCORE) *The pseudo score, \hat{S}_n , given in (6) and (7) evaluated at θ_0 converges weakly to a Gaussian limit, i.e., $\sqrt{n}\hat{S}_n(\theta_0) \Rightarrow \mathcal{G}$ where \mathcal{G} is a mean zero normally distributed random variable.*

Proof: It can be shown that

$$\sqrt{n} \left| \hat{S}_n(\theta_0) - S_n(\theta_0) - \sum_{k=1}^K \int_0^\tau \tilde{\pi}_k(s) d(\hat{\Lambda}_k - \Lambda_{0k})(s) - \tilde{F}^T(\hat{\beta} - \beta_0) \right| \rightarrow 0$$

almost surely where

$$\begin{aligned} \tilde{\pi}_k(t) := n^{-1} \sum_{i=1}^n \sum_{l=1}^L e^{\beta^T Z_{ikl}} Y_{ikl}(t) & \left[\theta_0^{-1} R_i^{-1}(\theta_0) e^{\theta_0 H_{ikl}} \right. \\ & - \{\theta_0^{-1} + N_{i..}(\tau)\} R_i^{-1}(\theta_0) \{1 + \theta_0 H_{ikl}\} e^{\theta_0 H_{ikl}} \\ & + \{1 + \theta_0 N_{i..}(\tau)\} R_i^{-2}(\theta_0) \left\{ \sum_{k=1}^K \sum_{l=1}^L H_{ikl} e^{\theta_0 H_{ikl}} \right\} e^{\theta_0 H_{ikl}} \\ & \left. + N_{ikl}(\tau) \right] \quad (9) \end{aligned}$$

and

$$\begin{aligned} \tilde{F} := n^{-1} \sum_{i=1}^n \sum_{k=1}^K \sum_{l=1}^L Z_{ikl} H_{ikl} e^{\beta^T Z_{ikl}} Y_{ikl}(t) & \left[\theta_0^{-1} R_i^{-1}(\theta_0) e^{\theta_0 H_{ikl}} \right. \\ & - \{\theta_0^{-1} + N_{i..}(\tau)\} R_i^{-1}(\theta_0) \{1 + \theta_0 H_{ikl}\} e^{\theta_0 H_{ikl}} \\ & + \{1 + \theta_0 N_{i..}(\tau)\} R_i^{-2}(\theta_0) \left\{ \sum_{k=1}^K \sum_{l=1}^L H_{ikl} e^{\theta_0 H_{ikl}} \right\} e^{\theta_0 H_{ikl}} \\ & \left. + N_{ikl}(\tau) \right] \end{aligned} \quad (10)$$

for $\theta_0 \neq 0$. For $\theta_0 = 0$, the formulas are given in the Appendix. It can be shown that $\sup_{t \in [0, \tau]} |\tilde{\pi}_k(t) - \pi_k(t)| \xrightarrow{P} 0$ and $\tilde{F} \xrightarrow{P} F$ for $(k = 1, \dots, K)$, where $\pi_k(t)$ is the pointwise limit of $\tilde{\pi}_k(t)$ and $F := E(\tilde{F})$. Applying the Skorohod-Dudley-Wichura almost sure representation (Shorack and Wellner (1986), p. 47), $\sqrt{n} \hat{S}_n(\theta_0)$ shares the same asymptotic distribution as

$$\sqrt{n} \left(S_n(\theta_0) + \sum_{k=1}^K \int_0^\tau \pi_k(s) d(\hat{\Lambda}_k - \Lambda_{0k})(s) + F^T (\hat{\beta} - \beta_0) \right).$$

By the results of Spiekerman and Lin (1998) the above is asymptotically equivalent to the sum of i.i.d. mean zero random variables with finite variance $n^{-1/2} \sum_{i=1}^n \Phi_i$, where Φ_1, \dots, Φ_n are given in the Appendix. By the central limit theorem $\sqrt{n} \hat{S}_n(\theta_0) \Rightarrow \mathcal{G}$ where \mathcal{G} is a mean 0 Gaussian random variable with variance $\sigma_\Phi^2 = E(\Phi_1^2)$. The variance of \mathcal{G} is estimated by $\hat{\sigma}_\Phi^2 = n^{-1} \sum_{i=1}^n \hat{\Phi}_i^2$. In the Appendix it is shown that if $\theta_0 = 0$, $\pi_k(\cdot) = 0$ ($k = 1, \dots, K$) and $F = 0$. In that case, the asymptotic variance of $\hat{S}(\theta_0)$ and $S(\theta_0)$ are equal. ■

