Biostatistics (2016), **17**, 3, *pp.* 576–588
doi:10.1093/biostatistics/kxw008
Advance Access publication on March 3, 2016

Accelerated failure time model under general biased sampling scheme

JANE PAIK KIM

Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA

TONY SIT*

Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR
tonysit@sta.cuhk.edu.hk

ZHILIANG YING

Department of Statistics, Columbia University, New York, NY 10027, USA

Summary

Right-censored time-to-event data are sometimes observed from a (sub)cohort of patients whose survival times can be subject to outcome-dependent sampling schemes. In this paper, we propose a unified estimation method for semiparametric accelerated failure time models under general biased estimating schemes. The proposed estimator of the regression covariates is developed upon a bias-offsetting weighting scheme and is proved to be consistent and asymptotically normally distributed. Large sample properties for the estimator are also derived. Using rank-based monotone estimating functions for the regression parameters, we find that the estimating equations can be easily solved via convex optimization. The methods are confirmed through simulations and illustrated by application to real datasets on various sampling schemes including length-bias sampling, the case–cohort design and its variants.

Keywords: Accelerated failure time model; Case–cohort design; Counting process; Estimating equations; Importance sampling; Length-bias; Regression; Survival data.

1. Introduction

Length/size-biased sampling has been recognized in statistics for decades in the studies of ecology (McFadden, 1962; Muttlak and McDonald, 1990), fiber length (Cox, 1969), and economic duration data (Kiefer, 1988). This kind of biased sampling arises when a positive-valued outcome variable is sampled with selection probability proportional to its size/survival time. The one-sample problem of estimating the survivor function has been studied in Vardi (1982, 1985). Further results can be found in Asgharian *and others* (2002) and Asgharian and Wolfson (2005).

Under the scope of semiparametric inference, Tsai (2009) carefully studied two intrinsically different types of length-biased sampling, depending on the assumption made on the recruitment/censoring

*To whom correspondence should be addressed.

procedure under the Cox proportional hazards model. Shen *and others* (2012) proposed two likelihood approaches for the estimation and assessment of the difference between two survival distributions under length-biased sampling. While there has been extensive effort to develop survival models in the context of length-biased sampling, e.g. proportional hazards and linear transformation models, less attention has been paid to the accelerated failure time (AFT) model. More recently, Chen (2010) and Shen *and others* (2009) specifically studied the AFT model for data that are length/size-biased.

Another outcome sampling scheme is the case–cohort design which has first proposed by Prentice (1986). Such a design is developed in order to provide a more cost-efficient sampling method, especially in large-scale (epidemiological) studies on rare diseases with certain covariate information that is difficult or costly to obtain. Recent works that study inference on survival data from the case–cohort design include Lu and Tsiatis (2006), Nan *and others* (2009), and Kong and Cai (2009); the first of these developed inference procedure on a general class of semiparametric transformation models, while the latter two proposed solutions on inferring regression parameters of the AFT model.

The AFT model, as a common alternative to linear transformation models, relates the logarithm of the failure time linearly to the covariates; see Tsiatis (1990), Wei *and others* (1990), Kalbfleisch and Prentice (2002), and Cox and Oakes (1984) among others. Under the AFT model, the effect of the covariates on the failure time is directly related to the acceleration or deceleration of time to failure. This feature facilitates an easy interpretation for clinicians and hence the model provides an alternative to the celebrated (Cox, 1972) proportional hazards model for the regression analysis of censored failure time data.

Several estimation and inferential procedures have been proposed for the estimation of the regression parameters, including rank-based estimating equations and martingale estimating equations for unbiased data; see Prentice (1978) and Buckley and James (1979). The large-sample properties of the Buckley–James-type and rank estimators were then studied by Ritov (1990), Tsiatis (1990), Ying (1993), and Jin *and others* (2003). Furthermore, Zeng and Lin (2007) proposed a kernel-smoothed profile likelihood function for the AFT model.

Despite the theoretical advances, semiparametric methods for the AFT model have rarely been developed to deal with various biased sampling schemes, either due to natural setting or trial design, under one general framework. This paper aims at proposing a unified approach that can handle many commonly encountered biased sampling schemes such as length-biased sampling, the case–cohort design as well as its variants. We show that our approach leads to estimators that are consistent and asymptotically normal and we provide consistent variance estimators. The one-size-fit-all solution introduced in this paper should benefit a wide range of practitioners who have to tackle problems on biased samples.

