Multivariate Accelerated Failure Time Model with Generalized Estimating Equations

Sy Han Chiou[1], Junghi Kim[2], and Jun Yan[1,3,4]

[1]Department of Statistics, University of Connecticut

[2]Division of Biostatistics, University of Minnesota

[3]Institute for Public Health Research, University of Connecticut Health Center

[4]Center for Environmental Sciences & Engineering, University of Connecticut

Abstract

The accelerated failure time model has not been as widely used as the Cox relative risk model mainly due to computation difficulties. Recent developments in least squares estimation and induced smoothing estimating equations provide promising tools to make the accelerate failure time models more attractive in practice. This paper focuses on multivariate accelerated failure time models. We propose a generalized estimating equation approach to account for the multivariate dependence through working correlation structures. The marginal error distributions can be either identical as in sequential event settings or different as in parallel event settings. Some regression coefficients can be shared across margins as needed. The initial estimator is a rank-based estimator with Gehan's weight, but obtained from an induced smoothing approach with computation ease. The resulting estimator are consistent and asymptotically normal, with a variance estimated through a multiplier resampling method. In a simulation study, our estimator is shown to be up to three times as efficient as the initial estimator, especially with stronger multivariate dependence and heavier censoring percentage. Two real examples are used to demonstrate the proposed method.

Key Words: efficiency; least squares; multivariate survival; AFT model; accelerated failure time model

1 Introduction

Multivariate failure times are commonly encountered in biomedical research where there are natural clusters. The failure times within the same cluster are correlated. Even though the primary interest most often lies

in the marginal effects of covariates on the failure times, accounting for the within-cluster dependence may lead to more efficient regression coefficients estimator. For non-censored multivariate data, the generalized estimating equations (GEE) approach (Liang and Zeger, 1986) has become an important piece in statisticians' toolbox for marginal regression. For censored multivariate failure times, the marginal accelerated failure time (AFT) model is a counterpart of the marginal model. This paper aims to develop a GEE approach to make inferences for multivariate AFT models by taking advantage of recent developments on AFT models with nice computation properties.

An AFT model is a linear model for the logarithm of the failure times with error distribution being unspecified. A nice interpretation of this model is that the effect of a covariate is to multiply the predicted event time by some constant. It provides an attractive alternative to the popular multivariate relative risk model (Cox, 1972). Three main classes of estimator exist for univariate AFT models. The Buckley–James (BJ) estimator extends the least squares principle to accommodate censoring, obtained by an EM algorithm which iterates between imputing the censored failure times and least squares estimation (Buckley and James, 1979). Despite the nice asymptotic properties (Ritov, 1990; Lai and Ying, 1991), the BJ estimator may be hard to get as the EM algorithm may not converge. Further, the limiting covariance matrix is difficult to estimate because it involves the unknown hazard function of the error term. The second type is the rank-based estimator motivated by inverting the weighted log-rank test (Prentice, 1978). Its asymptotic properties has been rigorously studied by Tsiatis (1990) and Ying (1993). Due to lack of efficient and reliable computing algorithm, the rank-based estimator has not been widely used in practice either. Numerical strategies for drawing inference were developed recently by Huang (2002) and Strawderman (2005). The third type is obtained by minimizing an inverse probability of censoring weighed (IPCW) loss function (Robins and Rotnitzky, 1992). The IPCW estimator is easy to compute, consistent and asymptotically normal (Zhou, 1992; Stute, 1993, 1996), but it requires correct specification of the conditional censoring distribution and overlapping of the supports of the censoring time and the failure time.

More recent works have led to a promising perspective on bringing AFT models into routine data analysis practice. For rank-based inference, Jin et al. (2003) proposed a linear programming approach, exploiting that the weighted rank estimating equation is the gradient of an objective function which can be readily solved by linear programming. Variances of the estimators are obtained from a resampling method. A computationally more efficient approach for rank-based inference with Gehan's weight (Gehan, 1965) is the induced smoothing procedure of Brown and Wang (2007). This approach is an application of the general induced smoothing method of Brown and Wang (2005), where the discontinuous estimating equations are replaced with a smoothed version, whose solutions are asymptotically equivalent to those of the former. The

smoothed estimating equations are differentiable, which facilitates rapid numerical solution and sandwich variance estimator. Jin et al. (2006a) suggested an iterative least-squared procedure that starts from a consistent and asymptotically normal initial estimator such as the one obtained from the rank-based method of Jin et al. (2003). The estimator is consistent and asymptotically normal, with variance estimated from a multiplier resampling approach.

For multivariate AFT models, Jin et al. (2006b) developed rank-based estimating equations that are solved via linear programming for marginal regression parameters. Johnson and Strawderman (2009) extended the induced smoothing approach for a rank-based estimator with Gehan's weight to the case of clustered failure times and showed that the smoothed estimates perform as well as those from the best competing methods at a fraction of the computational cost. Jin et al. (2006a) considered their least squares method with marginal models for multivariate failure times. All these approaches used independent working model and left the within-cluster dependence structure unspecified. Li and Yin (2009) developed a generalized method of moments approach for rank-based estimator using the quadratic inference function approach (Qu et al., 2000) to incorporate within-cluster dependence. Wang and Fu (2011) incorporated within-cluster ranks for the Gehan type estimator with the aid of induced smoothing. To the best of our knowledge, little work has been done to extend the GEE approach to setting of multivariate AFT models except a technical report (Hornsteiner and Hamerle, 1996), where the BJ estimator was combined with GEE. Nevertheless, having no access to recent advances on AFT models, they did not solve the convergence problems, and their asymptotic variance estimator formula could not be easily computed because it depends on the derivatives of imputed failure times with respect to regression parameters, which might explain why their variance estimator always overestimated the true variance.

We propose an iterative GEE procedure to account for multivariate dependence through a working covariance or weight matrix. This method has the same spirit as GEE in that misspecification of the working covariance matrix does not affect the consistency of the parameter estimator in the marginal AFT models; when the working covariance is close to the unknown truth, the estimator has higher efficiency than that from working independence as used in Jin et al. (2006a). Our initial estimator is the computationally efficient, rank-based estimator from Johnson and Strawderman (2009), whose consistency and asymptotic normality is inherited by the resulting GEE estimator. We develop methods for cases where all marginal distribution are identical and for cases where at least some margins are different. All the methods are implemented and made publicly available in an open source R package `aftgee` (Chiou and Yan, 2011).