THEOREM 2 (ASYMPTOTIC DISTRIBUTION OF PSEUDO LIKELIHOOD ESTIMATOR):

$\sqrt{n}(\hat{\theta} - \theta_0)$ converges weakly to a mean zero Gaussian random variable with variance σ^2 .

Proof: It can be shown that $\sqrt{n} |\hat{S}_n(\theta_0) - (\hat{\theta} - \theta_0) I(\theta_0)| \rightarrow 0$ as $n \rightarrow \infty$ where $I(\theta_0) = \lim_{n \rightarrow \infty} I_n(\theta_0)$ and $I_n(\theta)$ is the minus second derivative of (4). By the above, and Lemma 1, $I(\theta_0) \sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically equivalent to $n^{-1/2} \sum_{i=1}^n \Phi_i$. Since $I(\theta_0) > 0$ a.s., $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically equivalent to $I^{-1}(\theta_0) n^{-1/2} \sum_{i=1}^n \Phi_i$. The central limit theorem proves the theorem with $\sigma^2 = I^{-1}(\theta_0) \sigma_\Phi^2 I^{-1}(\theta_0)$. ■

The next theorem discusses the consistency of the estimator of the asymptotic variance

$$\hat{\sigma}^2 := \hat{I}^{-1}(\hat{\theta}) \left\{ n^{-1} \sum_{i=1}^n \hat{\Phi}_i^2 \right\} \hat{I}^{-1}(\hat{\theta}),$$

where $\hat{I}(\cdot)$ is the derivative of the pseudo score. The proof of the consistency of $\hat{\sigma}^2$ follows from the consistency of the parameter estimators.

THEOREM 3 (CONSISTENT ESTIMATOR OF THE ASYMPTOTIC VARIANCE) *The estimator of the asymptotic variance $\hat{\sigma}^2$ satisfies $|\hat{\sigma}^2 - \sigma^2| \rightarrow 0$ almost surely as $n \rightarrow \infty$.*

Proof: The consistency follows in two steps: first, we show that $\hat{\sigma}_\Phi^2$ is consistent and that $\hat{I}_n(\hat{\theta})$ converges to $I(\theta_0)$. The strong law of large numbers gives $n^{-1} \sum_{i=1}^n \Phi_i^2 \rightarrow \sigma_\Phi^2$. It can be shown that $n^{-1} \sum_{i=1}^n \hat{\Phi}_i^2 - \Phi_i^2$ converges to zero, almost surely. It is sufficient to show that $|\hat{\Phi}_1 + \Phi_1|$ is bounded and that $n^{-1} \sum_{i=1}^n |\hat{\Phi}_i - \Phi_i|$ converges to zero almost surely. This can be shown term by term. Second, since $\hat{I}_n(\cdot)$ is continuous on Θ and $\hat{I}_n(\cdot)$ converges to $I_n(\cdot)$ uniformly in Θ , then $I_n(\hat{\theta})$ converges to $I_n(\theta_0)$ which converges to $I(\theta_0)$. ■

THEOREM 4 (PSEUDO LIKELIHOOD TEST STATISTIC) *The difference $-2\{\hat{l}_n(\theta_0) - \hat{l}_n(\hat{\theta})\}$ converges weakly to a constant times a chi-square on one degree of freedom.*