The approach presented in this paper is related to that of Kim *and others* (2013), which deals with biased sampling under the semiparametric linear transformation models. However, it is by no means straightforward to apply their framework directly to the AFT model setting. This is due to the fact that the bias involved in sampling is usually characterized via the event time, whereas inference for the regression parameter of the AFT model is carried out through the residual term. Furthermore, it has still been unclear about the feasibility of extending the existing rank-based inference procedure for censored data (Jin *and others*, 2003, 2006) including finding equivalent objective functions and resampling schemes for variance estimation to datasets with sampling bias. Our main contribution in this paper is to provide a careful treatment that extends (Jin *and others*, 2003) methodology to a biased sampling setting so that the corresponding estimator is still numerically tractable, while its standard error can also be estimated via a resampling scheme as discussed in Jin *and others* (2006). Ignoring the sampling bias or uniqueness of the point estimates may result in substantial bias in estimating the survival time distribution and hence potentially fallacious inference.

The rest of the article is organized as follows: Section 2 specifies the model setup, necessary notation as well as the estimating procedures. The asymptotic properties of the proposed estimator are studied. In particular, we presented a Gehan-type estimating equation whose root can be obtained via convex optimization. We present a resampling procedure for variance estimation. Solution to a more general weighting scheme will also be elaborated. Simulation studies under practical sample sizes and real data analyses are conducted to assess the performance of the proposed estimator; the corresponding discussion is included in Section 3. Section 4 concludes the paper. In addition to specific simulation settings and results, all the technical details are presented in supplementary materials (available at *Biostatistics* online).

2. Proposed methods

We shall assume throughout this paper that there are two underlying random variables T^* and C^* that correspond to unbiased time to failure and time to censoring as seen in typical right-censored data problems. Only $\tilde{T}^*$, the minimum between T^* and C^*, is observed. The AFT model relates the logarithm of the failure time linearly to the concomitant covariates, say $Z = (Z_1, \ldots, Z_p)$ in the following sense:

$$\log T^* = -\beta_0' Z + \epsilon, \tag{2.1}$$

where β_0 is a p-vector of regression coefficients and ϵ, in contrast to the transformation models, has an unknown distribution. The parameter β_0 appears to be easy to interpret because it directly relates to the level of $\log T^*$. The primary goal of this paper is to find semiparametric estimates of β_0, denoted by $\hat{\beta}$, under a generalized biased sampling scheme. The data observed will consist of n biased iid random vectors $\{(\tilde{T}_i, \Delta_i, Z_i)\}_{i=1,\ldots,n}$, where $\tilde{T}_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leqslant C_i)$. Following Kim *and others* (2013), we introduce $q_Z(t, \delta)$, $t \geqslant 0$, $\delta \in \{0, 1\}$, the joint conditional density of $(\tilde{T}^*, \Delta^*)$ given the covariates Z, where $\Delta^* = I(T^* \leqslant C^*)$. Since T^* and C^* are assumed to be conditionally independent given Z, we can write

$$q_Z(t, \delta) = \left\{ f_Z(t) \bar{G}_Z(t) \right\}^{\delta} \left\{ g_Z(t) \bar{F}_Z(t) \right\}^{1-\delta}, t \geqslant 0, \delta \in \{0, 1\},$$

where f_Z ($\bar{F}_Z$) and g_Z ($\bar{G}_Z$) denote the conditional density (survival) functions of T^* and C^*, respectively. Under a biased sampling scheme, i.e. the data collected are subject to a certain type of biased sampling, which is characterized by a biasing function $w(t, \delta)$, $t \geqslant 0$, $\delta \in \{0, 1\}$, the conditional joint density of $(\tilde{T}, \Delta)$ given Z then changes to

$$\tilde{q}_Z(t, \delta) = \frac{w(t, \delta) q_Z(t, \delta)}{\int w(s, 0) q_Z(s, 0)\, \mathrm{d}s + \int w(s, 1) q_Z(s, 1)\, \mathrm{d}s}. \tag{2.2}$$

Typical examples of biasing functions include: (a) length-biased sampling, where $w(t, \delta) = t$ and (b) case–cohort sampling, where $w(t, \delta) = \delta + (1 - \delta)p$, $p \in (0, 1)$ is a predefined constant. Readers should be noted that the biasing function can be generalized so that it depends on $\tilde{T}$, Δ, or both.