The rest of the article is organized as follows. The multivariate accelerated failure time model and the notations are introduced in Section 2. In Section 3, we propose the GEE update of a consistent initial

estimator and our choice of the initial estimator. Both section 2 and Section 3 are further divided into subsections based on the identities of error distributions. A large scale simulation study is reported in Section 4 to assess the properties of the proposed estimator. Two real applications are illustrated in Section 5. A discussion is concluded in Section 6. The proofs are relegated to the appendix.

2 Multivariate Accelerated Failure Time Model

There are two types of multivariate failure times depending on whether the multiple events are parallel or sequential. The difference between the two types is that the dimension is fixed for parallel data while random for sequential data. In a regression model, we often expect to have different covariates and different coefficients at each margin for parallel data. For sequential data, however, some or all covariates and coefficients may be the same across margins. In general, it is desirable to allow some of the regression coefficients to be shared across margins as needed. We develop the methodology for parallel data for notational simplicity but comment when appropriate on how to adapt to sequential data.

Consider a random sample formed by n clusters. For parallel data, all clusters are of size K while for sequential data, cluster i may have size K_i. For ease of notation, assume at the moment that the cluster sizes are all equal to K. For $i = 1, 2, \cdots, n$ and $k = 1, 2, \cdots, K$, let T_{ik} and C_{ik} be, respectively, the log-transformed failure time and censoring time for margin k in cluster i. Let $Y_{ik} = \min(T_{ik}, C_{ik})$ and $\Delta_{ik} = I(T_{ik} < C_{ik})$. We stack Y_{ik}, T_{ik}, C_{ik}, and Δ_{ik}, $k = 1, \ldots, K$, to form vector Y_i, T_i, C_i, and Δ_i, respectively. Let $X_i = \{X_{i1}, \ldots, X_{iK}\}^\top$ be a $K \times p$ covariate matrix, with the kth row denoted by X_{ik}. The observed data are independent and identically distributed copies of $\{Y, \Delta, X\}$: $\{(Y_i, \Delta_i, X_i) : i = 1, \ldots, n\}$. We assume that T_i and C_i are conditionally independent given X_i.

Our multivariate accelerated failure time model is

$$T_i = X_i\beta + \epsilon_i, \tag{1}$$

where β is a $p \times 1$ vector of regression coefficients, and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iK})^\top$ is a random error vector with an unspecified multivariate distribution. This formulation accommodates margin-specific regression coefficients, in which case, β is a stack of all marginal coefficients, and X_i is a block diagonal matrix. The error vectors ϵ_i's, $i = 1, \ldots, n$, are independent and identically distributed. For parallel data, all K marginal distributions are different, while for sequential data, the number of unique marginal distributions may be smaller or even one as in a recurrent event setting.

With incomplete data from censoring, Buckley and James (1979) replaced each response T_{ik} with its conditional expectation $\hat{Y}_{ik}(\beta) = E_\beta(T_{ik}|Y_{ik}, \Delta_{ik}, X_{ik})$, where the expectation is evaluated at regression

coefficients β. The conditional expectation used to replace Y_{ik} is computed differently depending on the identifications of error terms. Let $\hat{Y}_i(\beta) = \left(\hat{Y}_{i1}(\beta), \ldots, \hat{Y}_{iK}(\beta)\right)^\top$. Jin et al. (2006a) defined

$$U(\beta, b) = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^\top \left(\hat{Y}_i(b) - X_i^\top \beta\right) = 0, \tag{2}$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$, and b is an initial estimator of β. The solution for $U(\beta, \beta)$ is the Buckley-James estimator. The advantage for fixing the initial value b is to avoid solving for $U(\beta, \beta)$ which is neither continuous nor monotone in β. Let the $L(b)$ be the solution for $U(\beta, b) = 0$. Then $L(b)$ has a closed form solution,

$$L(\beta) = \left\{\sum_{i=1}^{n} (X_i - \bar{X})^\top (X_i - \bar{X})\right\}^{-1} \left[\sum_{i=1}^{n} (X_i - \bar{X})^\top (\hat{Y}_i(b) - \bar{Y}(b))\right], \tag{3}$$

where $\bar{Y}(b) = \sum_{i=1}^{n} \hat{Y}_i(b)/n$. With a consistent initial value, equation (3) leads to an iterative algorithm: $\hat{\beta}_{(m)} = L(\hat{\beta}_{(m-1)})$, $m \geq 1$. Once converged, $\hat{\beta}_{(m)}$ is consistent and asymptotically normal and has closed form for every m.

Although the estimator from equation (2) is consistent, it may be inefficient as it completely ignores the dependence. We propose to accomodate dependence using the GEE approach. The method of Jin et al. (2006a) will become a special case with working independence.

3 Inference with GEE

Define $\hat{Y}_{ik}(\beta) = E_\beta(T_{ik}|Y_{ik}, \Delta_{ik}, X_{ik})$, the conditional expectation of T_{ik} evaluated at β, and form $\hat{Y}_i(\beta)$ by staking $\hat{Y}_{ik}(\beta)$, $k = 1, \ldots, K$. For a given initial estimator b of β, we obtain an updated estimator by solving the GEE

$$U(\beta, b, \alpha) = \sum_{i=1}^{n} (X_i - \bar{X})^\top \Omega_i^{-1}(\alpha) \left(\hat{Y}_i(b) - X_i\beta\right) = 0, \tag{4}$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$, and $\Omega_i^{-1}(\alpha)$ is a $K \times K$ working weight matrix which may involve additional working parameters α. For given α and b, the solution of the GEEs (4) has a closed-form

$$L(b, \alpha) = \left\{\sum_{i=1}^{n} (X_i - \bar{X})^\top \Omega_i^{-1}(\alpha)(X_i - \bar{X})\right\}^{-1} \left[\sum_{i=1}^{n} (X_i - \bar{X})^\top \Omega_i^{-1}(\alpha)(\hat{Y}_i(b) - \bar{Y}(b))\right], \tag{5}$$

where $\bar{Y}(b) = \sum_{i=1}^{n} \hat{Y}_i(b)/n$. We propose to estimate β using the following iterative procedure:

1. Obtain an initial estimate $\hat{\beta}_{(0)}$ of β and initialize with $m = 1$.
2. Obtain an estimate $\hat{\alpha}_{(m-1)}$ of α given $\hat{\beta}_{(m-1)}$.
3. Update with $\hat{\beta}_{(m)} = L(\hat{\beta}_{(m-1)}, \hat{\alpha}_{(m-1)})$.

