Journal of Statistical Software

November 2014, Volume 61, Issue 11. http://www.jstatsoft.org/

Fitting Accelerated Failure Time Models in Routine Survival Analysis with R Package aftgee

Sy Han Chiou
University of Minnesota, Duluth

Sangwook Kang
Yonsei University

Jun Yan
University of Connecticut

Abstract

Accelerated failure time (AFT) models are alternatives to relative risk models which are used extensively to examine the covariate effects on event times in censored data regression. Nevertheless, AFT models have been much less utilized in practice due to lack of reliable computing methods and software. This paper describes an R package **aftgee** that implements recently developed inference procedures for AFT models with both the rank-based approach and the least squares approach. For the rank-based approach, the package allows various weight choices and uses an induced smoothing procedure that leads to much more efficient computation than the linear programming method. With the rank-based estimator as an initial value, the generalized estimating equation approach is used as an extension of the least squares approach to the multivariate case. Additional sampling weights are incorporated to handle missing data needed as in case-cohort studies or general sampling schemes. A simulated dataset and two real life examples from biomedical research are employed to illustrate the usage of the package.

Keywords: case-cohort, efficiency, Gehan weight, generalized estimating equation, G^ρ class, induced smoothing, least squares, log-rank, Prentice-Wilcoxon weight, rank-based, weighted estimating equation.

1. Introduction

The linear regression model is the most commonly used regression model in data analysis for uncensored data. When survival data are right-censored, two of the most frequently used regression models are the relative risk model (Cox 1972) and the accelerate failure time (AFT) model (e.g., Kalbfleisch and Prentice 2002, Chapter 4). The AFT model is appealing because it is analogous to the classical linear regression approach, directly linking the expected failure time to covariates. The AFT model with an unspecified error distribution is known as the semiparametric AFT model, which has been studied extensively and is an alternative to the relative risk model with an unspecified baseline hazard function. Two methods for fitting such models have been popular. One is the rank-based approach motivated by inverting the weighted log-rank test (Prentice 1978). Its asymptotic properties have been rigorously studied by Tsiatis (1990) and Ying (1993). The other method is an extension of the least squares principle, such as the Buckley-James (BJ) estimator (Buckley and James 1979). The theoretical properties of the BJ estimator were investigated in Ritov (1990) and Lai and Ying (1991). Due to lack of efficient and reliable computing algorithms, both approaches have not been widely used in practice until recently (Jin, Lin, Wei, and Ying 2003; Jin, Lin, and Ying 2006b,c). Our R package **aftgee** (Chiou, Kang, and Yan 2014c) aims to provide an easy access to AFT models with both methods based on the recent methodological developments. Package **aftgee** is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=aftgee`.

Several packages for AFT models are available for the R environment (R Core Team 2014). For parametric AFT models, where the error distribution is parametrically specified, one can use `survreg` in package **survival** (Therneau 2014), `psm` in package **rms** (Harrel 2014) or `aftreg` in package **eha** (Brostöm 2014). Misspecified error distributions in parametric AFT modeling may lead to bias in estimation and false conclusion under the presents of censoring. For semiparametric AFT models with unspecified error distribution, one can use `bj` in package **rms** (Harrel 2014) or `lss` in package **lss** (Jin and Huang 2007). Function `bj` provides the BJ estimator but it has several limitations: it computes the variance estimator based on non-censored observations only which, although this has been reported to behave well in simulation studies, lacks theoretical justification (Wei 1992); its convergence is slow and not guaranteed; and it is only implemented for univariate failure time data. Package **lss** provides a rank-based estimator with Gehan's weight obtained from a linear programming approach (Jin *et al.* 2003) and a least squares estimator with an iterative algorithm starting from the rank-based estimator (Jin *et al.* 2006b). The variance estimators for both methods are bootstrap based with validity theoretically justified. Nevertheless, there are several features **lss** fall short. Its rank-based estimator is limited to Gehan's weight which may not be the optimal weight (Tsiatis 1990). The linear programming approach used for the rank-based estimator is computationally very intensive, which also affects the least squares estimator through the initial estimator. The bootstrap based variance estimation is very time consuming. Although easily fixable, the package does not support user-specified initial values for the least squares estimator. For clustered failure times, it operates with working independence and disregards the within-cluster dependence, which may lead to efficiency loss especially when the within-cluster dependence is strong (Chiou, Kang, Kim, and Yan 2014a).

Our package **aftgee** overcomes the aforementioned limitations in existing implementations and provides a set of comprehensive tools for semiparametric AFT models in practical survival analysis. For the rank-based estimator with Gehan's weight, we implemented the induced smoothing approach which is much faster than the linear programming approach without loss in accuracy (Brown and Wang 2005, 2007). The induced smoothing approach has been extended to work with any general weight (in addition to Gehan's weight; Chiou, Kang, and Yan 2013). Our efficient sandwich variance estimators provide much faster alternatives to the full bootstrap variance estimation (Chiou, Kang, and Yan 2014b). With the fast rank-based estimators as initial estimators, we implemented an iterative least squares procedure method that extends generalized estimating equations (GEE) to clustered censored data (Chiou *et al.* 2014a). The resulting estimator is robust to misspecification of the working covariance matrix, and the efficiency is higher when the working covariance structure is closer to the truth. Furthermore, these methodologies are generalized to incorporate additional sampling weights for handing missing data and various sampling schemes (Chiou, Kang, and Yan 2014d). Because of these features, the **aftgee** package is appealing to analysts who would like to fit AFT models in their routine analysis of survival data.

The rest of the article is organized as follows. In the next section, we introduce the notations and model formulation for the univariate AFT model. The multivariate extension is presented in Section 3. Incorporation of sampling weight with application to case-cohort data is extended in Section 4. Detailed usages are described in Section 5. A simulated dataset, an univariate example and a multivariate example are used for illustration in Section 6. Conclusion and some remarks are summarized in Section 7.

2. Univariate AFT model

For $i = 1, \ldots, n$, let T_i, C_i and X_i be the log-transformed failure time, censoring time and the $p \times 1$ covariate vector for the ith subject. It is assumed that T_i is conditionally independent of C_i given X_i. An univariate semiparametric AFT model has the form

$$T_i = X_i^\top \beta + \epsilon_i, \qquad i = 1, \ldots, n,$$

where β is an unknown $p \times 1$ vector of regression parameters, ϵ_i's are independent and identically distributed random variables with an unspecified distribution. It is also assumed that ϵ_i's are independent of X_i. In the presence of right censoring, the observed data are independent copies of (Y_i, Δ_i, X_i), $i = 1, \ldots, n$, where $Y_i = \min(T_i, C_i)$, $\Delta_i = I(T_i < C_i)$, and $I(\cdot)$ is the indicator function.