Proof: Let $W = -2\{\hat{l}_n(\theta_0) - \hat{l}_n(\hat{\theta})\}$. Following the expansion of Cox and Hinkley (1974) pp. 317–318 and the results of Theorems 1 and 2, we obtain $W = n(\hat{\theta} - \theta_0)^2 I(\theta_0) + o_p(1)$. This implies W is asymptotically equivalent to $\sigma^2 I(\theta_0)$ times a χ_1^2 . $I(\theta_0)$ is consistently estimated by $\hat{I}(\hat{\theta})$ and σ^2 is consistently estimated by $\hat{\sigma}^2$. Thus by Slutsky's theorem, W is asymptotically equivalent to $\hat{\sigma}^2 \hat{I}(\hat{\theta})$ times a χ_1^2 . ■

4. Simulation Studies

Simulation studies were carried out to assess the behavior of the estimator of θ_0 in moderate sample sizes and to compare the properties of Wald and pseudo likelihood-ratio (PLR) based confidence intervals. The first simulation considered $K = 1$ and $L = 2$ with varying numbers of clusters, namely $n = 50, 100$ and 200 . The joint survivor function for the failure times (T_1, T_2) took the form:

$$S(t_1, t_2; \theta_0) = (e^{-\theta_0 t_1} + e^{-\theta_0 R t_2} - 1)^{-\theta_0^{-1}}$$

which is a Clayton-Oakes model with a unit exponential margin distribution for T_1 and a hazard ratio, $R = e^{\beta_0} = 0.5$ between T_2 and T_1 . The second simulation considered $K = 1$ and $L = 5$ for $n = 50, 100$ and 200 . The joint survivor function for the failure times (T_1, \dots, T_5) had the form:

$$S(t_1, \dots, t_5; \theta_0) = \left(\sum_{l=1}^2 e^{-\theta_0 t_l} + \sum_{l=3}^5 e^{-R \theta_0 t_l} - 4 \right)^{-\theta_0^{-1}}.$$

For both simulations the dependence parameter θ_0 took values of 0.1, 0.7, and 2.0. Three

Table 1. Small Sample Behavior of $\hat{\theta}$ and associated confidence intervals for varying dependence, n and censoring patterns, $K = 1, L = 2$.

θ_0	Cens.	Mean $\hat{\theta}$	SD($\hat{\theta}$)	Mean \hat{SD}	Wald Cov.	PLR Cov.
0.10						
$n = 50$	None	0.12	0.19	0.17	0.88	0.93
	$t = 2$	0.13	0.21	0.19	0.90	0.94
	Unif.	0.13	0.26	0.25	0.92	0.95
	$n = 100$					
	None	0.11	0.13	0.12	0.92	0.94
	$t = 2$	0.11	0.13	0.13	0.93	0.95
	Unif.	0.11	0.17	0.16	0.93	0.95
	$n = 200$					
	None	0.10	0.09	0.08	0.93	0.94
$t = 2$	0.10	0.09	0.09	0.94	0.94	
Unif.	0.11	0.12	0.11	0.94	0.95	
0.70						
$n = 50$	None	0.72	0.30	0.29	0.92	0.93
	$t = 2$	0.74	0.33	0.31	0.93	0.94
	Unif.	0.77	0.42	0.38	0.94	0.95
	$n = 100$					
	None	0.70	0.21	0.20	0.93	0.93
	$t = 2$	0.72	0.23	0.22	0.94	0.94
	Unif.	0.72	0.28	0.27	0.94	0.94
	$n = 200$					
	None	0.70	0.15	0.15	0.94	0.94
$t = 2$	0.71	0.16	0.16	0.95	0.95	
Unif.	0.71	0.19	0.19	0.95	0.95	
2.00						
$n = 50$	None	1.93	0.52	0.52	0.91	0.92
	$t = 2$	2.05	0.58	0.58	0.94	0.94
	Unif.	2.03	0.70	0.69	0.92	0.92
	$n = 100$					
	None	1.95	0.38	0.37	0.92	0.92
	$t = 2$	2.03	0.42	0.41	0.94	0.94
	Unif.	2.02	0.50	0.48	0.93	0.94
	$n = 200$					
	None	1.96	0.27	0.26	0.93	0.94
$t = 2$	2.01	0.29	0.29	0.94	0.94	
Unif.	2.01	0.34	0.34	0.95	0.95	

censoring patterns were explored: 1) uniform $[0, 2.3]$, yielding approximately 30% censoring; 2) censoring at $t = 2$, yielding approximately 13% censoring; and 3) no censoring. The mean of the estimates $\hat{\theta}$, their empirical standard deviation, the mean of their estimated standard error $n^{-1/2}\hat{\sigma}$, and the coverage of the 0.95 confidence intervals are given in Table 1 (for $L = 2$) and Table 2 (for $L = 5$). The results are based on 2,000 simulated datasets for each θ_0, n combination.