2.1 *Notation and background*

For our purposes, it will be convenient to consider the counting process approach; see Gill (1980). Here, and in the sequel, we shall use the notation $N(t) = \Delta I(\tilde{T} \leqslant t)$ to represent the counting process which jumps by one when a failure occurs. Hazard and cumulative hazard functions of T are denoted by $\lambda(\cdot)$ and $\Lambda(\cdot)$, respectively. Let $Y(t) = I(\tilde{T} \geqslant t)$ denote the at-risk indicator. Moreover, for the AFT model, it is helpful to re-express the counting process via the residual term. Specifically, we define $e_i(\beta) = \log \tilde{T}_i + \beta' Z_i$, $N_i(\beta; t) = \Delta_i I\{e_i(\beta) \leqslant t\}$, and $Y_i(\beta; t) = I\{e_i(\beta) \geqslant t\}$ for $i = 1, 2, \ldots, n$. Unless there exists ambiguity, the subscript Z that appears in q and $\tilde{q}$ will be suppressed.

In full cohort design, the regression parameters can be estimated via the weighted log-rank estimating function (Ying, 1993; Jin *and others*, 2003):

$$U_\phi(\beta) = \sum_{i=1}^{n} \Delta_i \phi\{\beta; e_i(\beta)\}[Z_i - \bar{Z}\{\beta, e_i(\beta)\}] = \sum_{i=1}^{n} \int \phi(\beta; u)\{Z_i - \bar{Z}(\beta, u)\}\,\mathrm{d}N_i(\beta; u) = 0,$$

where ϕ is a (data-dependent) weight function, $\bar{Z}(\beta, u) := S^{(1)}(\beta, u)/S^{(0)}(\beta, u)$ with $S^{(\kappa)}(\beta, u) = (1/n)\sum_{j=1}^{n} Y_j(\beta; u)Z_j^{\kappa}, \kappa = 0, 1$; $Z^0 = 1$, and $Z^1 = Z$. The choices of $\phi(\beta, u) = 1$ and $\phi(\beta, u) = S^{(0)}$ correspond to log-rank and Gehan statistics, respectively.

One of the technical issues that needs to be tackled for the above root-finding problem is that the estimating equation $U_\phi(\beta)$ is not necessarily component-wise monotone in β. Non-differentiability and non-monotonicity involved with respect to β also lead to the potential of having multiple solutions to the equation, amongst which some of them are inconsistent. Such a problem is even more prominent when β is of higher dimension. In the following subsections, apart from introducing the estimation procedure for β, we shall also investigate the numerical issues involved in this root-finding procedure.

2.2 *Estimation procedure*

Biased sampling appears in many applications, either naturally or by design. To the best of our knowledge, there has not yet been any unified inference procedure for the regression parameter in the AFT model (2.1) under general biased sampling schemes. A key element in the counting process approach is the use of $E\{\mathrm{d}N(t) - Y(t)\,\mathrm{d}\Lambda(t)\} = 0$ as the basis for constructing an unbiased estimating equation. Under the biased sampling setting (2.2), however, the above quantity is no longer a zero-mean process. In fact, Kim *and others* (2013) derived the following lemma that shows that the compensator of the counting process $N(t)$ requires an extra adjustment term to achieve the mean-zero property.

LEMMA 2.1 Under the biased sampling scheme, i.e. $(\tilde{T}, \Delta)$ follows $\tilde{q}_Z$ given by (2.2) and $\omega(t, \tilde{T}, \Delta, \beta)$, we have

$$E_Z\{\mathrm{d}N(t)\} = E_Z\{\omega(t, \tilde{T}, \Delta, \beta_0)Y(t)\lambda(t)\,\mathrm{d}t\}, \tag{2.3}$$

where E_Z denotes the conditional expectation given $\boldsymbol{Z}$ and $\omega(t, \tilde{T}, \Delta) = \{q(\tilde{T}, \Delta)/\tilde{q}(\tilde{T}, \Delta)\}\{\tilde{q}(t, 1)/q(t, 1)\}$.