2.1. Rank-based estimator

The regression parameters can be estimated by solving the rank-based weighted estimating equation

$$U_{n,\varphi}(\beta) = \sum_{i=1}^{n} \varphi_i(\beta)\Delta_i \left[X_i - \frac{\sum_{j=1}^{n} X_j I[e_j(\beta) \geq e_i(\beta)]}{\sum_{j=1}^{n} I[e_j(\beta) \geq e_i(\beta)]} \right] = 0, \tag{1}$$

where $e_i(\beta) = Y_i - X_i^\top \beta$ and $\varphi_i(\beta)$ is a possibly data-dependent nonnegative weight function with values between 0 and 1. Let $\hat{F}_{e_i(\beta)}(t)$ be the estimated cumulative distribution function based on the censored residual $e_i(\beta)$'s. Some common choices of $\varphi_i(\beta)$ are 1, $n^{-1} \sum_{i=1}^{n} I[e_j(\beta) \geq e_i(\beta)]$, $1 - \hat{F}_{e_i(\beta)}(t)$ and $[1 - \hat{F}_{e_i(\beta)}(t)]^\rho$, $\rho \geq 0$, corresponding to log-rank (Prentice 1978), Gehan (Gehan 1965), Prentice-Wilcoxon (Prentice 1978) and the more general G^ρ class (Harrington and Fleming 1982), respectively. The Kaplan-Meier estimator is typically used to obtain $\hat{F}_{e_i(\beta)}(t)$. The solution of Equation 1, $\hat{\beta}_{n,\varphi}$, is consistent to the true parameter, β_0, and is asymptotically normal (Tsiatis 1990; Ying 1993). Noting that Equation 1 with Gehan's weight is the gradient of an objective function, Jin *et al.* (2003) used a linear programming approach to obtain the estimator, which is computationally demanding, especially for larger datasets and for obtaining variance estimators through bootstrap. In our implementation, we used the Barzilai-Borwein spectral method implemented in package **BB** (Varadhan and Gilbert 2009) to solve Equation 1 directly.

A computationally more efficient approach is the induced smoothing procedure of Brown and Wang (2005, 2007). The idea is to replace the nonsmooth estimating equations with a smoothed version, whose solutions are asymptotically equivalent to those of the former. Define an independent $p \times 1$ standard normal random vector Z and a $p \times p$ matrix Γ_n such that $\Gamma_n^2 = \Sigma_n$ where Σ_n is a symmetric positive definite matrix. The induced smoothing procedure replaces $U_{n,\varphi}(\beta)$ with $\mathsf{E}_Z[U_{n,\varphi}(\beta + n^{-1/2}\Gamma_n Z)]$, where the expectation is taken with respect to Z. A choice of Σ_n is the identity matrix (Brown and Wang 2007). Some other forms of Σ_n might be considered but the differences are minimal in practice (Chiou *et al.* 2014b).

With Gehan's weight, the denominator of the ratio in Equation 1 gets canceled, and the resulting smooth estimating equation is

$$\tilde{U}_{n,G}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i (X_i - X_j)\Phi\left[\frac{e_j(\beta) - e_i(\beta)}{r_{ij}}\right] = 0, \qquad (2)$$

where $r_{ij}^2 = (X_i - X_j)^\top \Sigma_n (X_i - X_j)$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The estimating equation in Equation 2 is monotone and continuously differentiable with respect to β, hence, its root can be found with standard numerical methods such as the Barzilai-Borwein spectral method. The solution to Equation 2, denoted by $\tilde{\beta}_{n,G}$, is consistent to β_0 and has the same asymptotic distribution as that obtained from the nonsmooth version (Brown and Wang 2005, 2007). Its variance can be approximated from a resampling procedure similar to that in Jin *et al.* (2003). Even with the smoothed equations, the resampling procedure can still be very time consuming. Alternative variance estimation procedures, such as those proposed by Chiou *et al.* (2014b), are recommended.

Deriving the smoothed estimating equations with general weights is challenging because $\mathsf{E}_Z[U_{n,\varphi}(\beta + n^{-1/2}\Gamma_n Z)]$ involves the expectation of the ratio of two random quantities. In addition, $\varphi_i(\beta)$ also depends on β. Holding $\varphi_i(\beta)$ evaluated at some initial estimator b, we propose an approximation which replaces the expectation of the ratio with the ratio of the expectations of the two terms:

$$\tilde{U}_{n,\varphi}(b,\beta) = \sum_{i=1}^{n} \Delta_i \varphi_i(b)\left[X_i - \frac{\sum_{j=1}^{n} X_j \Phi[\kappa_{ij}(\beta)]}{\sum_{j=1}^{n} \Phi[\kappa_{ij}(\beta)]}\right] = 0, \qquad (3)$$

where $\kappa_{ij}(\beta) = [e_j(\beta) - e_i(\beta)]/r_{ij}$. The asymptotic equivalence between Equation 3 and the smooth version of Equation 1 for the log-rank weight is established in Chiou *et al.* (2013). For general weights, the regression parameters can be estimated from an iterative induced smoothing procedure with the following steps:

1. Obtain an initial estimate $\tilde{\beta}_{n,\varphi}^{(0)} = b_n$ of β and initialize with $m = 1$.

2. Update $\hat{\beta}_{n,\varphi}^{(m)}$ by solving $\tilde{U}_{n,\varphi}(\hat{\beta}_{n,\varphi}^{(m-1)}, \hat{\beta}_{n,\varphi}^{(m)}) = 0$.

3. Increase m by one and repeat Step 2 until $|\tilde{\beta}_{n,\varphi,q}^{(m-1)} - \tilde{\beta}_{n,\varphi,q}^{(m)}| < t$ for all $q = 1, \ldots, p$, where $\tilde{\beta}_{n,\varphi,q}^{(m)}$ is the qth component of $\tilde{\beta}_{n,\varphi}^{(m)}$ and t is a prefixed tolerance.

A simple choice of the initial estimator is the easy-to-compute Gehan's estimator, $\tilde{\beta}_{n,G}$.

Since estimating Equation 3 is not necessarily monotone in β, it might cause numerical problems in solving the estimating equations. Inspired by a discussion in Jin *et al.* (2003), a reweighted induced smoothing strategy based on monotone estimating functions is considered (Chiou *et al.* 2013):

$$U_{n,\phi}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i \phi_i(\beta)(X_i - X_j) I[e_j(\beta) \geq e_i(\beta)] = 0, \tag{4}$$

where $\phi_i(\beta) = \varphi_i(\beta)/\sum_{j=1}^{n} I[e_j(\beta) \geq e_i(\beta)]$. Fixing the weight $\phi_i(b)$ evaluated at b and applying induced smoothing on Equation 4 lead to

$$\tilde{U}_{n,\phi}(b,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i \phi_i(b)(X_i - X_j)\Phi\left[\frac{e_j(\beta) \geq e_i(\beta)}{r_{ij}}\right] = 0. \tag{5}$$

This is the same as Equation 2 except for the weight $\phi_i(b)$ which is free from β. For an initial estimator b of β, an estimator $\tilde{\beta}_{n,\phi}$ can be obtained from the iterative procedure with $\tilde{U}_{n,\varphi}(b,\beta)$ replaced by $\tilde{U}_{n,\phi}(b,\beta)$. Using the arguments in Jin *et al.* (2003), the consistency and asymptotic normality of the resulting estimators can be established (Chiou *et al.* 2013). The equations within each iteration can be solved with package **BB**. The convergence is usually fast with the initial Gehan's estimator. Variance estimation can be done with the full resampling method (Jin *et al.* 2003) or a fast sandwich variance estimator (Chiou *et al.* 2014b, 2013).

2.2. Least squares approach

The other method for fitting AFT model is the least squares approach. With survival data from right censoring, Buckley and James (1979) replaced each response T_i with the conditional expectation $\hat{Y}_i(\beta) = \mathsf{E}_\beta(T_i|Y_i, \Delta_i, X_i)$, where the expectation is evaluated at regression coefficients β. In particular,

$$\hat{Y}_i(b) = \Delta_i Y_i + (1 - \Delta_i)\left[\frac{\int_{e_i(b)}^{\infty} t \mathrm{d}\hat{F}_{e_i(\beta)}(t)}{1 - \hat{F}_{e_i(\beta)}[e_i(b)]} + X_i^\top b\right].$$

The theoretical properties of the BJ estimator have been studied by Ritov (1990) and Lai and Ying (1991). The method, however, is rarely used in practice due to numerical challenges.