Table 2. Small Sample Behavior of $\hat{\theta}$ and associated confidence intervals for varying dependence, n and censoring patterns, $K = 1, L = 5$.

θ_0	Cens.	Mean $\hat{\theta}$	SD($\hat{\theta}$)	Mean \hat{SD}	Wald Cov.	PLR Cov.		
0.10	$n = 50$	None	0.10	0.07	0.07	0.91	0.94	
		$t = 2$	0.10	0.07	0.07	0.91	0.95	
		Unif.	0.10	0.08	0.08	0.92	0.94	
	$n = 100$	None	0.10	0.05	0.05	0.92	0.94	
		$t = 2$	0.10	0.05	0.05	0.92	0.94	
		Unif.	0.10	0.06	0.06	0.93	0.94	
	$n = 200$	None	0.10	0.03	0.03	0.93	0.95	
		$t = 2$	0.10	0.04	0.03	0.93	0.94	
		Unif.	0.10	0.04	0.04	0.94	0.95	
	0.70	$n = 50$	None	0.68	0.18	0.16	0.87	0.88
			$t = 2$	0.69	0.19	0.17	0.89	0.89
			Unif.	0.69	0.20	0.19	0.91	0.92
		$n=100$	None	0.70	0.13	0.12	0.90	0.90
			$t = 2$	0.70	0.13	0.12	0.91	0.91
			Unif.	0.70	0.14	0.14	0.92	0.93
$n = 200$		None	0.70	0.09	0.09	0.92	0.92	
		$t = 2$	0.70	0.09	0.09	0.93	0.93	
		Unif.	0.70	0.10	0.10	0.93	0.93	
2.00		$n = 50$	None	1.90	0.41	0.36	0.86	0.87
			$t = 2$	1.94	0.41	0.38	0.89	0.89
			Unif.	1.93	0.44	0.41	0.88	0.89
		$n = 100$	None	1.94	0.29	0.27	0.90	0.91
			$t = 2$	1.97	0.29	0.27	0.92	0.92
			Unif.	1.96	0.32	0.30	0.91	0.91
	$n = 200$	None	1.97	0.20	0.20	0.92	0.92	
		$t = 2$	1.98	0.20	0.20	0.93	0.93	
		Unif.	1.98	0.22	0.21	0.93	0.93	

The method gave approximately unbiased estimates of $\hat{\theta}$. The methods of Nielsen et al. (1992) and Prentice and Hsu (1997) showed a more pronounced negative bias in estimators of θ particularly in the absence of censoring. The coverage of the PLR intervals was as good or better than Wald intervals in every scenario and was appreciably better when $\theta_0 = 0.10$ and $n = 50$.