The inclusion of the adjustment term $\omega(t, \tilde{T}, \Delta, \beta)$ in (2.3) can be regarded as the Radon–Nikodym derivative between the true and the biased densities for the risk set and the counting process, respectively. Due to the fact that both the counting process and the risk set are observed under a biased probability measure, whereas the hazard function is expressed with respect to the true density, we have to adjust both $\mathrm{d}N(t)$ and the at-risk indicator $Y(t)$ so that all the components in estimating equation (2.3) are evaluated under the same measure. The proof of this lemma is presented in Kim *and others* (2013).

Due to (2.3), we can obtain the following set of estimating equations:

$$\sum_{i=1}^{n} \phi(t; \beta)\{\mathrm{d}N_i(t) - \omega(t, \tilde{T}_i, \Delta_i; \beta)Y_i(t)\,\mathrm{d}\Lambda(t)\} = 0, \tag{2.4}$$

$$\sum_{i=1}^{n} \int_0^{\tau} Z_i\phi(t; \beta)\{\mathrm{d}N_i(t) - \omega(t, \tilde{T}_i, \Delta_i; \beta)Y_i(t)\,\mathrm{d}\Lambda(t)\} = 0, \tag{2.5}$$

where τ is constant such that $P(\tilde{T} \geqslant \tau) > 0$. By choosing $\phi(\cdot) = 1$ for (2.4), we obtain $\mathrm{d}\hat{\Lambda}(t) = \sum_{i=1}^{n} \mathrm{d}N_i(t) / \sum_{i=1}^{n} \omega(t, \tilde{T}_i, \Delta_i; \beta) Y_i(t)$. It follows that if we consider a transformation on the time scale: $t \mapsto te^{-\beta' Z}$, then (2.5) can be written as

$$\sum_{i=1}^{n} \int_{0}^{\tau \exp(\beta' Z_i)} \phi(te^{-\beta' Z_i}; \beta) \left\{ Z_i - \frac{\sum_{j=1}^{n} Z_j \omega(te^{-\beta' Z_j}, \tilde{T}_j, \Delta_j; \beta) Y_j(te^{-\beta' Z_j})}{\sum_{j=1}^{n} \omega(te^{-\beta' Z_j}, \tilde{T}_j, \Delta_j; \beta) Y_j(te^{-\beta' Z_j})} \right\} \mathrm{d}N_i(te^{-\beta' Z_i}) = 0,$$

or equivalently,

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Delta_i \phi(\tilde{T}_i; \beta) \left[Z_i - \frac{\sum_{j=1}^{n} Z_j \omega\{\tilde{T}_i e^{\beta'(Z_i - Z_j)}, \tilde{T}_i, \Delta_i, \beta\} Y_j\{\tilde{T}_i e^{\beta'(Z_i - Z_j)}\}}{\sum_{j=1}^{n} \omega\{\tilde{T}_i e^{\beta'(Z_i - Z_j)}, \tilde{T}_i, \Delta_i, \beta\} Y_j\{\tilde{T}_i e^{\beta'(Z_i - Z_j)}\}} \right] \\ &= \sum_{i=1}^{n} \Delta_i \phi(\tilde{T}_i; \beta) \{Z_i - \bar{Z}_\omega(\beta)\}, \quad \text{say.} \end{aligned} \tag{2.6}$$

The unbiasedness property exhibited in (2.6) is important in obtaining an asymptotically unbiased estimator for β_0. It is noteworthy that since the weight function ω, which may contain $\tilde{T}$ and/or Δ, is not necessarily $\mathcal{F}_t$-measurable, the process $M(t) = N(t) - \int_{-\infty}^{t} \omega(u, \tilde{T}_i, \Delta_i, \beta) Y_i(u)\, \mathrm{d}\Lambda(u)$ is not a martingale but a mean-zero process instead.