Jin, Lin, Wei, and Ying (2006a) proposed a more practical solution that generalizes the BJ estimator. Given an initial estimator b_n of β, the least squares estimator is the solution of the following estimating equation

$$U_{n,ls}(\beta, b) = \sum_{i=1}^{n}(X_i - \bar{X})^\top(\hat{Y}_i(b) - X_i\beta) = 0, \tag{6}$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$. The solution to $U_{n,ls}(\beta,\beta) = 0$ is the BJ estimator. The advantage for fixing the initial value b_n is to avoid numerical complexity caused by solving Equation 6 which is neither continuous nor monotone in β. Jin *et al.* (2006a) devised an iterative procedure $\hat{\beta}_{n,ls}^{(m)} = L_n(\hat{\beta}_{n,ls}^{(m-1)})$ for $m > 1$ with $\hat{\beta}_{n,ls}^{(0)} = b_n$ where

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

and $\bar{Y}(b) = \sum_{i=1}^{n} \hat{Y}_i(b)/n$. If the initial estimator b_n is consistent and asymptotically normal, then $\hat{\beta}_{n,ls}^{(m)}$ is also consistent and asymptotically normal for every m (Jin *et al.* 2006b). A good candidate for the initial estimator is the induced smoothing Gehan estimator. The variance of the resulting estimator can be approximated by a resampling procedure (Jin *et al.* 2006b).

3. Multivariate AFT model

Suppose now we have a random sample of n independent clusters with K_i margin in the ith cluster. For $i = 1, \ldots, n$ and $k = 1, \ldots, K_i$, let T_{ik}, C_{ik} and X_{ik} be the log-transformed failure time, censoring time and the $p \times 1$ covariate vector for margin k in cluster i. Further define the censored failure time and censoring indicator to be $Y_{ik} = \min(T_{ik}, C_{ik})$ and $\Delta_{ik} = I(T_{ik} < C_{ik})$. The multivariate AFT model takes the following form

$$T_{ik} = X_{ik}^{\top}\beta + \epsilon_{ik}, \qquad i = 1, \ldots, n, \ k = 1, \ldots, K_i, \tag{7}$$

where β is an unknown $p \times 1$ vector of regression parameters and the error terms, $\epsilon_i = \{\epsilon_{i1}, \ldots, \epsilon_{iK_i}\}$ are independent and identically distributed random variables with an unspecified distribution throughout clusters. In the presence of censoring, the observed data consists of copies of $\{Y_{ik}, \Delta_{ik}, X_{ik}\}$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K_i$ where $Y_{ik} = \min(T_{ik}, C_{ik})$ and $\Delta_{ik} = I(T_{ik} < C_{ik})$.

3.1. Rank-based estimator

The regression parameters in Equation 7 can be estimated from the following rank-based weighted estimating equation

$$U_{n,\varphi}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i} \varphi_{ik}(\beta)\Delta_{ik}\left[X_{ik} - \frac{\sum_{j=1}^{n}\sum_{l=1}^{K_j} X_{jl} I[e_{jl}(\beta) \geq e_{ik}(\beta)]}{\sum_{j=1}^{n}\sum_{l=1}^{K_j} I[e_{jl}(\beta) \geq e_{ik}(\beta)]}\right] = 0, \tag{8}$$

where $e_{ik}(\beta) = Y_{ik} - X_{ik}^{\top}\beta$ and $\varphi_{ik}(\beta)$ is a possibly data-dependent nonnegative weight function. If $K_i = 1$ for all $i = 1, \ldots n$, Equation 8 will reduce to Equation 1. This estimating equation also yields a consistent estimator for β_0 (Jin *et al.* 2006a). Applying the aforementioned induced smoothing technique, the smoothed version of Equation 8 with Gehan's weight is

$$\tilde{U}_{n,G}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\sum_{j=1}^{n}\sum_{l=1}^{K_j} \Delta_{ik}(X_{ik} - X_{jl})\Phi\left[\frac{e_{jl}(\beta) - e_{ik}(\beta)}{r_{ikjl}}\right] = 0, \tag{9}$$

where $r_{ikjl}^2 = (X_{ik} - X_{jl})^{\top}\Sigma_n(X_{ik} - X_{jl})$. The consistency and asymptotic properties continue to hold (Johnson and Strawderman 2009). The multivariate version of Equations 3 and 5 are

$$\tilde{U}_{n,\varphi}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i} \Delta_{ik}\varphi_{ik}(b)\left[X_{ik} - \frac{\sum_{j=1}^{n}\sum_{l=1}^{K_j} X_j\Phi[\kappa_{ikjl}(\beta)]}{\sum_{j=1}^{n}\sum_{l=1}^{K_j} \Phi[\kappa_{ikjl}(\beta)]}\right] = 0, \tag{10}$$

and

$$\tilde{U}_{n,\phi}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\sum_{j=1}^{n}\sum_{l=1}^{K_j} \Delta_{ik}\phi_{ik}(\beta)(X_{ik} - X_{jl})\Phi\left[\frac{e_{jl}(\beta) \geq e_{ik}(\beta)}{r_{ikjl}}\right] = 0, \tag{11}$$

where $\kappa_{ikjl}(\beta) = [e_{jl}(\beta) - e_{ik}(\beta)]/r_{ikjl}$ and $\phi_{ik}(\beta) = \varphi_{ik}(\beta)/\sum_{j=1}^{n}\sum_{l=1}^{K_j} I(e_{jl}(\beta) \geq e_{ik}(\beta))$. The same iterative procedure as for the univariate case can be used and the asymptotic properties of the resulting estimator, $\tilde{\beta}_{n,\phi}$, continue to hold.

3.2. GEE approach

Depending on the set up of the design matrix X_i, the multivariate AFT model accommodates margin-specific regression coefficients, identical regression coefficients across margins, and their mix. For instance, a covariate could be appropriate for one margin but unsuitable for the other; a covariate could have a different effect in the regression model of different margins. The error term ϵ_i's are assumed to be independent and identically distributed across clusters, but within a cluster, the components of $\epsilon_{i1}, \ldots, \epsilon_{iK_i}$ do not need to follow a common distribution and may be correlated. We generalize the GEE approach to multivariate AFT modeling which accounts for multivariate dependence through working correlation structures to improve efficiency (Chiou *et al.* 2014a).

Define $\Omega_i^{-1}\big(\alpha(b)\big)$ to be an $K_i \times K_i$ nonsingular working weight matrix that may involve additional working parameters α that depends on an initial value b of β_0. For $i = 1, \ldots, n$ and $k = 1, \ldots, K_i$, let $\hat{\mathbf{Y}}_i(b)$, $\mathbf{Y}_i$, $\mathbf{T}_i$, $\mathbf{C}_i$ and $\boldsymbol{\Delta}_i$ be $K_i \times 1$ vectors formed by stacking $\hat{Y}_{ik}(b)$, Y_{ik}, C_{ik} and Δ_{ik}, where

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

A generalization of Equation 6 can be expressed as

$$U_{n,GEE}(\beta, b, \alpha) = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})^{\top} \Omega_i^{-1}\big(\alpha(b)\big)(\hat{\mathbf{Y}}_i(b) - \mathbf{X}_i \beta) = 0, \tag{12}$$

where $\bar{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{X}_i / n$. Given α and b, the solution to Equation 12 has a closed form

$$L_n(b, \alpha) = \left[\sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})^{\top} \Omega_i^{-1}\big(\alpha(b)\big)(\mathbf{X}_i - \bar{\mathbf{X}}) \right]^{-1} \left[\sum_{i=1}^{n} (\mathbf{X}_i - \bar{X})^{\top} \Omega_i^{-1}\big(\alpha(b)\big) \Big(\hat{\mathbf{Y}}_i(b) - \bar{\mathbf{Y}}(b)\Big) \right],$$

where $\bar{\mathbf{Y}}(b) = \sum_i^n \hat{\mathbf{Y}}_i(b)/n$.