4. Increase m by one and repeat 2 and 3 until convergence.

Details of step 2, construction of working weight, will be given below. As in Jin et al. (2006a), a consistent and asymptotically normal estimator is important for avoiding convergence problems.

The generalized estimating equations are most efficient when Ω_i is chosen to be the covariance of $\hat{Y}_i(b)$. When Ω_i's are the identity matrix (working independence with all marginal variances the same), our estimator reduces to the least squares estimator of Jin et al. (2006a). The working covariance matrix Ω_i's are the same when all clusters have the same size K; they only vary with i when the cluster sizes are not equal.

For convenience, we assume from now on that $E(\epsilon_{ik}) = 0$, $n = 1, \ldots, n$, $k = 1, \ldots, K$. This can be done by incorporating appropriate columns of ones in X_i, and, hence, adding intercepts in β. Our construction of working covariance involves filling element Ω_{kl}, $1 \le k, l \le K$, of the covariance matrix Ω, To allow arbitrary number of unique marginal distributions, let $m_k \in \{1, \ldots, K\}$ be the index of the kth margin among the unique marginal distributions. The conditional expectation $E(T_{ik} \mid Y_{ik}, \Delta_{ik}, X_{ik})$ evaluated at regression coefficients β is computed as

$$\hat{Y}_{ik}(\beta) = \Delta_{ik} Y_{ik} + (1 - \Delta_{ik}) \left[\frac{\int_{e_{ik}(\beta)}^{\infty} u \mathrm{d}\hat{F}_{k,\beta}(u)}{1 - \hat{F}_{k,\beta}\{e_{ik}(\beta)\}} + X_{ik}^{\top}\beta \right],$$

where $e_{ik}(\beta) = Y_{ik} - X_{ik}^{\top}\beta$ and $\hat{F}_{k,\beta}$ is the pooled Kaplan–Meier estimator of the distribution function $F_{k,\beta}$ from the transformed data $\{e_{ik}(\beta), \Delta_{ik}\}$. Specifically, $\hat{F}_{k,\beta}$ is obtained by pooling data from all margins that share the same marginal distribution,

$$\hat{F}_{k,\beta}(t) = 1 - \prod_{1 \le i \le n, 1 \le r \le K : m_r = m_k, e_{ik} < t} \left(1 - \frac{\Delta_{ir}}{\sum_{j=1}^{n} \sum_{1 \le l \le K : m_l = m_k} I(e_{jl}(\beta) \ge e_{ir}(\beta))} \right).$$

To fill Ω_{kk}, $1 \le k \le K$, evaluate the conditional second moment of $\epsilon_{ik}(b)$ at b given the observed data:

$$\hat{V}_{ik}(b) = \Delta_{ik} e_{ik}^2(b) + (1 - \Delta_{ij}) \frac{\int_{e_{ik}(b)}^{\infty} u^2 \mathrm{d}\hat{F}_{k,b}(u)}{1 - \hat{F}_{k,b}\{e_{ik}(b)\}}, \qquad i = 1, \ldots, n, \quad k = 1, \ldots, K. \tag{6}$$

For a given β, we fill Ω_{kk} by

$$\hat{\Omega}_{kk}(\beta) = \frac{\sum_{1 \le i \le n, 1 \le r \le K : m_r = m_k} \hat{V}_{ik}(\beta)}{n \sum_{1 \le r \le K} I\{m_r = m_k\}}. \tag{7}$$

To fill Ω_{kl}, $k \ne l$, define

$$\hat{e}_{ik}(b) = \hat{Y}_{ik}(b) - X_{ik}^{\top} b \qquad i = 1, \ldots, n, \quad k = 1, \ldots, K. \tag{8}$$

Note that $\hat{e}_{ik}(b)$ is the conditional expectation of the ϵ_i evaluated at b given the observed data. For a given b, we fill Ω_{kl} by

$$\hat{\Omega}_{kl} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{ik}(b) \hat{e}_{il}(b). \tag{9}$$

Parsimonious working covariance structures such as exchangeable (EX) or autoregressive with order 1 (AR1) can be imposed. Parameters α in the working covariance can be estimated based on $\hat{V}_{ik}$'s and $\hat{e}_{ik}$'s using the method of moments as in the non-censored case (Liang and Zeger, 1986). When there were no censoring, the working covariance matrix $\hat{\Omega}$ would converge to the true covariance matrix. This is no longer true when censoring is present. Nevertheless, $\hat{\Omega}$, and consequently, $\hat{\alpha}$, still converges to some limit which still helps to improve the efficiency of the GEE estimation.

Extension to unequal cluster sizes as in a recurrent event setting is straightforward. In this case, it is reasonable to assume identical error distributions, hence, identical variances, across all margins. The working covariance matrix V_i with dimension $K_i \times K_i$ can be constructed with an initial estimator for α for a specified working covariance structure.

Under certain regularity conditions, the asymptotic properties of the resulting estimator are consistent and asymptotically normal. These properties are summarized in the following theorems and their proofs are relegated to the Appendix.