Before discussing more general weight functions $\omega(\cdot)$ in Section 2.4, we first study the special case of $\phi(te^{-\beta' Z_i}; \beta) = \sum_{j=1}^{n} \omega(te^{-\beta' Z_j}, \tilde{T}_j, \Delta_j, \beta) Y_j(te^{-\beta' Z_j})$, which is regarded as the Gehan-type weight function. Such a formulation can considerably simplify (2.6) and lead to

$$U_G(\beta) := \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (Z_i - Z_j) \omega\{\tilde{T}_i e^{\beta'(Z_i - Z_j)}, \tilde{T}_i, \Delta_i, \beta\} Y_j\{\tilde{T}_i e^{\beta'(Z_i - Z_j)}\} = 0. \tag{2.7}$$

There are a number of typical biasing sampling designs that can be covered under the purposed generalized framework. Common examples include (i) length-biased sampling, (ii) case–cohort design, (iii) case–cohort sampling on a length-biased sampling, and (iv) generalized case–cohort design. Readers are referred to Kim *and others* (2013) for a more elaborated description of the weight function $\omega(\cdot)$ formulation. We shall elaborate the subtlety of the proposed method and its difference between that discussed in Kim *and others* (2013) via two examples.

Length-biased sampling. In the length-biased sampling case, since the density of $(\tilde{T}, \Delta)$ is proportional to $tq(t, \delta)$, the weight function is given by

$$\omega(t, \tilde{T}, \Delta) = \frac{q(\tilde{T}, \Delta)\tilde{q}(t, 1)}{\tilde{q}(\tilde{T}, \Delta) q(t, 1)} = \frac{t}{\tilde{T}}.$$

Due to the fact that the time horizon considered in (2.7) is viewed from the perspective of η, the time variable t expressed in the above weight function should be adjusted to $te^{-\beta' Z_i}$ for subject i $(i = 1, \ldots, n)$, instead. With specifically $\omega(t; \beta) = te^{-\beta' Z} / \tilde{T}$ and hence $\phi(te^{-\beta' Z_i}; \beta) = \sum_{j=1}^{n} te^{-\beta' Z_j} Y_j(te^{-\beta' Z_j}) / \tilde{T}_j$, we can rewrite (2.7) into the following estimating equation that corresponds to the length-biased sampling setting:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (Z_i - Z_j) \left\{ \frac{\tilde{T}_i}{\tilde{T}_j} e^{\beta'(Z_i - Z_j)} \right\} I \left\{ \frac{\tilde{T}_i}{\tilde{T}_j} e^{\beta'(Z_i - Z_j)} - 1 \leqslant 0 \right\} = 0. \tag{2.8}$$

We cannot, however, directly adopt the approach of Jin *and others* (2003) in solving (2.8) via linear programming due to the extra term $\tilde{T}_i / \tilde{T}_j e^{\beta'(Z_i - Z_j)}$ involved for the bias correction.

It can be shown in the supplementary material (available at *Biostatistics* online) that the root of (2.8) is still a solution of a monotone estimating equation. In fact, the root-finding problem presented in (2.8) can be translated into an optimization problem of a convex function:

$$L_G(\beta) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \frac{\tilde{T}_i e^{\beta'(Z_i - Z_j)}}{\tilde{T}_j}\{e_i(\beta) - e_j(\beta)\}^- = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \omega(\tilde{T}_i e^{\beta' Z_i}; \beta)\{e_i(\beta) - e_j(\beta)\}^-,$$

where $a^- = |a|I(a < 0)$. The minimizer of $L_G(\beta)$ is denoted by $\hat{\beta}_G$. Note that although $\hat{\beta}_G$ is not necessarily unique, all these minimizers are asymptotically equivalent.

We would like to emphasize that our estimation procedure is different from that proposed in Chen (2010). In particular, Chen (2010) only considered the uncensored case and made use of the invariance principle (see Property 1 of Chen, 2010, p. 150) to provide an inference procedure for the regression covariates. Shen *and others* (2009) proposed the estimating equation $\sum_{i=1}^{n} q(Z_i)\Delta_i Z_i(\log \tilde{T}_i - \beta' Z_i)/\hat{w}(\tilde{T}_i) = 0$, where $\hat{w}(\tilde{T}_i)$ is a consistent estimator of $w(\tilde{T}_i) = \int_0^{\tilde{T}_i} \hat{G}(u)\,\mathrm{d}u$ with $\hat{G}$ the Kaplan–Meier estimator for the censoring variable C. For the cases where censoring times are dependent on a large number of covariates, the estimation of $\hat{G}(t)$ may not be stable, so will be the estimates of the regression parameters.