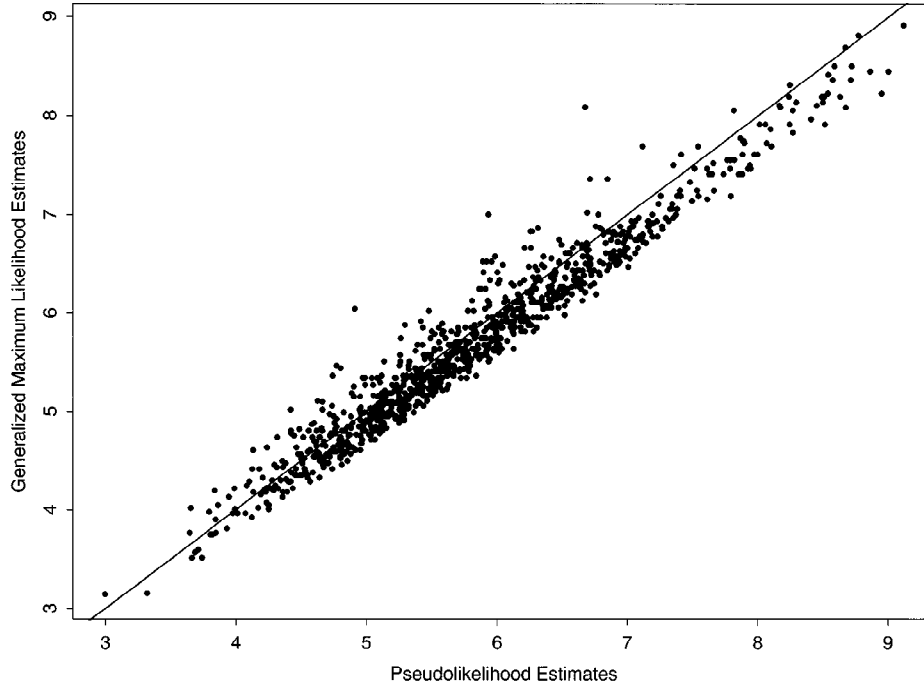


Figure 1. Scatterplot of 1,000 GMLE values versus 1,000 PLE values for $\theta_0 = 6.0$ and $n = 100$. Line indicates values $\tilde{\theta} - \hat{\theta} = 0$.

To assess the efficiency loss due to estimating θ from a pseudo likelihood rather than a full likelihood, a third simulation study was undertaken. The $M = 5,000$ datasets each with 100 clusters and $K = 1, L = 2$ were simulated from a Clayton-Oakes model with unit exponential margins. Values were censored at $t = 2$. The values of θ_0 included $\theta_0 = 0.2, 1.0, 6.0$. Two-Stage (pseudo likelihood) estimates (PLE) $\hat{\theta}$ are compared with generalized maximum likelihood estimates (GMLE) $\tilde{\theta}$ calculated using the method described by Nielsen et al. (1992). Table 3 gives sample mean and standard deviations for $\hat{\theta}$ and $\tilde{\theta}$ as well as relative efficiency $SD(\tilde{\theta})/SD(\hat{\theta})$ and ‘closeness’, i.e.,

$$\hat{\text{pr}}(|\tilde{\theta} - \theta_0| \leq |\hat{\theta} - \theta_0|) = M^{-1} \sum_{m=1}^M I(|\tilde{\theta}_m - \theta_0| \leq |\hat{\theta}_m - \theta_0|).$$

The results show that the two methods give similar answers. Figure 1 plots the GMLE $\tilde{\theta}$ versus the PLE $\hat{\theta}$ for a randomly chosen subset (1,000 datasets) when $\theta_0 = 6.0$ along with the line $\tilde{\theta} - \hat{\theta} = 0$. From the plot it is clear that the GMLE estimates are consistently smaller than the PLE estimates. The two-stage estimator is at least 95% efficient for θ_0 relative to

Table 3. Relative Efficiency of the GMLE and the PLE for varying θ_0 .

θ_0	GMLE		PLE		Relative Efficiency	Close-ness
	Mean $\tilde{\theta}$	SE $\tilde{\theta}$	Mean $\hat{\theta}$	SE $\hat{\theta}$		
0.2	0.183	0.145	0.202	0.151	0.960	0.464
1.0	0.971	0.264	1.016	0.272	0.972	0.453
6.0	5.667	0.935	5.804	0.976	0.958	0.438

the GMLE. The loss of efficiency is offset by reduced bias; in all cases the PLE is more frequently closer to θ_0 . Genest, Ghoudi, and Rivest (1995) found that the efficiency of a PLE was poor compared with a parametric maximum likelihood estimator when dependence was strong. The relative efficiency compared with the GMLE is not surprising. Glidden and Self (1999) showed that the independence-working estimators of marginal parameters were highly efficient relative to the GMLE for the Clayton-Oakes model. The methods differ greatly in their computational burden. The GMLE $\tilde{\theta}$ is obtained by maximizing the profile likelihood in θ_0 using the golden section search. This method requires at least 20 evaluations of the profile likelihood to bracket the maximum within a narrow interval. Each evaluation of the profile likelihood involves iterating an EM algorithm to convergence. In contrast, finding the PLE by Newton-Raphson requires perhaps 7 evaluations of (5), (6) and the derivative of (6).