Theorem 1. *Under regularity conditions described in (Lai and Ying, 1991), $\hat{\beta}_{(m)}$ is a strongly consistent estimator of the true parameter β_0.*

Theorem 2. *Under regularity conditions described in (Lai and Ying, 1991) $n^{1/2}(\hat{\beta}_{(m)} - \beta_0)$ converges in distribution to multivariate normal with zero mean.*

The resampling approach developed by Jin et al. (2006a) is used to estimate the covariance matrix of $\hat{\beta}_{(m)}$. Let Z_i, $i = 1, \cdots, n$, be independent and identically distributed positive random variables with $E(Z_i) = \mathrm{Var}(Z_i) = 1$. Define

$$L^*(b) = \left\{ \sum_{i=1}^{n} Z_i (X_i - \bar{X}) \Omega_i\big(\alpha(b)\big) (X_i - \bar{X}) \right\}^{-1} \left[\sum_{i=1}^{n} Z_i (X_i - \bar{X}) \left\{ \hat{Y}_i^*(b) - \bar{Y}^*(b) \right\} \right],$$

where $\alpha(b)$ is an estimator of working correlation parameter given regression coefficients evaluated at b, $\bar{Y}^*(b) = \sum_{i=1}^{n} Z_i \hat{Y}_i(b)/n$,

$$\hat{Y}_{ik}^*(b) = \Delta_{ik} Y_{ik} + (1 - \Delta_{ik}) \left[\frac{\int_{e_{ik}(b)}^{\infty} u \mathrm{d}\hat{F}_{\beta}^*(u)}{1 - \hat{F}_{\beta}^* \{e_{ik}(b)\}} + X_{ik}^{\top} b \right],$$

and

$$\hat{F}_{k,\beta}(t) = 1 - \prod_{1 \le i \le n, 1 \le r \le K: M(r) = M(k), e_{ik} < u} \left(1 - \frac{Z_i \Delta_{ir}}{\sum_{j=1}^{n} \sum_{1 \le l \le K: M(l) = M(k)} Z_i I(e_{jl}(\beta) \ge e_{ir}(\beta))} \right).$$

For any realization of $(Z_1, \ldots, Z_n)$, a bootstrap estimator of β is obtained from $\hat{\beta}_{(m)}^* = L^*(\hat{\beta}_{(m-1)}^*)$. The covariance matrix of $\hat{\beta}_{(m)}$ can be estimated from the sample covariance matrix of a bootstrap sample of $\hat{\beta}_{(m)}^*$.

The consistency of this variance estimator can be proved following arguments similar to those in Jin et al. (2006a, Theorem 5).

4 Simulation Study

We conducted two simulation studies to assess the performance of proposed estimators and compared its efficiency with estimators in Johnson and Strawderman (2009).

The first simulation setting is clustered failure times with identical regression coefficients and identical marginal error distributions. The cluster sizes are fixed at three. For cluster i, the multivariate failure time $T_i = (T_{i1}, T_{i2}, T_{i3})$ is generated from

$$\log T_{ik} = 2 + X_{1ik} + X_{2ik} + \epsilon_{ik},$$

where covariates X_{1ik} is Bernoulli with rate 0.5, X_{2ik} is $N(0, 0.5^2)$, and $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3})$ is a trivariate random vector specified by identical marginal error distributions and a copula for dependence structure. Three marginal error distributions are considered: standard normal, standard logistic, and standard Gumbel, abbreviated by N, L, and G, respectively. The within cluster dependence structure is specified by a Clayton copula with three levels dependence level measured by Kendall's tau: 0, 0.3, and 0.6. Censoring times are independently generated from a uniform distribution over $(0, c)$, where c is selected to achieve three levels of censoring percentage: 0%, 25%, and 50%. We considered random samples of size $n = 200$ clusters. Rank-based estimator with Gehan's weight from the induced smoothing approach of Johnson and Strawderman (2009), denoted as JS, was used as the initial estimator for GEE estimators. The covariance matrix of the estimator was obtained from the resampling approach with 200 bootstrap size in Section 3. For each configuration, we did 1000 replicates.

The results are summarized in Table 1. Two working covariance structures, EX and AR1, were used for the proposed iterative GEE procedure. To save space, only results for nonzero Kendall's tau were reported. All estimators appear to be virtually unbiased. The empirical variation of the estimates and the estimated variation based on the resampling procedure agree closely for all estimators. For a given censoring percentage, as the dependence level increases, the variance of the JS estimator changes little, but the variance of the GEE estimators with both working covariance structures decreases. Further, the variance from the EX structure is in general smaller than that from the AR1 structure, which is expected because the true covariance structure is exchangeable in this simulation setting. For a fixed dependence level, the effect of censoring percentage on the variances of the estimator depends on the marginal error distributions. The variance increases clearly as the censoring gets heavier when the errors are normally distributed, but this pattern is not observed with

Gumbel or logistic marginal error distributions. The relative efficiency of the proposed GEE estimator in relative to the rank-based JS estimator is up to 3.5 in the table (with logistic margin and Kendall's tau 0.6 for β_2).

The second simulation setting is multiple event data with different regression coefficients and different marginal error distributions. The cluster sizes are still fixed at three. For cluster i, the multivariate failure times are generated from

$$\log T_{ik} = \beta_{0k} + \beta_{1k} X_{1ik} + \beta_{2k} X_{2ik} + \epsilon_{ik},$$

where $(\beta_{0k}, \beta_{1k}, \beta_{2k})$, $k = 1, 2, 3$, is the regression coefficient vector for margin k, and $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3})$ is a trivariate random vector specified by three marginal distributions and a copula for dependence. The marginal distributions of ϵ_i are standard normal, standard logistic, and standard Gumbel, respectively, for the first, second and third margin; their copula is Clayton with three dependence levels measured by Kendall's tau: 0, 0.3, and 0.6. The regression coefficients $(\beta_{0k}, \beta_{1k}, \beta_{2k})$ are set to be $(-1, 1, -1)$, $(1, -1, 1)$, and $(1, 1, 1)$, respectively for $k = 1$, 2, and 3. Other settings such as the covariates, censoring time, sample size, initial estimator, bootstrap sample size for variance estimation, replication size are all the same as in the first simulation setting.