While the weight function introduced for bias-adjustment resembles the risk set resampling proposed in Wang (1996), our methodology is developed upon an estimating equation derived from the joint density of $(\tilde{T}, \Delta)$ and hence the problem of censoring, which is not covered under the pseudo-likelihood framework as considered in Wang (1996), can be properly tackled.

As an additional remark, the current setup is designed for handling censoring first and followed by length-biased sampling. In fact, this setting coincides with the first type of length-biased sampling considered in Tsai (2009). Adopting the notation used in Tsai (2009), we let $h_2(a, t, \delta \mid Z)$ be the conditional probability density function of the observed data (A, T, Δ), where A denotes the truncation time. Under the assumption that A follows a uniform distribution marginally, one can write $\tilde{q}_Z(t, \delta) \propto \int_0^t h_2(a, t, \delta \mid Z)\,\mathrm{d}a$. This formulation occurs naturally when a cross-sectional sampling (censoring) is performed in which the probability for a sample to be selected is proportional to the follow-up period $\tilde{T}$ instead of the event time T. Under such a model assumption, subjects with short follow-up times, which may be the result of dropouts or death due to other unrelated causes, will be selected with lower probabilities. For the second type of censoring discussed in Tsai (2009), we shall introduce the corresponding treatment in Section 3.2.

Classical case–cohort sampling and its variants. For those weight functions that do not involve time, it can be easily proved that the estimating equations for these cases are also monotonically non-decreasing as there is no time-transformation involved in the bias-adjustment term $\omega(t; \beta)$. Similar estimation procedure as shown above can be applied to obtain $\hat{\beta}_G$. In particular, for the traditional case–cohort biased sampling setting, the corresponding $L_G(\beta)$ is given by

$$L_G(\beta) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\Delta_i}{\Delta_i + (1 - \Delta_i)p}\{e_i(\beta) - e_j(\beta)\}^- = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\Delta_i \omega(\tilde{T}_i e^{\beta' Z_i}; \beta)\{e_i(\beta) - e_j(\beta)\}^-,$$

where p denotes the sampling probability amongst the censored subjects. Furthermore, in our model, $(\tilde{T}_i, \Delta_i)_{i=1,\ldots,n}$ refer to the samples selected in the subcohort. For case–cohort data on the AFT model, Kong and Cai (2009) also proposed an estimation procedure based on the convergence result of Wei *and others* (1990). Their method covers the case of sampling without replacement, which is slightly different from the current setting. It is noteworthy that Nan *and others* (2006) and Nan *and others* (2009) view case–cohort sampling from the missing data perspective. Indeed, one can actually derive (2.3) by considering $\Pr\{\tilde{T} \in (t \pm \delta), \Delta \mid D = 1, Z\}$, where D denotes the indicator of observing an individual or not as demonstrated in Kim *and others* (2013). While our formulation coincides with the construction of Nan *and others* (2009) with a non-predictable weight function for the classical case–cohort sampling, such a generalization enables us to handle variants of case–cohort sampling, including stratified case–cohort sampling, generalized case–cohort sampling, and case–cohort sampling on length-biased samples; see Section 3.1 and Kim *and others* (2013) for further details. Neither of the previously mentioned last two settings has been studied in the existing literature to the best of our knowledge.

Before concluding this subsection, the authors would like to emphasize that the translation from Kim *and others* (2013) to the current model setting is not as direct as it appears to be. In addition to the fine treatment for the time variable t due to a different counting process considered, the proposed method, under a relatively high-dimensional setting, requires extra attention for cases like left truncation. Recall that for the case where there is left truncation for the sampled data, the weight function needed for bias correction is given by $\omega(t, \tilde{T}, \Delta) = I(U \leqslant t)$, where U is the truncation threshold. In order to be selected into the sample, subjects must survive until at least U. Correspondingly, the estimating equation, under the AFT model setting, will become

$$0 = \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i (Z_i - Z_j) I\{U \leqslant \tilde{T}_i e^{-\beta'(Z_i - Z_j)}\} Y_j\{\tilde{T}_i e^{\beta'(Z_i - Z_j)}\}.$$

It can be shown that the right-hand side of the above equation is not a monotonic non-decreasing field. There can be multiple inconsistent roots, especially for relatively high-dimensional Z's. For other sampling procedures that introduce biases, which include case–cohort, stratified/generalized and combo of case–cohort with length-biased sampling, we can still obtain a monotone estimating equation. Once we have such a desirable numerical property, the root-finding procedure can be carried out via convex optimization.