The GEE estimator, denoted by $\hat{\beta}_{n,GEE}$, can be obtained from an iterative procedure:

1. Obtain an initial estimate $\hat{\beta}_{n,GEE}^{(0)} = b_n$ of β and initialize with $m = 1$.

2. Obtain an estimate $\hat{\alpha}_n$ of α given $\hat{\beta}_{n,GEE}^{(m-1)}$, $\hat{\alpha}_n(\hat{\beta}_{n,GEE}^{(m-1)})$.

3. Update with $\hat{\beta}_{n,GEE}^{(m)} = L_n\big(\hat{\beta}_{n,GEE}^{(m-1)}, \hat{\alpha}_n(\hat{\beta}_{n,GEE}^{(m-1)})\big)$.

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

The iteration proceeds with the aid of function `geese` in package **geepack** (Højsgaard, Halekoh, and Yan 2014; Halekoh, Højsgaard, and Yan 2006). The estimator reduces to the least squares estimator of Jin *et al.* (2006a) when the working weight matrices Ω_i's are the identity matrices. We refer to Chiou *et al.* (2014a) for more details. The working parameter estimate $\hat{\alpha}_n$ does not affect the consistency of the GEE estimator, but may affect its efficiency. Higher efficiency can be achieved if Ω_i is closer to the covariance matrix of $\hat{\mathbf{Y}}_i(b)$ and even an imperfect working weight still improves the efficiency (Chiou *et al.* 2014a). The variance of the estimator can again be estimated by resampling procedures.

4. Incorporating sampling weight

So far, we have assumed that we have full access to the whole data $(Y_{ik}, \Delta_{ik}, X_{ik})$, $i = 1, \ldots, n$, $k = 1, \ldots, K_i$. In many cases, however, the full cohort data may not be available. For example, a case-cohort design (Prentice 1978) is known to be cost-effective when the proportion of the event of interest is rare or covariates are expensive to measure. Under this design, covariates are measured for all the subjects who experienced the event of interest by the end of the observation period but only for a subset of subjects who did not. This design is a special case of the more general stratified case-cohort design where a subcohort is selected via a stratified random sampling from S mutually exclusive strata in the original full cohort. For both designs, covariates are measured only for those who were selected into the sample. Thus, the observed data from each design are not complete and statistical methods which do not account for this missingness in covariates could result in biased estimates. One typical method employed to adjust for biases is to weight a complete observation by the inverse of the inclusion probability.

Suppose a simple random sampling was used at the cluster level. Let ψ_{is} be the strata indicator ($\psi_{is} = 1$ if the ith cluster is in the sth stratum and $\psi_{is} = 0$ otherwise) and ξ_{is} be the sampling indicator ($\xi_i = 1$ if the ith cluster is sampled and $\xi_i = 0$ otherwise). The case-cohort weight, h_i, for cluster i is $h_i = \sum_{s=1}^{S} \xi_i \psi_{is}/p_{n,s}$, where $p_{n,s}$ is the inclusion probability for the sth stratum for $s = 1, \ldots, S$. The weight-adjusted version of Equations 10 and 11 are, respectively,

$$\tilde{U}_{n,\varphi}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i} h_{ik}\Delta_{ik}\varphi_{ik}(b)\left[X_{ik} - \frac{\sum_{j=1}^{n}\sum_{l=1}^{K_j} h_{jl}X_j\Phi(\kappa_{ikjl}(\beta))}{\sum_{j=1}^{n}\sum_{l=1}^{K_j} h_{jl}\Phi(\kappa_{ikjl}(\beta))}\right] = 0, \qquad (13)$$

and

$$\tilde{U}_{n,\phi}(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K_i}\sum_{j=1}^{n}\sum_{l=1}^{K_j} h_i h_j \Delta_i \phi_i(\beta)(X_{ik} - X_{jl})\Phi\left[\frac{e_{jl}(\beta) \geq e_{ik}(\beta)}{r_{ikjl}}\right] = 0. \qquad (14)$$

Note that if we sample all subjects within each strata, then Equations 13 and 14 reduce to Equations 10 and 11, respectively. The variance estimation can be obtained via resampling procedures or fast sandwich variance estimators similar to the unweighted versions (Chiou *et al.* 2014b,d).

5. Package implementation

The two major functions in package **aftgee** are `aftsrr` for the rank-based approach and `aftgee` for the least squares or GEE approach. The synopsis of `aftsrr` is:

```
aftsrr(formula, data, subset, id = NULL, contrasts = NULL,
  strata = NULL, weights = NULL, rankWeights = "gehan", method = "sm",
  variance = "ISMB", B = 100, SigmaInit = NULL, control = aftgee.control())
```

The required arguments are `formula` and `data`. Argument `formula` specifies the model to be fit with the variables coming with `data`. The formula is the same as the argument of function `survreg` in package **survival**, with response created from `Surv`. The '`Surv`' object consists of two columns, where the first column is the survival time or censored time and the second column is the censoring indicator, indicating right censored data. Since ranks are invariant to location shift, the intercept cannot be estimated and the estimation will ignore the intercept term whether it is specified or not. Clusters are defined by vector `id`. The `weights` argument is a vector containing sampling weights (h_i) as described in Section 4. When data arise from a stratified design, a vector of integers that specifies the stratification is indicated in `strata`. The length of the arguments `id`, `weights` and `strata` needs to be the same as the number of observations. The rank weight, controlled by argument `rkWeight`, includes the aforementioned log-rank weight (`"logrank"`), Gehan's weight (`"gehan"`), Prentice-Wilcoxon weight (`"PW"`) and general G^ρ class weight (`"GP"`). Argument `method` determines the type of weighted estimating equations to be used. When `method = "nonsm"`, regression parameters are estimated by directly solving the nonsmooth estimating Equations 1 or 8. When `method = "sm"` and `rkWeight = "gehan"`, the induced smoothing estimating Equations 2 or 9 are used. For the non-Gehan's weights, `method = "sm"` and `method = "monosm"` apply the iterative procedure with the smooth estimating Equations 3 and 5, respectively. The initial values for the variance estimator, or the Σ_n in the smoothing progress, are determined by `sigmainit`. The identity matrix is used for `sigmainit`, if it is left unspecified.