The results are summarized in Tables 2. In addition to the JS estimator, GEE estimators with two working covariance structures were reported: EX and unstructured (UN). Similar to the first simulation study, all estimators are virtually unbiased, and their variance estimators are generally close to the empirical variances of the replicates. The variance of the GEE estimators decreases as the dependence gets stronger at any level of censoring percentage. Holding the dependence level, as the censoring percentage increases, the variance increases at the normal margin, but the pattern is different for the other two margins. The variance has little changes at the logistic margin. At the Gumbel margin, it remains its level as the censoring percentage increases from 0 to 25%, but increases notably as the censoring percentage increases from 25% to 50%. There is almost no difference between the two working covariance structures, both leding to about the same relative efficiency compared to the rank-based JS estimator. The relative efficiency of both GEE estimators almost double as Kendall's tau is increased from 0.3 to 0.6.

5 Applications

The first application is the diabetic retinopathy study (DRS) that was started in 1971 (Diabetic Retinopathy Study Research Group, 1976). The research objective was to investigate the efficacy of laser photocoagulation in delaying the onset of severe vision loss. Diabetic retinopathy is the most common and serious eye

Table 1: Summary of simulation results with identical regression coefficients and identical marginal error distributions based on 1000 replications. Empirical SE is the standard deviation of the parameter estimates; Estimated SE is the mean of the standard error of the estimator; RE is the empirical relative efficiencies in relative to the JS estimator.

Marg	τ	Cens	β	Bias			Empirical SE			Estimated SE			RE	
				JS	EX	AR1	JS	EX	AR1	JS	EX	AR1	EX	AR1
N	0.3	0%	β_1	−0.002	−0.003	−0.004	0.087	0.072	0.075	0.084	0.068	0.072	1.492	1.376
			β_2	0.001	0.002	0.002	0.083	0.072	0.074	0.084	0.068	0.071	1.349	1.264
		25%	β_1	−0.008	−0.012	−0.013	0.091	0.073	0.076	0.089	0.073	0.077	1.543	1.415
			β_2	−0.003	−0.005	−0.003	0.093	0.075	0.079	0.090	0.075	0.078	1.550	1.384
		50%	β_1	−0.006	−0.011	−0.011	0.101	0.084	0.088	0.099	0.086	0.090	1.467	1.316
			β_2	−0.004	−0.009	−0.010	0.102	0.084	0.090	0.102	0.089	0.093	1.484	1.281
	0.6	0%	β_1	0.002	0.001	0.001	0.082	0.047	0.050	0.083	0.046	0.050	3.130	2.691
			β_2	0.005	0.001	0.001	0.082	0.045	0.050	0.084	0.046	0.050	3.316	2.697
		25%	β_1	−0.007	−0.009	−0.009	0.092	0.050	0.055	0.088	0.052	0.057	3.322	2.826
			β_2	−0.003	−0.008	−0.007	0.090	0.053	0.058	0.090	0.054	0.058	2.931	2.432
		50%	β_1	−0.003	−0.008	−0.008	0.101	0.063	0.069	0.100	0.069	0.074	2.567	2.144
			β_2	0.000	−0.005	−0.004	0.103	0.070	0.077	0.102	0.071	0.077	2.142	1.815
L	0.3	0%	β_1	−0.001	0.002	0.004	0.138	0.123	0.130	0.142	0.124	0.130	1.258	1.128
			β_2	−0.006	−0.004	−0.004	0.145	0.125	0.130	0.142	0.123	0.128	1.352	1.250
		25%	β_1	−0.020	−0.022	−0.021	0.140	0.117	0.121	0.145	0.121	0.128	1.442	1.341
			β_2	−0.013	−0.017	−0.018	0.153	0.124	0.131	0.147	0.121	0.128	1.512	1.369
		50%	β_1	−0.011	−0.012	−0.012	0.164	0.133	0.140	0.162	0.135	0.143	1.524	1.363
			β_2	−0.008	−0.013	−0.014	0.164	0.137	0.148	0.166	0.137	0.145	1.428	1.231
	0.6	0%	β_1	0.006	0.001	0.000	0.145	0.084	0.093	0.141	0.085	0.093	2.966	2.419
			β_2	0.001	0.002	0.001	0.142	0.082	0.090	0.142	0.085	0.092	3.020	2.505
		25%	β_1	−0.011	−0.014	−0.015	0.145	0.080	0.088	0.145	0.080	0.087	3.245	2.679
			β_2	−0.014	−0.013	−0.013	0.149	0.080	0.088	0.146	0.081	0.088	3.494	2.868
		50%	β_1	−0.009	−0.011	−0.012	0.164	0.089	0.099	0.162	0.094	0.102	3.439	2.778
			β_2	−0.006	−0.011	−0.012	0.161	0.092	0.102	0.165	0.095	0.104	3.036	2.479
G	0.3	0%	β_1	−0.001	0.004	0.005	0.092	0.092	0.096	0.094	0.093	0.096	0.982	0.911
			β_2	0.000	−0.004	−0.005	0.093	0.094	0.096	0.094	0.093	0.096	0.973	0.942
		25%	β_1	−0.007	−0.015	−0.017	0.095	0.086	0.089	0.093	0.085	0.088	1.221	1.155
			β_2	−0.007	−0.012	−0.014	0.094	0.088	0.092	0.094	0.086	0.089	1.140	1.048
		50%	β_1	−0.008	−0.012	−0.012	0.099	0.089	0.091	0.095	0.090	0.093	1.255	1.187

Table 2: Summary of simulation results with different regression coefficients and different marginal error distributions based on 1000 replications. Emprical SE is the standard deviation of the parameter estimates; Estimated SE is the mean of the standard error of the estimator; RE is the empirical relative efficiencies in relative to the JS estimator.