2.3 *Variance estimation*

The random vector $n^{1/2}(\hat{\beta}_G - \beta_0)$ is asymptotically mean zero with covariance matrix of the form $A_G^{-1} B_G A_G^{-1}$, where A_G and B_G are functions of $\lambda_\epsilon(\cdot)$, the unknown hazard of the residual:

$$A_G := \lim_{n\to\infty} A_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int \{Z_i - \bar{Z}_\omega(u, \beta_0)\}^{\otimes 2} \{\dot{\lambda}_\epsilon(u)/\lambda_\epsilon(u)\}\, \mathrm{d}N_i(\beta; u),$$

$$B_G := \lim_{n\to\infty} B_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \int \{Z_i - \bar{Z}_\omega(u, \beta_0)\}^{\otimes 2}\, \mathrm{d}N_i(\beta; u),$$

where $\dot{\lambda}_\epsilon(u) = \mathrm{d}\lambda_\epsilon(u)/\mathrm{d}u$. It is therefore challenging to obtain an analytical expression for the covariance matrix. In this subsection, we will adopt a resampling scheme discussed in Jin *and others* (2006). Since the perturbation method proposed in Jin *and others* (2003) is developed upon a martingale structure, the weight function $\omega(\cdot)$ that adjusts the bias makes the direct adaptation of Jin *and others* (2003) not possible.

We, therefore, define a loss function

$$U_G^*(\beta) = \frac{1}{n} \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i (Z_i - Z_j)\, \omega(\tilde{T}_i e^{\beta' Z_i}, \tilde{T}_i, \Delta_i; \beta)\, I\{e_i(\beta) \leqslant e_j(\beta)\} \xi_i \xi_j,$$

where $(\xi_1, \ldots, \xi_n)$ are independent positive random variables with $E(\xi_i) = \mathrm{Var}(\xi_i) = 1$, $i = 1, \ldots, n$. Denote by $\hat{\beta}_G^*$ a root that solves $U_G^* = 0$, which can be obtained from the convex optimization discussed in Section 2.2. Note that U_G^* has the same mean and approximately the same variance as its unperturbed counterpart U_G as defined in (2.7). It can be shown that the asymptotic distribution of $n^{1/n}(\hat{\beta}_G - \beta_0)$ can be approximated by the conditional distribution of $n^{1/2}(\hat{\beta}_G^* - \hat{\beta}_G)$ given the data $(\tilde{T}_i, \Delta_i, Z_i)$. In order to approximate the distribution of β_G, a large number of realizations of $\hat{\beta}_G^*$ are generated by randomizing the perturbation $(\xi_1, \xi_2, \ldots, \xi_n)$ while holding the data at their observed values. The asymptotic variance of $\hat{\beta}_G$ can then be estimated by the empirical covariance matrix of $\hat{\beta}_G^*$, enabling construction of Wald-type confidence intervals for $\hat{\beta}_G$. This procedure is applicable to cases in which there are high censoring rates and hence is useful for the biased sampling schemes that our framework can cover.

2.4 *General weight functions*

Upon the basis of $\hat{\beta}_G$, we introduce the following modification that allows the construction of weighted Gehan-type loss functions. It is a modified version of Jin *and others* (2003) with the variable t due to the bias-adjusting weight function which may involve time. The most popular weight functions, for example log-rank, Prentice–Wilcoxon and G^p class of Harrington and Fleming (1982), are special cases of this framework. Let $\psi(t; b) = \phi(t; b)/S_\omega^{(0)}(t; b)$; one can write