5. Examples

5.1. Australian Twins Study

The Australian Twins Study (Duffy et al., 1990) was conducted to analyze the association between monozygotic twins and dizygotic twins in various diseases. The study collected information on a number of diseases, including appendicitis. Subjects were asked if they had undergone appendectomy and at what age the procedure was performed. We analyzed a subset previously reported by Hsu and Prentice (1996), data on all female twin pairs. The sample included 1214 monozygotic and 735 dizygotic twin pairs. Five hundred ninety and 334 respondents underwent appendectomy in the monozygotic and dizygotic pairs respectively. We fit the Clayton-Oakes model with a common baseline hazard among all subjects ($K = 1$) and ($L = 2$). The estimate of θ_0 was 1.8 with an estimated standard error of 0.27 among monozygotic twin. For dizygotic twins $\hat{\theta}$ was 0.83 with an estimated standard error 0.25. The parameter estimates were very similar to those of Hsu and Prentice (1996) who estimated θ_0 to be 1.7 and 0.82 with a standard errors of 0.25 and 0.25 for the monozygotic and dizygotic twins respectively.

In a bivariate Clayton-Oakes model, the hazard functions at time t_1 for the conditional distributions of T_1 given $T_2 = t_2$ and given $T_2 \geq t_2$ have ratio $1 + \theta_0$, not depending on t_1 or t_2 . The parameter $1 + \theta_0$ describes the modification in the hazard of appendectomy for one twin induced by appendectomy in the other. This ratio is estimated to be 2.8 for monozygotic twins and 1.83 for dizygotic twins.

5.2. *Familial Aggregation of Mental Illness*

The Maryland Schizophrenia Study (Pulver and Liang, 1991) has recruited a cohort of female schizophrenic probands and their first-degree relatives to study the familial tendency of age at onset of mental illness. A primary question of the study is to determine whether the age of onset of schizophrenia in the probands was associated with age of onset of affective illness in the relatives. Age of affective illness in the study is defined as the age at which the relative developed a diagnosis of depression, mania or bipolar.

We studied a subset of 93 female proband and 487 of their first-degree relatives (273 male and 214 females). Thirty one episodes of affective illness were recorded in the relatives. There is some evidence of familial aggregation with 50% of events occurring in 9% of the families. The sex of the relatives is also predictive with 74% events occurring in female subjects. We fit a Clayton-Oakes model with distinct marginal hazard functions by sex ($k = 1, 2$) and a common hazard function among subjects of the same sex ($l = 1, \dots, 10$). This model nonparametrically accounts for the different rates of affective illness between the sexes. The estimate of θ by the two-stage estimator is 1.6 with a standard error of 1.2.

6. Discussion

This paper applies the approaches of Shih and Louis (1995) and Genest, Ghouli and Rivest (1995) to marginal distributions which follow a general proportional hazards model specification. This paper proves that the two-stage estimator is consistent, asymptotically normal, with an asymptotic representation as the sum of i.i.d. random variables. The variance of the estimator has a sandwich-like form. The “meat” of the sandwich includes a penalty for the estimation of $\hat{\beta}$ and $\hat{\Lambda}_k(\cdot)$, ($k = 1, \dots, K$). However, as has been noted previously, the penalty is 0 when failures in a cluster are independent. The method also allows for confidence intervals to be built based on pseudo likelihood ratio statistics. The theory developed for such a statistic parallels the results of Liang and Self (1996). Specifically, the statistic converges weakly to a constant times a chi-squared distribution.

The method is very flexible and computationally simple. The small sample results show the method has a lesser negative bias than the previous methods, and retains high efficiency relative to the computationally demanding likelihood estimator developed by Nielsen et al. (1992). The author is developing portable software to implement the method described in the paper.