			EST			Empirical SE			Estimated SE			RE	
τ	Cen	β	JS	EX	UN	JS	EX	UN	JS	EX	UN	EX	UN
0.3	0%	β_{11}	0.008	0.003	0.003	0.143	0.122	0.123	0.146	0.120	0.119	1.370	1.351
		β_{21}	0.000	−0.003	−0.004	0.151	0.130	0.130	0.146	0.120	0.119	1.340	1.346
		β_{12}	−0.000	−0.003	−0.002	0.164	0.163	0.164	0.166	0.160	0.159	1.014	1.006
		β_{22}	−0.001	−0.005	−0.005	0.162	0.160	0.161	0.166	0.158	0.157	1.023	1.012
		β_{13}	0.002	−0.004	−0.003	0.242	0.219	0.219	0.247	0.217	0.217	1.221	1.220
		β_{23}	0.007	−0.001	−0.003	0.254	0.227	0.228	0.249	0.217	0.217	1.257	1.248
	25%	β_{11}	0.008	0.004	0.003	0.154	0.131	0.132	0.156	0.127	0.127	1.374	1.368
		β_{21}	−0.005	−0.007	−0.006	0.160	0.132	0.132	0.158	0.129	0.128	1.476	1.478
		β_{12}	−0.006	−0.001	−0.000	0.161	0.151	0.151	0.165	0.148	0.147	1.147	1.150
		β_{22}	−0.003	−0.010	−0.010	0.170	0.154	0.154	0.167	0.149	0.149	1.217	1.209
		β_{13}	0.002	−0.006	−0.006	0.262	0.228	0.230	0.260	0.220	0.219	1.315	1.295
		β_{23}	−0.000	−0.011	−0.012	0.262	0.229	0.228	0.264	0.221	0.221	1.310	1.321
	50%	β_{11}	0.010	0.001	−0.000	0.170	0.144	0.145	0.177	0.146	0.145	1.381	1.376
		β_{21}	−0.018	−0.008	−0.007	0.180	0.150	0.150	0.181	0.148	0.147	1.443	1.434
		β_{12}	−0.006	−0.005	−0.004	0.176	0.153	0.152	0.169	0.149	0.148	1.319	1.342
		β_{22}	0.014	0.004	0.002	0.185	0.165	0.166	0.172	0.153	0.152	1.261	1.241
		β_{13}	0.018	0.001	0.000	0.315	0.270	0.271	0.309	0.262	0.260	1.364	1.352
		β_{23}	0.029	0.006	0.007	0.327	0.283	0.283	0.314	0.264	0.262	1.339	1.339
0.6	0%	β_{11}	0.004	−0.000	−0.001	0.149	0.089	0.087	0.146	0.084	0.092	2.813	2.919
		β_{21}	−0.015	−0.003	−0.002	0.140	0.085	0.085	0.146	0.082	0.090	2.700	2.722
		β_{12}	−0.010	0.000	−0.001	0.167	0.126	0.126	0.165	0.120	0.142	1.754	1.744
		β_{22}	−0.001	−0.000	−0.000	0.169	0.124	0.124	0.165	0.119	0.166	1.873	1.853
		β_{13}	0.003	−0.004	−0.005	0.245	0.159	0.156	0.248	0.156	0.192	2.370	2.451
		β_{23}	−0.003	−0.001	−0.000	0.238	0.158	0.156	0.248	0.154	0.189	2.279	2.326
	25%	β_{11}	0.009	0.003	0.002	0.155	0.093	0.092	0.157	0.091	0.113	2.783	2.858
		β_{21}	−0.007	−0.004	−0.005	0.155	0.093	0.092	0.159	0.093	0.112	2.763	2.798
		β_{12}	0.000	−0.003	−0.002	0.166	0.113	0.113	0.166	0.111	0.114	2.145	2.168
		β_{22}	−0.003	−0.006	−0.006	0.168	0.118	0.118	0.167	0.112	0.114	2.036	2.033
		β_{13}	0.001	0.000	0.000	0.266	0.160	0.160	0.260	0.155	0.175	2.769	2.771

complication of diabetes, which may lead to poor vision or even blindness. A subset of the DRS data for patients with "high-risk" diabetic retinopathy, categorized by risk group 6 or higher, has been analyzed by many authors (e.g., Huster et al., 1989; Liang et al., 1993; Lee and Wei, 1993; Spiekerman and Lin, 1996). Each of the 197 patients in this subset had one eye randomized to laser treatment and the other eye received no treatment. The outcomes of interest were the actual times from initiation of treatment to the time when visual acuity dropped below 5/200 at two visits in a row (defined as "blindness"). By the end of the study, 73% of the times for treated eyes The scientific interest was the effectiveness of the laser treatment and the influence of other risk factors. In addition to the treatment indicator, three covariates are available: age at diagnosis of diabetes, type of diabetes (1 = adult, 0 = juvenile), and risk group (6 to 12, rescaled to 0.5 to 1.0). Since the interaction between treatment and diabetes type was found to be significant in Spiekerman and Lin (1996), we also include this interaction in the model.

We first fit an AFT model with identical error margins and identical regression coefficients for both left and right eyes. The second AFT model we fit is the opposite with different error margins and different regression coefficients for the two eyes. For each model, we report GEE estimators with working independence and working exchangeable covariance structures, in addition to the rank-based JS estimator in Table 3. GEE estimator with exchangeable working structure from the first model suggest that the treatment is significant in delaying the onset of vision loss, it has a significant higher effect for adult than for juvenile, and patients in higher risk groups tend to lose vision sooner. Note that the treatment effect is not significant if working independence is used in the GEE estimator. The second model offers a possibility to check whether the marginal error distributions and regression coefficients should indeed be identical as assumed in the first model. Figure 5 shows the the Kaplan–Meier survival curves of the censored residuals for the left margin and right margin respectively, overlaid with the pooled estimate from the first model. All three curves appear to be mingled together in a tight range. A naive log-rank test to compare the two margins, ignoring that the regression coefficients were not known but estimated, yields a p-value of 0.907, confirming the visual observation. Our joint model also allows testing hypothesis of equal coefficients for each covariate across the two margins. The coefficients of treatment, risk group, and treatment-diabetes interaction were found to be not significantly different across the two margins, with p-values 0.400, 0.278, and 0.147, respectively. The coefficients of age and diabetes were found to be significantly different across the two margins, with p-values 0.036 and 0.042, respectively.

We then fit an AFT model with identical error margins, same coefficients for treatment, risk group and treatment-diabetes interaction, and different coefficients for age and diabetes. This is one of the many models with intermediate complexity between the first and the second models. Results are summarized

Table 3: Results of analyzing Diabetic Retinopathy Study.