Given a point estimate, variance estimates can be obtained from several approaches which are specified by argument `variance`. A straightforward but computationally inefficient variance estimator is the multiplier bootstrap approach (`"MB"`). A more efficient method is to consider sandwich variance estimators (Chiou *et al.* 2014b). Suppose the variance of the estimator has a sandwich form, $\Sigma = A^{-1}V(A^{-1})^\top$ where V is the asymptotic variance of the estimating function and A is the slope matrix. Chiou *et al.* (2014b) proposed to estimate V by either a closed-form formulation (`CF`) or through bootstrap the estimating equations (`MB`). The bootstrapping estimate of V is much less demanding than the full multiplier bootstrap, because it only involves evaluations of estimating equations instead of solving them. On the other hand, to estimate the slope matrix A, Chiou *et al.* (2014b) proposed three methods based on the induced smoothing approach (`IS`), smoothed Huang's approach (`sH`) motivated by Huang (2002) or Zeng and Lin's approach (`ZL`) by Zeng and Lin (2008). Combinations between estimating V and A yield six sandwich estimators, `"ISCF"`, `"ISCF"`, `"ZLCF"`, `"ZLMB"`, `"sHCF"`, `"sHMB"` for `variance`. When a bootstrap is needed, the bootstrap size is controlled by `B` with default value `100`.

The convergence for the procedure is controlled by relative tolerance. The iteration stops and the output is given when the tolerance is met or iteration reaches the pre-specified maximum iteration number. The default relative tolerance is set at 0.001 and the default maximum number of iterations is set to 50.

```
aftgee.control(maxiter = 50, reltol = 0.001, trace = FALSE)
```

The maximum number of iterations is controlled by `maxiter` and relative convergence tolerance is controlled by `reltol`. A logical value, `trace`, is used to determine whether to print the output for each iteration.

The least squares estimator can be obtained by calling `aftgee` with the following arguments

```
aftgee(formula, data, subset, id = NULL, contrasts = NULL,
  weights = NULL, margin = NULL, corstr = "independence",
  binit = "srrgehan", B = 100, control = aftgee.control())
```

Most of the arguments and the convergence criterion of `aftsrr` are shared by `aftgee`. With `aftgee`, the intercept, if included, is estimated by the mean of the estimated cumulative distribution function based on the censored residual computed from the slope estimator. The `margin` argument is a vector with the same length as data. It is used to specify the marginal distribution within clusters. Identical marginal distributions are assumed with unspecified `margin`. A character string, `corstr`, is used to specify the working correlation structure, as offered by package **geepack**. Four working correlation structures are independence (`"independence"`), exchangeable (`"exchangeable"`), autoregressive model of order one (`"ar1"`) and unstructured (`"unstructured"`). The default is `"independence"`. The initial value is specified by `binit` with default `"srrgehan"` giving the induced smooth rank-based estimator with Gehan's weight. Alternatively, although not recommended, the simple linear regression with censored observations ignored (`"lm"`), can also be used for faster results.

In the uncensored case, `aftgee` with independent working correlation structure will return an ordinary least squares estimate. In the multivariate case, efficiency can be improved in `aftgee` when the working correlation structure is close to the true correlation even in the absent of censoring. A more detailed multivariate illustration is presented in a kidney cather data in Section 6.3.

6. Illustrations

6.1. Simulated data

Let T be the log-transformed failure time generated from the univariate AFT model

$$T = 2 + X_1 + X_2 + \epsilon,$$

where X_1 is Bernoulli with rate 0.5 and X_2 is a standard normal variable. The error term, ϵ, follows an exponential distribution with mean 3. The censoring time was generated from Uniform$(0, \tau)$ with τ adjusted to yield approximately 50% censoring rate. A dataset with 500 subjects was generated with the following code:

```
R> datgen <- function(n = 500, tau = 327) {
+    x1 <- rbinom(n, 1, 0.5)
+    x2 <- rnorm(n)
+    e <- rweibull(n, 1, 3)
+    T <- exp(2 + x1 + x2 + e)
+    cstime <- runif(n, 0, tau)
+    delta <- (T < cstime) * 1
+    Y <- pmin(T, cstime)
+    out <- data.frame(T = T, Y = Y, delta = delta, x1 = x1, x2 = x2)
+  }
R> set.seed(1)
R> mydata <- datgen()
```

On a 3.3 GHz linux machine, we start with the comparison between two versions of Gehan's estimators: one is the nonsmooth version estimated from Equation 1 fitted with `lss` and the other is the smooth version estimated from Equation 2 fitted with `aftsrr`. For `lss`, variances are estimated with the multiplier bootstrap approach with bootstrap sample size 100. For `aftsrr`, in addition to the fully bootstrapping variance estimator from Equation 2 (`"MB"`), the sandwich variance estimator using induced smoothing approach (`"ISMB"`) is also considered.

```
R> library("aftgee")
R> library("survival")
R> library("lss")
R> system.time(rk.lss <- lss(Surv(log(Y), delta) ~ x1 + x2, data = mydata,
+    gehanonly = TRUE, mcsize = 100))

   user  system elapsed
 412.08    1.27  414.88

R> system.time(rk.srrMB <- aftsrr(Surv(Y, delta) ~ x1 + x2,
+    data = mydata, variance = "MB"))

   user  system elapsed
111.040   0.311 111.877

R> system.time(rk.srrISMB <- aftsrr(Surv(Y, delta) ~ x1 + x2,
+    data = mydata, variance = "ISMB"))

   user  system elapsed
  3.276   0.015   3.319

R> rbind(rk.lss = c(rk.lss$betag), srrMB = coef(rk.srrMB),
+    srrISMB = coef(rk.srrISMB), deparse.level = 2)

            x1     x2
rk.lss  0.9412 0.9496
srrMB   0.9399 0.9499
srrISMB 0.9399 0.9499
```

```
R> rbind(rk.lss = rk.lss$gehansd, srrMB = sqrt(diag(vcov(rk.srrMB)$MB)),
+    srrISMB = sqrt(diag(vcov(rk.srrISMB)$ISMB)), deparse.level = 2)

            x1      x2
rk.lss  0.1552 0.07141
srrMB   0.1333 0.07133
srrISMB 0.1385 0.06940
```

The output indicates that `aftsrr` clearly outperforms `lss` in timing. The timing result also suggests that the efficient sandwich estimator is substantially faster.

We next fit the simulated data with the least squares approach. For the parametric approach, `survreg` is used with `dist = "lognormal"`. For the semiparametric approach, the BJ estimator (`bj` from package **rms**), the `lss` estimator and the `aftgee` estimator are considered.

```
R> library("rms")
R> system.time(ls.bj <- bj(Surv(Y, delta) ~ x1 + x2, data = mydata))

   user  system elapsed
  0.089   0.000   0.090

R> system.time(ls.lss <- lss(Surv(log(Y), delta) ~ x1 + x2,
+    data = mydata, mcsize = 100))

Warning: Solution may be nonunique

 Converged. Criteria Satisfied:  0.001
   user  system elapsed
401.506   1.065 403.746

R> system.time(ls.gee <- aftgee(Surv(Y, delta) ~ x1 + x2, data = mydata))

   user  system elapsed
 17.428   0.042  17.504

R> system.time(ls.sur <- survreg(Surv(Y, delta) ~ x1 + x2, data = mydata,
+    dist = "lognormal"))

   user  system elapsed
  0.011   0.000   0.011

R> rbind(bj = coef(ls.bj), lss = c(NA, ls.lss$lse), gee = coef(ls.gee),
+    sur = coef(ls.sur), deparse.level = 2)

    Intercept     x1     x2
bj      4.509 0.9833 0.9332
lss        NA 0.9838 0.9338
gee     4.510 0.9838 0.9338
sur     4.313 0.9126 0.8652
```