Appendix

Formulae

When $\theta_0=0$, the formula given in (9) has the form

$$\begin{aligned} \tilde{\pi}_k(t) = n^{-1} \sum_{i=1}^n \left[\sum_{l=1}^L e^{\beta_0^T Z_{ikl}} Y_{ikl}(t) \left\{ \sum_{k=1}^K \sum_{l=1}^L H_{ikl} - N_{ikl}(\tau) \right\} \right. \\ \left. + \sum_{l=1}^L e^{\beta_0^T Z_{ikl}} Y_{ikl}(t) \{N_{ikl}(\tau) - H_{ikl}\} \right] \end{aligned}$$

and formula (10) has the form

$$\begin{aligned} \tilde{F} = n^{-1} \sum_{i=1}^n \left[\sum_{k=1}^K \sum_{l=1}^L Z_{ikl} H_{ikl} \left\{ \sum_{k=1}^K \sum_{l=1}^L H_{ikl} - N_{ikl}(\tau) \right\} \right. \\ \left. + \sum_{k=1}^K \sum_{l=1}^L Z_{ikl} H_{ikl} \{N_{ikl}(\tau) - H_{ikl}\} \right]. \end{aligned}$$

The random variables Φ_1, \dots, Φ_n have the form

$$\Phi_i = \phi_i(\theta_0) + \sum_{k=1}^K \int_0^\tau \pi_k(s) d\Psi_{ik}(s) + F^T A^{-1} w_{i..}$$

and random variables $\hat{\Phi}_1, \dots, \hat{\Phi}_n$ have the form

$$\hat{\Phi}_i = \hat{\phi}_i + \sum_{k=1}^K \int_0^\tau \hat{\pi}_k(s) d\hat{\Psi}_{ik}(s) + \hat{F}^T \hat{A}^{-1} \hat{w}_{i..}$$

where

$$\begin{aligned} \hat{\phi}_i = \int_0^\tau \frac{N_{i..}(u-)}{1 + \hat{\theta} N_{i..}(u-)} dN_{i..}(u) + \hat{\theta}^{-2} \log\{\hat{R}_i(\hat{\theta})\} \\ - \{\hat{\theta}^{-1} + N_{i..}(\tau)\} \hat{R}_i^{-1}(\hat{\theta}) \sum_{k=1}^K \sum_{l=1}^L \hat{H}_{ikl} e^{\hat{\theta} \hat{H}_{ikl}} + \sum_{k=1}^K \sum_{l=1}^L N_{ikl}(\tau) \hat{H}_{ikl}. \end{aligned}$$

and

$$\hat{\Psi}_{ik}(t) = \int_0^t \frac{d\hat{M}_{ik.}(u)}{S_k^{(0)}(\hat{\beta})} + \hat{h}^T \hat{A}^{-1} \hat{w}_{i..}$$

The functions $\hat{\pi}_k(t)$ are equal to $\tilde{\pi}_k(t)$ ($k = 1, \dots, K$) replacing parameter values by their estimators. The vector \hat{F} replaces parameter values in \tilde{F} by their sample estimators. The

expressions above involve formulae given below:

$$\hat{M}_{ik.}(t) = \sum_{l=1}^L \hat{M}_{ikl}(t) \quad \hat{M}_{ikl}(t) = N_{ikl}(t) - \int_0^t Y_{ikl}(s) e^{\hat{\beta}^T Z_{ikl}} d\hat{\Lambda}_k(s)$$

$$\hat{h} = - \int_0^t E_k(\hat{\beta}, s) d\hat{\Lambda}_k(s) \quad \hat{A} = n^{-1} \sum_{i=1}^n \sum_{k=1}^k \sum_{l=1}^L V_k(\hat{\beta}, u) dN_{ikl}(u)$$

$$\hat{w}_{i..} = \sum_{k=1}^K \sum_{l=1}^L \int_0^\tau \{Z_{ikl} - E_k(\hat{\beta}, u)\} d\hat{M}_{ikl}(u)$$

$$E_k(\beta, t) = \frac{S_k^{(1)}(\beta, t)}{S_k^{(0)}(\beta, t)}, \quad V_k(\beta, t) = \frac{S_k^{(2)}(\beta, t)}{S_k^{(0)}(\beta, t)} - E_k(\beta, t)^{\otimes 2},$$

$$S_k^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^n \sum_{l=1}^L Y_{ikl}(t) e^{\beta^T Z_{ikl}} Z_{ikl}^{\otimes r}$$

where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$. The representation for the theoretical quantities are given by replacing estimators by their estimands and empiricals by limits.