		JS		IND		EX	
Marg	Effects	EST	SE	EST	SE	EST	SE
Identical error margins and identical regression coefficients:							
pooled	risk group	−2.659	0.739	−2.408	0.859	−2.306	0.775
	age	−0.010	0.012	−0.010	0.013	−0.010	0.014
	diabetes	−0.140	0.349	−0.065	0.440	−0.065	0.369
	treatment	0.520	0.197	0.545	0.330	0.542	0.263
	interaction	1.116	0.301	0.961	0.466	0.964	0.410
Different error margins and different regression coefficients:							
left	risk group	−2.819	1.114	−2.832	1.195	−2.654	1.242
	age	−0.042	0.016	−0.037	0.019	−0.036	0.020
	diabetes	0.825	0.463	0.706	0.554	0.702	0.544
	treatment	0.925	0.422	0.645	0.549	0.652	0.489
	interaction	1.719	0.650	1.742	0.855	1.739	0.820
right	risk group	−2.087	1.013	−1.944	1.316	−1.805	1.283
	age	0.011	0.014	0.009	0.016	0.009	0.018
	diabetes	−0.770	0.432	−0.640	0.528	−0.639	0.656
	treatment	0.383	0.326	0.481	0.381	0.477	0.446
	interaction	0.752	0.476	0.600	0.639	0.603	0.646
Identical error margins with partial common regression coefficients:							
left	age	−0.039	0.015	−0.036	0.021	−0.036	0.022
	diabetes	0.892	0.406	0.848	0.607	0.846	0.621
right	age	0.011	0.015	0.009	0.019	0.009	0.017
	diabetes	−0.870	0.435	−0.837	0.499	−0.835	0.574
common	treatment	0.630	0.227	0.606	0.250	0.607	0.267
	risk group	−2.588	0.747	−2.409	1.034	−2.264	0.938
	interaction	1.067	0.318	1.014	0.344	1.014	0.409

in the last section of Table 3. This time, the shared coefficients of treatment, risk group, and treatment-diabetes interaction remained significant as before. An interesting finding is that the difference between the coefficient of diabetes (0.846 versus −0.835) is significantly nonzero with a p-value 0.002, suggesting that the adult diabetes have sooner onset of vision loss in right eye than in left eye. This finding has not been reported in existing analyses.

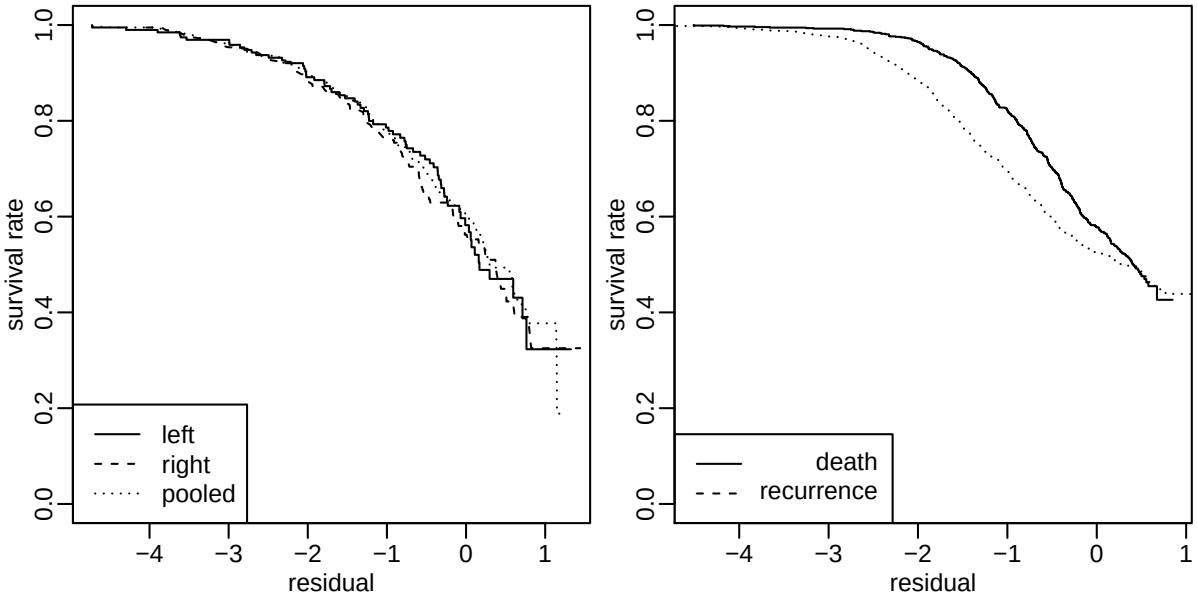

Figure 1: Kaplan–Meier survival curves for censored residuals of the two applications. Left: the DRS Study. Right: the colon cancer stydy.

The second application is the colon cancer study (Lin, 1994). Through randomization, 315, 310 and 304 patients with stage C colon cancer received observation, levamisole alone (Lev), and levamisole combined with fluorouracil (Lev + 5FU), respectively. Lin (1994) considered bivariate models for the time to first recurrence and the time to death. The research interest was the effectiveness of the treatment in prolonging the time to recurrence and time to death. Gender and age are available as covariates besides treatment.

In this application, the error distributions and regression coefficients have no reason to be identical across margins. We report results with different error margin and different regression coefficients in Table 4. Since all covariates are at the cluster level, the exchangeable and independent working covariance structure give the same results (e.g., Hin et al., 2007). The Kaplan–Meier survival curves for the two error margins are shown in Figure 5, which clearly exhibits no similarity; a naive log-rank test gives p-value 0.0008. The treatment of levamisole combined with fluorouracil appears to have a significant positive effect on both event times. The gender and age are found not be significant for either time. The estimated difference between

Table 4: Result of analyzing Colon Cancer Study

		JS		EX	
Event	Effects	EST	SE	EST	SE
recurrence	Lev	0.010	0.124	0.012	0.173
	Lev + 5FU	0.940	0.138	0.931	0.185
	gender	0.310	0.111	0.274	0.161
	age	0.011	0.004	0.012	0.006
dealth	Lev	−0.009	0.104	−0.038	0.131
	Lev + 5FU	0.458	0.108	0.307	0.136
	gender	0.064	0.090	0.066	0.111
	age	−0.003	0.004	−0.004	0.004

the combined treatment effect on recurrence and on death (0.931 versus 0.307) has a standard error 0.103, suggesting that the combined treatment has a higher effect on recurrence than on death.