		`bj`		`lss`		`aftgee`		`survreg`	
$\hat{\beta}$	β_0	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{\beta}_0$	5.000	−0.480	0.261			−0.473	0.252	−0.701	0.506
$\hat{\beta}_1$	1.000	−0.005	0.035	−0.008	0.029	−0.005	0.033	−0.080	0.031
$\hat{\beta}_2$	1.000	−0.007	0.008	−0.001	0.008	−0.001	0.007	−0.082	0.014

Table 1: Comparison of the `bj`, `lss`, `aftgee` and `survreg` estimators. Bias and MSE represent the bias and mean squared error of the estimator, respectively. The true regression coefficient is β_0. Each cell is the average of 1000 replicates.

```
R> rbind(bj = sqrt(diag(vcov(ls.bj))), lss = c(NA, ls.lss$sd),
+    gee = sqrt(diag(vcov(ls.gee))), sur = sqrt(diag(vcov(ls.sur)))[1:3],
+    deparse.level = 2)
```

```
    Intercept     x1      x2
bj    0.07846 0.1261 0.06448
lss        NA 0.1932 0.09152
gee   0.10588 0.1758 0.09061
sur   0.11078 0.1672 0.08321
```

Estimation for both the `lss` estimator and the `aftgee` estimator are based on a rank-based initial value that is invariant to the intercept. Once the slope estimator is obtained, the `lss` estimator left out the intercept whereas the `bj` and `aftgee` approach estimated the intercept by the mean of the estimated cumulative distribution function based on the censored residual computed from the slope estimator. The semiparametric methods from `bj`, `lss` and `aftgee` provide fairly close point estimates. In terms of timing, the `lss` estimator took the longest with more than six minutes. For further investigation, the estimation performance is assessed via bias and mean squared error with a full scale simulation. Table 1 summarizes the results for 1000 replicates.

The performance of `bj`, `lss` and `aftgee` are similar in terms of the biases and mean squared errors. As expected, when the error distribution is misspecified in the parametric model, the `survreg` approach produced a biased estimate.

6.2. National Wilms' tumor study

We next illustrate the usage of the arguments `weights` and `rkWeight` in `aftsrr` with cohort studies conducted by the national Wilms' tumor study group (NWTSG; D'Angio *et al.* 1989; Green *et al.* 1998). The dataset is available in the **survival** package as `nwtco`. The interest of the study is to assess the relationship between the likability of central lab histology measurement (`histol`) and days to tumor relapse (`edrel`). In addition to the likability of central lab histology measurement (1 = unfavorable, 0 = favorable), we also include patient's age (`age`) in years as covariate. There are two study groups (`study`), NWTSG-3 and NWTSG-4, denoting the third and the fourth Wilms' tumor studies. Patients are further categorized into four stages (`stage`) with stage 4 being the latest and most severest. The dataset consists of 4028 patients, among which, 571 patients experienced tumor relapse (`rel` = 1) and 3457 patients did not (`rel` = 0).

To take advantage of the full cohort data, we fit the full-cohort data with `aftsrr` using two types of sandwich variance estimators (`"ISMB"` and `"ISCF"`).

```
R> data("nwtco", package = "survival")
R> nwtco$age <- nwtco$age/12
R> head(nwtco, 5)

  seqno instit histol stage study rel edrel   age in.subcohort
1     1      2      2     1     3   0  6075 2.083        FALSE
2     2      1      1     2     3   0  4121 4.167        FALSE
3     3      2      2     1     3   0  6069 0.750        FALSE
4     4      2      1     4     3   0  6200 2.333         TRUE
5     5      2      2     2     3   0  1244 4.583        FALSE

R> set.seed(1)
R> system.time(fit.IS <- aftsrr(Surv(edrel, rel) ~ histol + age,
+    data = nwtco, variance = c("ISCF", "ISMB")))

   user  system elapsed
 75.628   1.209  76.917
```

The summary gives the following information:

```
R> summary(fit.IS)

Call:
aftsrr(formula = Surv(edrel, rel) ~ histol + age, data = nwtco,
    variance = c("ISCF", "ISMB"))

Variance Estimator: ISCF
       Estimate StdErr z.value p.value
histol   -3.221  0.144  -22.40  <2e-16 ***
age      -0.231  0.026   -9.03  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Variance Estimator: ISMB
       Estimate StdErr z.value p.value
histol   -3.221  0.152  -21.13  <2e-16 ***
age      -0.231  0.024   -9.49  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All point estimators and variance estimators are close to each other. The coefficients of central histological lab diagnosis is found to be significantly different from zero. In addition, the coefficient for the central lab histological diagnosis is negative. This suggests patients who do not favor the central lab histological diagnosis tend to have shorter time to tumor relapse.

With the same dataset, we next demonstrate incorporating weights via a case-cohort design. Define cases and controls as those who experience the event of interest by the end of the study period and who do not, respectively. In nwtco, there are 571 cases who experienced the relapse of tumor and 3457 controls who did not experience the relapse of tumor. The case-cohort sample is the union of all the cases and the sub-cohort sample selected via a simple random sampling. The case-cohort sample of the data had 1154 subjects, including all 571 cases and 583 controls. This gave sampling weights 1 and 5.93 for the cases and controls, respectively. The following codes give a summary of the case-cohort weight, h_i.

```
R> table(nwtco$in.subcohort, nwtco$rel)
```

```
           0    1
  FALSE 2874  486
  TRUE   583   85
```

```
R> nwtco$in.casecohort <- (nwtco$in.subcohort | nwtco$rel == 1)
R> nwtco$hi <- 0
R> nwtco$hi <- ifelse(nwtco$in.casecohort & nwtco$rel == 1, 1, nwtco$hi)
R> nwtco$hi <- ifelse(nwtco$in.casecohort & nwtco$rel == 0, 5.93, nwtco$hi)
R> table(nwtco$hi)
```

```
   0    1 5.93
2874  571  583
```

For the case-cohort design, we also demonstrate the usage of different rank weights in a rank-based approach; we considered the Gehan's, log-rank and PW weights. For the log-rank and PW weights, the monotone function approach was used. Jin and Huang (2007)'s lss was not considered in this analysis because it does not have the capability of handling general rank weights and sampling weights. Standard errors are estimated with the efficient sandwich variance estimator, ZLMB. Commands for these estimators are presented below followed by a summary.

```
R> system.time(fit.gh <- aftsrr(Surv(edrel, rel) ~ histol + age,
+    weights = hi, data = nwtco, variance = "ZLMB", subset = in.casecohort))
```

```
   user  system elapsed
 29.311   0.083  29.542
```

```
R> system.time(fit.lk <- aftsrr(Surv(edrel, rel) ~ histol + age,
+    weights = hi, data = nwtco, variance = "ZLMB", rankWeights = "logrank",
+    subset = in.casecohort))
```

```
   user  system elapsed
 35.490   0.097  35.697
```

```
R> system.time(fit.pw <- aftsrr(Surv(edrel, rel) ~ histol + age,
+    weights = hi, data = nwtco, variance = "ZLMB", rankWeights = "PW",
+    method = "monosm", subset = in.casecohort))
```

```
   user  system elapsed 
 84.716   0.299  85.453 
```

```
R> summary(fit.gh)

Call:
aftsrr(formula = Surv(edrel, rel) ~ histol + age, data = nwtco,
    subset = in.casecohort, weights = hi, variance = "ZLMB")

Variance Estimator: ZLMB
       Estimate StdErr z.value p.value    
histol   -3.133  0.164  -19.06  <2e-16 ***
age      -0.204  0.038   -5.44  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> summary(fit.lk)

Call:
aftsrr(formula = Surv(edrel, rel) ~ histol + age, data = nwtco,
    subset = in.casecohort, weights = hi, rankWeights = "logrank",
    variance = "ZLMB")

Variance Estimator: ZLMB
       Estimate StdErr z.value p.value    
histol   -3.891  0.191  -20.33  <2e-16 ***
age      -0.208  0.057   -3.62  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> summary(fit.pw)

Call:
aftsrr(formula = Surv(edrel, rel) ~ histol + age, data = nwtco,
    subset = in.casecohort, weights = hi, rankWeights = "PW",
    method = "monosm", variance = "ZLMB")

Variance Estimator: ZLMB
       Estimate StdErr z.value p.value    
histol   -3.793  0.105  -36.10  <2e-16 ***
age      -0.209  0.026   -8.14  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although the differences in standard errors among the three weights are noticeable, they all lead to the same conclusion. The p values suggest that the coefficient of central histological diagnosis is significantly different from zero and had a significant effect on the time to relapse. Compared to the full cohort analysis, all point estimates are reasonably close. This result is also found in the full cohort analysis.