Acknowledgments

The author thanks David Duffy, Kung-Yee Liang, and Karen Bandeen-Roche. The paper was greatly improved by comments given by the Associate Editor and two reviewers. This paper was supported by grant AI38855 from the National Institutes of Health.

References

- D. G. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika* vol. 65 pp. 141–151, 1978.
- D. R. Cox, "Regression models and life-tables (with discussion)," *Journal of the Royal Statistical Society, Series B* vol. 34 pp. 187–220, 1972.
- D. R. Cox and D. V. Hinkley, *Theoretical Statistics*, Chapman and Hall: New York, 1974.
- D. L. Duffy, N. G. Martin, and J. D. Mathews, "Appendectomy in Australian twins," *American Journal of Human Genetics* vol. 47 pp. 590–592, 1990.
- C. Genest, K. Ghoudi, and L.-P. Rivest, "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions," *Biometrika* vol. 82 pp. 543–552, 1995.
- D. V. Glidden and S. G. Self, "Semiparametric likelihood estimation in the Clayton-Oakes model," *Scandinavian Journal of Statistics*, vol. 26 pp. 363–372, 1999.
- G. Gong and F. J. Samaniego, "Pseudo maximum likelihood estimation: theory and applications," *Annals of Statistics* vol. 9 pp. 861–869, 1981.

- L. Hsu and R. L. Prentice, "On assessing the strength of dependency between failure time variates," *Biometrika* vol. 83 pp. 491–506, 1996.
- E. L. Lehmann, *Theory of Point Estimation*, Wiley: New York, 1983.
- K.-Y. Liang and S. G. Self, "On the asymptotic behaviour of the pseudolikelihood ratio test statistic," *Journal of the Royal Statistical Society, Series B* vol. 58 pp. 785–796, 1996.
- D. Y. Lin, "Cox regression analysis of multivariate failure time data: the marginal approach," *Statistics in Medicine* vol. 13 pp. 2233–2247, 1994.
- G. G. Nielsen, R. D. Gill, P. K. Andersen, and T. I. A. Sørensen, "A counting process approach to maximum likelihood estimation in frailty models," *Scandinavian Journal of Statistics* vol. 19 pp. 25–43, 1992.
- D. Oakes, "A model for association in bivariate survival data," *Journal of the Royal Statistical Society, Series B* vol. 44 pp. 414–422, 1982.
- D. Oakes, "Semiparametric inference in a model for association in bivariate survival data," *Biometrika* vol. 73 pp. 353–361, 1986.
- R. L. Prentice and L. Hsu, "Regression on hazard ratios and cross ratios in multivariate failure time analysis," *Biometrika* vol. 84 pp. 349–363, 1997.
- A. E. Pulver and K.-Y. Liang, "Estimating effects of probands' characteristics on familial risk: II. the association between age at onset and familial risk in the Maryland schizophrenia sample," *Genetic Epidemiology* vol. 8 pp. 339–350, 1991.
- J. H. Shih and T. A. Louis, "Inferences on the association parameter in copula models for bivariate survival data," *Biometrics* vol. 51 pp. 1384–1399, 1995.
- G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*, Wiley: New York, 1986.
- C. F. Spiekerman and D. Y. Lin, "Marginal regression models for multivariate failure time data," *Journal of the American Statistical Association* vol. 93 pp. 1164–1175, 1998.
- J. W. Vaupel, K. G. Manton, and E. Stallard, "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography* vol. 16 pp. 439–454, 1979.
- L. J. Wei, D. Y. Lin, and L. Weissfeld, "Regression analysis of multivariate incomplete failure time data by modeling marginal distributions," *Journal of the American Statistical Association* vol. 84 pp. 1065–1073, 1989.