6 Discussion

The working covariance structure of the proposed GEE approach is different from that in a generalized linear model setting, where the variance is assumed to be a function of the mean. The errors at each margin are assumed to be independent and identically distributed, and hence have the same variance. This assumption might be relaxed by imposing a structure on the variance of the errors. For instance, in model (1), we replace ϵ_{ik} with $\sigma_{ik} \cdot \nu_{ik}$, where ν_{ik}'s is independent and identically distributed for $i = 1, \ldots, n$ with mean zero and variance one, and the scale σ_{ik} may be described by a regression model. Such specification would lead to heteroskedasticity in errors and merits further investigation.

For applications like the DRS study, where there are reasons to impose identical distribution across margins, a rigorous test to compare the survival curves of the residuals would be desirable. We used naive tests that ignore the fact that the residuals were calculated based on estimated regression coefficients. A test procedure should take into account of the variation caused by the estimation procedure.

References

Brown, B. M. and Wang, Y.-G. (2005), "Standard Errors and Covariance Matrices for Smoothed Rank Estimators," *Biometrika*, 92, 149–158.

— (2007), "Induced Smoothing for Rank Regression with Censored Survival Times," *Statistics in Medicine*, 26, 828–836.

Buckley, J. and James, I. (1979), "Linear Regression with Censored Data," *Biometrika*, 66, 429–436.

Chiou, S. H. and Yan, J. (2011), *aftgee:Accelerated Failure Time Model with Generalized Estimating Equations*, r package version 0.2-8.

Cox, D. R. (1972), "Regression Models and Life-Tables (with discussion)," *Journal of the Royal Statistical Society, Series B, Methodological*, 34, 187–220.

Diabetic Retinopathy Study Research Group (1976), "Preliminary Report on Effects of Photocoagulation Therapy," *American Journal of Ophthalmology*, 81, 383–396.

Gehan, E. A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-censored Samples," *Biometrika*, 52, 203–223.

Hin, L.-Y., Carey, V. J., and Wang, Y.-G. (2007), "Criteria for Working Correlation Structure Selection in GEE," *The American Statistician*, 61, 360–364.

Hornsteiner, U. and Hamerle, A. (1996), "A Combined GEE/Buckley-James Method for Estimating an Accelerated Failure Time Model of Multivariate Failure Times," .

Huang, Y. (2002), "Calibration Regression of Censored Lifetime Medical Cost," *Journal of the American Statistical Association*, 97, 318–327.

Huster, W. J., Brookmeyer, R., and Self, S. G. (1989), "Modelling Paired Survival Data with Covariates," *Biometrics*, 45, 145–156.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003), "Rank-based Inference for the Accelerated Failure Time Model," *Biometrika*, 90, 341–353.

Jin, Z., Lin, D. Y., and Ying, Z. (2006a), "On Least-squares Regression with Censored Data," *Biometrika*, 93, 147–161.

— (2006b), "Rank Regression Analysis of Multivariate Failure Time Data Based on Marginal Linear Models," *Scandinavian Journal of Statistics*, 33, 1–23.

Johnson, L. M. and Strawderman, R. L. (2009), "Induced Smoothing for the Semiparametric Accelerated Failure Time Model: Asymptotics and Extensions to Clustered Data," *Biometrika*, 96, 577–590.

Lai, T. L. and Ying, Z. (1991), "Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data," *The Annals of Statistics*, 19, 1370–1402.

Lee, E. W. and Wei, L. J. aand Ying, Z. (1993), "Linear Regression Analysis for Highly Stratified Failure Time Data," *Journal of the American Statistical Association*, 88, 557–565.

Li, H. and Yin, G. (2009), "Generalized Method of Moments Estimation for Linear Regression with Clustered Failure Time Data," *Biometrika*, 96, 293–306.

Liang, K.-Y., Self, S. G., and Chang, Y.-C. (1993), "Modelling Marginal Hazards in Multivariate Failure Time Data," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 55, 441–453.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, D. Y. (1994), "Cox Regression Analysis of Multivariate Failure Time Data: The Marginal Approach," *Statistics in Medicine*, 13, 2233–2247.

Prentice, R. L. (1978), "Linear Rank Tests with Right Censored Data (Corr: V70 P304)," *Biometrika*, 65, 167–180.

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving Generalised Estimating Equations Using Quadratic Inference Functions," *Biometrika*, 87, 823–836.

Ritov, Y. (1990), "Estimation in a Linear Regression Model with Censored Data," *The Annals of Statistics*, 18, 303–328.

Robins, J. M. and Rotnitzky, A. (1992), "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers," in *AIDS Epidemiology – Methodological Issues*, eds. N., J., K., D., and V., F., Boston, MA: Birkhäuser, pp. 297–331.

Spiekerman, C. F. and Lin, D. Y. (1996), "Checking the Marginal Cox Model for Correlated Failure Time Data," *Biometrika*, 83, 143–156.

Strawderman, R. L. (2005), "The Accelerated Gap Times Model," *Biometrika*, 92, 647–666.

Stute, W. (1993), "Consistent Estimation under Random Censorship When Covariables Are Present," *Journal of Multivariate Analysis*, 45, 89–103.

— (1996), "Distributional Convergence under Random Censorship When Covariables Are Present." *Scandinavian Journal of Statistics*, 23, 461–471.

Tsiatis, A. A. (1990), "Estimating Regression Parameters Using Linear Rank Tests for Censored Data," *The Annals of Statistics*, 18, 354–372.

Wang, Y.-G. and Fu, L. (2011), "Rank Regression for Accelerated Failure Time Model with Clustered and Censored Data," *Computational Statistics and Data Analysis*, 55, 2334–2343.

Ying, Z. (1993), "A Large Sample Study of Rank Estimation for Censored Regression Data," *The Annals of Statistics*, 21, 76–99.

Zhou, M. (1992), "M-estimation in Censored Linear Models," *Biometrika*, 79, 837–841.