6.3. Kidney catheter data

A bivariate failure time example from a kidney catheter study is used to illustrate the least squares approach feature of **aftgee**. The kidney catheter data are available in the **survival** package as `kidney` (McGilchrist and Aisbett 1991). The interest of the study is to examine the time to infection from the point of catheter insertion for patients using portable dialysis equipment. The catheter is removed when the infection occurs and reinserted after some pre-determined time. If catheters are removed for reasons other than infection, then the time to infection is treated as censored. The data contain 38 patients, each having exactly two insertions, where two observations on time to infection was recorded. The two covariates considered are patient's age (`age` in years) and gender (`sex` = 0 if male, `sex` = 1 if female).

We first fit bivariate AFT models with identical error margins and identical regression coefficients for the two margins. Since it is reasonable to expect some correlation between the two recurrence times for a given patient, we can model this in the least squares approach with some dependent working covariance structure. We will first fit the least squares approach with a working independent covariance structure and then with an exchangeable working structure. For both least squares approaches, we use the induced smoothing rank-based estimator with Gehan's weight as the initial estimator. The standard errors are estimated by the multiplier resampling method with bootstrap size 100.

```
R> data("kidney", package = "survival")
R> set.seed(123)
R> fit.ind <- aftgee(Surv(time, status) ~ age + sex, id = id, data = kidney)
R> fit.ex <- aftgee(Surv(time, status) ~ age + sex, id = id, data = kidney,
+    corstr = "ex")
R> summary(fit.ind)

Call:
aftgee(formula = Surv(time, status) ~ age + sex, data = kidney,
    id = id)

AFTGEE Estimator
            Estimate StdErr z.value p.value
(Intercept)    2.071  0.609    3.40   0.001 ***
age           -0.005  0.008   -0.64   0.523
sex            1.374  0.346    3.96  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> summary(fit.ex)

Call:
aftgee(formula = Surv(time, status) ~ age + sex, data = kidney,
    id = id, corstr = "ex")

AFTGEE Estimator
            Estimate StdErr z.value p.value
(Intercept)    2.070  0.688    3.01   0.003 **
age           -0.005  0.010   -0.54   0.589
sex            1.374  0.369    3.72  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient of `sex` is found to be significantly different from zero for both models. This suggests that female patients tend to have longer recurrence times to infection. The efficiency gain is expected on average but, unfortunately, this dataset does not show an efficiency gain.

In addition to the common marginal error distribution and common coefficient assumption, we also consider the case where the marginal error distributions and regression coefficients are different. In this case, we need to specify `margin` and construct the corresponding block diagonal design matrix. After the block diagonal design matrix is constructed, least squares estimators with both independent covariance working structure and exchangeable working structure are fitted. For each model, we use the smooth Gehan estimator as the initial value.

```
R> kidney$margin <- as.factor(rep(1:2, 38))
R> fit2.ind <- aftgee(Surv(time, status) ~ age:margin + sex:margin +
+    margin - 1, id = id, margin = margin, data = kidney)
R> fit2.ex <- aftgee(Surv(time, status) ~ age:margin + sex:margin +
+    margin - 1, id = id, margin = margin, data = kidney, corstr = "ex")
R> summary(fit2.ind)

Call:
aftgee(formula = Surv(time, status) ~ age:margin + sex:margin +
    margin - 1, data = kidney, id = id, margin = margin)

AFTGEE Estimator
            Estimate StdErr z.value p.value
margin1        1.676  0.804    2.08   0.037 *
margin2        2.542  0.881    2.89   0.004 **
age:margin1   -0.013  0.011   -1.18   0.238
age:margin2    0.005  0.013    0.41   0.679
margin1:sex    1.744  0.439    3.97  <2e-16 ***
margin2:sex    0.895  0.451    1.99   0.047 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R> summary(fit2.ex)

Call:
aftgee(formula = Surv(time, status) ~ age:margin + sex:margin +
    margin - 1, data = kidney, id = id, margin = margin, corstr = "ex")

AFTGEE Estimator
            Estimate StdErr z.value p.value
margin1        1.672  0.897    1.86   0.062 .
margin2        2.544  0.901    2.82   0.005 **
age:margin1   -0.013  0.012   -1.14   0.253
age:margin2    0.005  0.012    0.47   0.638
margin1:sex    1.744  0.467    3.73  <2e-16 ***
margin2:sex    0.887  0.473    1.88   0.060 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model allows hypothesis testing of equal coefficiencts for each covariate across the two margins from Wald-type tests with covariance matrix estimates. However, the covariates of `age` and `sex` are found to be not significantly different across the two margins, with p values of 0.87 and 0.06, respectively, under the exchangeable structure. The coefficients of `sex` for both margins and the two working structures are found to be significantly different from zero. These results coincide with those from the common margin model. The efficiency gain from the exchangeable structure in the margin-specific case is also absent. This is probably because there is not much strength to borrow with distinctive marginal fits.

7. Conclusion

Package **aftgee** provides an easy access to fitting semiparametric AFT models for possibly clustered failure times with both rank-based approaches and the least squares approach. For rank-based approaches, we implemented the induced smoothing method with Gehan's weight and extended it to allow arbitrary rank weight. The method is much faster than those based on linear programming. Computationally efficient sandwich variance estimators are provided for all the estimators, and additional sampling weight can be incorporated for various sampling schemes. Our least squares approach uses rank-based estimators as initial estimators in an iterative estimation procedure. For clustered data, we exploited within-cluster dependence through working correlation structure in a GEE framework which enhances efficiency when within-cluster dependence is strong. The implementation is fast and reliable, making it possible for AFT models to be much more widely applied in routine survival analysis.

Our package can be expanded in several directions. The current version allows weights for handling missing data in the rank-based approach, similar weights can also be made available to our GEE approach. For the rank-based approach with clustered data, Wang and Fu (2011) considered estimating equations that can be decomposed into between- and within-cluster estimating equations for better efficiency. An implementation of this method would be desirable. To account for measurement errors in covariates, package **simexaft** (He, Xiong, and Yi 2012) implemented a simulation-extrapolation approach for AFT models. Such an approach can be extended to the semiparametric AFT model. Furthermore, our methods can also be extended to accommodate survival data other than with right censoring.

Acknowledgments

This research was partially supported by NSF DMS grant 120922.

References

Brostöm G (2014). ***eha**: Event History Analysis.* R package version 2.4-1, URL http://CRAN.R-project.org/package=eha.

Brown BM, Wang YG (2005). "Standard Errors and Covariance Matrices for Smoothed Rank Estimators." *Biometrika*, **92**(1), 149–158.

Brown BM, Wang YG (2007). "Induced Smoothing for Rank Regression with Censored Survival Times." *Statistics in Medicine*, **26**(4), 828–836.

Buckley J, James I (1979). "Linear Regression with Censored Data." *Biometrika*, **66**(3), 429–436.

Chiou SH, Kang S, Kim J, Yan J (2014a). "Marginal Semiparametric Multivariate Accelerated Failure Time Model with Generalized Estimating Equations." *Lifetime Data Analysis*, **20**(4), 599–618.

Chiou SH, Kang S, Yan J (2013). "Rank-Based Estimating Equations with General Weight for Accelerated Failure Time Models: An Induced Smoothing Approach." *Technical Report 39*, Department of Statistics, University of Connecticut.

Chiou SH, Kang S, Yan J (2014b). "Fast Accelerated Failure Time Modeling for Case-Cohort Data." *Statistics and Computing*, **24**(4), 559–568.

Chiou SH, Kang S, Yan J (2014c). ***aftgee**: Accelerated Failure Time Model with Generalized Estimating Equations.* R package version 1.0-0, URL http://CRAN.R-project.org/package=aftgee.

Chiou SH, Kang S, Yan J (2014d). "Semiparametric Accelerated Failure Time Modeling for Clustered Failure Times from Stratified Sampling." *Technical report*. doi:10.1080/01621459.2014.917978. Forthcoming.

Cox DR (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society B*, **34**(2), 187–220.

D'Angio GJ, Breslow N, Beckwith JB, Evans A, Baum E, Delorimier A, Fernbach D, Hrabovsky E, Jones B, Kelalis P, Othersen HB, Tefft M, Thomas PRM (1989). "Treatment of Wilms' Tumor. Results of the Third National Wilms' Tumor Study." *Cancer*, **64**(2), 349–360.

Gehan EA (1965). "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples." *Biometrika*, **52**(1/2), 203–223.

Green DM, Breslow NE, Beckwith JB, Finklestein JZ, Grundy PE, Thomas PR, Kim T, Shochat SJ, Haase GM, Ritchey ML, Kelalis PP, D'Angio GJ (1998). "Comparison between Single-Dose and Divided-Dose Administration of Dactinomycin and Doxorubicin for Patients with Wilms' Tumor: A Report from the National Wilms' Tumor Study Group." *Journal of Clinical Oncology*, **16**(1), 237–245.

Halekoh U, Højsgaard S, Yan J (2006). "The R Package geepack for Generalized Estimating Equations." *Journal of Statistical Software*, **15**(2), 1–11. URL http://www.jstatsoft.org/v15/i02/.

Harrel Jr FE (2014). ***rms**: Regression Modeling Strategies.* R package version 4.2-1, URL http://CRAN.R-project.org/package=rms.

Harrington DP, Fleming TR (1982). "A Class of Rank Test Procedures for Censored Survival Data." *Biometrika*, **69**(3), 133–143.

He W, Xiong J, Yi GY (2012). "SIMEX R Package for Accelerated Failure Time Models with Covariate Measurement Error." *Journal of Statistical Software, Code Snippets*, **46**(1), 1–14. URL http://www.jstatsoft.org/v46/c01/.

Højsgaard S, Halekoh U, Yan J (2014). ***geepack**: Generalized Estimating Equation Package.* R package version 1.2-0, URL http://CRAN.R-project.org/package=geepack.

Huang Y (2002). "Calibration Regression of Censored Lifetime Medical Cost." *Journal of the American Statistical Association*, **97**(457), 318–327.

Jin Z, Huang L (2007). "**lss**: An S-PLUS/R Program for the Accelerated Failure Time Model to Right Censored Data Based on Least-Squares Principle." *Computer Methods and Programs in Biomedicine*, **86**(1), 45–50.

Jin Z, Lin DY, Wei LJ, Ying Z (2003). "Rank-Based Inference for the Accelerated Failure Time Model." *Biometrika*, **90**(2), 341–353.

Jin Z, Lin DY, Wei LJ, Ying Z (2006a). "Rank Regression Analysis of Multivariate Falure Time Data Based on Marginal Linear Models." *Scandinavian Journal of Statistics*, **33**(1), 1–23.

Jin Z, Lin DY, Ying Z (2006b). "On Least-Squares Regression with Censored Data." *Biometrika*, **93**(1), 147–161.

Jin Z, Lin DY, Ying Z (2006c). "Rank Regression Analysis of Multivariate Failure Time Data Based on Marginal Linear Models." *Scandinavian Journal of Statistics*, **33**(1), 1–23.

Johnson LM, Strawderman RL (2009). "Induced Smoothing for the Semiparametric Accelerated Failure Time Model: Asymptotics and Extensions to Clustered Data." *Biometrika*, **96**(3), 577–590.

Kalbfleisch JD, Prentice RL (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Lai TL, Ying Z (1991). "Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data." *The Annals of Statistics*, **19**(3), 1370–1402.

McGilchrist CA, Aisbett CW (1991). "Regression with Frailty in Survival Analysis." *Biometrics*, **47**(2), 461–466.

Prentice RL (1978). "Linear Rank Tests with Right Censored Data." *Biometrika*, **65**(1), 167–180.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Ritov Y (1990). "Estimation in a Linear Regression Model with Censored Data." *The Annals of Statistics*, **18**(1), 303–328.

Therneau T (2014). ***survival**: A Package for Survival Analysis in S*. R package version 2.37-7, URL `http://CRAN.R-project.org/package=survival`.

Tsiatis AA (1990). "Estimating Regression Parameters Using Linear Rank Tests for Censored Data." *The Annals of Statistics*, **18**(1), 354–372.

Varadhan R, Gilbert P (2009). "**BB**: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function." *Journal of Statistical Software*, **32**(4), 1–26. URL `http://www.jstatsoft.org/v32/i04/`.

Wang YG, Fu L (2011). "Rank Regression for Accelerated Failure Time Model with Clustered and Censored Data." *Computational Statistics & Data Analysis*, **55**(7), 2334–2343.

Wei LJ (1992). "The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis." *Statistics in Medicine*, **11**(14–15), 1871–1879.

Ying Z (1993). "A Large Sample Study of Rank Estimation for Censored Regression Data." *The Annals of Statistics*, **21**(1), 76–99.

Zeng D, Lin DY (2008). "Efficient Resampling Methods for Nonsmooth Estimating Functions." *Biostatistics*, **9**(2), 355–363.

Affiliation:

Sy Han Chiou
Department of Mathematics and Statistics
University of Minnesota, Duluth
1117 University Drive,
Duluth, MN 55812-3000, United States of America
Telephone: 218/726-7032
Fax: 218/726-8399
E-mail: `schiou@d.umn.edu`

Sangwook Kang
Department of Applied Statistics
Yonsei University
50 Yonsei Road
Seodaemun-Gu, Seoul 120-749, Korea
E-mail: kanggi1@yonsei.ac.kr

Jun Yan
Department of Statistics
University of Connecticut
215 Glenbrook Road U-4120
Storrs, CT 06279-4120, United States of America
Telephone: 860/486-3414
Fax: 860/486-4113
E-mail: jun.yan@uconn.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 61, Issue 11 *Submitted:* 2012-11-15
November 2014 *Accepted:* 2014-06-13