$$\begin{aligned}\tilde{U}_\phi(\beta; \hat{\beta}) &= \sum_{i=1}^{n} \Delta_i \psi\{\tilde{T}_i e^{\beta' Z_i} \times \mathrm{e}^{(\hat{\beta}-\beta)' Z_i}; \hat{\beta}\} S_\omega^{(0)}\{\tilde{T}_i e^{\beta' Z_i}; \beta\}\{Z_i - \bar{Z}_\omega(\beta)\} \\ &= \sum_{i=1}^{n} \Delta_i \psi(\tilde{T}_i e^{\hat{\beta}' Z_i}; \hat{\beta}) S_\omega^{(0)}(\tilde{T}_i e^{\beta' Z_i}; \beta)\{Z_i - \bar{Z}_\omega(\beta)\},\end{aligned}$$

where $S_\omega^{(k)}(te^{-b'Z_i}; b) = \sum_{j=1}^n Z_j^\kappa \omega(te^{-b'Z_j}, \tilde{T}_j, \Delta_j; b) Y_j(te^{-b'Z_j})$, $\kappa = 0, 1$. The resulting $\tilde{U}_\phi(\beta; \hat{\beta})$ is similar to that of (2.7) except for the additional weights ϕ introduced. This term is, however, free of β. Hence, $\tilde{U}_\phi(\beta; \hat{\beta})$ is still monotone in each component of β and the root-finding procedure can be carried out in a similar fashion as discussed in Section 2.2. The estimate $\hat{\beta}$ has to be estimated iteratively with $\hat{\beta}_G$ as the initial value. Numerical results suggest that the number of iterations can be as small as three in order to achieve the convergence; see Jin *and others* (2003). Similar to the discussion in Section 2.2, despite the fact that $\hat{\beta}$ can be easily obtained via iterations, it is difficult to obtain an analytical expression for the limiting covariance matrix. As in the case of $\hat{\beta}_G$, one can adopt the resampling approach and carry out inference on β_0 based on the empirical distribution of $\hat{\beta}^*$, the set of estimates under perturbation.

3. Numerical results

3.1 *Simulations*

Results of simulations of four study designs are detailed in the supplementary material available at *Biostatistics* online. In particular, we considered four biased sampling designs that are under the scope of the proposed framework, namely (i) length-biased sampling, (ii) case–cohort design, (iii) generalized case–cohort design, and (iv) a combo biased sampling: case–cohort analysis on length-biased data. For generalized case–cohort design (Kim *and others*, 2013), case–cohort samples are selected from the full cohort by sampling from cases with a probability of $p_1(\Delta, \tilde{T}, \boldsymbol{X})$ and controls with a possibly different probability of $p_2(\Delta, \tilde{T}, \boldsymbol{X})$, where $\boldsymbol{X}$ denotes covariates that may include (part of) the model covariates $\boldsymbol{Z}$. Such a design can be regarded as a generalization of the stratified case–cohort design; see Borgan *and others* (2000). In all the examples presented, our method produces nearly unbiased parameter estimates with empirical coverage probabilities of the confidence intervals constructed close to their nominal values. As suggested by an anonymous referee, we also compared our results on length-biased data with Mandel and Ritov (2010); for various case–cohort settings, our approach was compared with that proposed in Nan *and others* (2009).

Our approach, in general, provides more desirable estimates in terms of lower biasedness and/or smaller standard errors incurred.

3.2 *Real examples*

In this subsection, we analyze three datasets, namely (i) a shrub dataset of Muttlak and McDonald (1990) for size-biased data, (ii) a dementia study carried out by Canadian Study of Health and Aging, and (iii) South Wales Nickel Refinery Study on nasal sinus. The first two sets of data correspond to length-/size-biased data, while the last one involves case–cohort analysis and its variants.

The original dataset of shrub with contains data on 89 shrub samples. According to Wang (1996), the dataset includes 46 samples. Data were collected using a line-intercept sampling method for vegetation. Under this sampling technique, the probability a shrub was included in the sample was proportional to the width. Two indicator covariates were used to denote the three groups of transects to which the shrubs belonged. In the analysis with results reported in Table 1(a), we defined Z_1 and Z_2 to be indicators that the shrub belonged to transect I and transect III, respectively, in which case the second transect was the reference group. A total of 500 resamplings were carried out for estimating the standard errors of the estimators. The estimates obtained are in concordance with the conclusion drawn in Kim *and others* (2013) for the Cox proportional hazards model and show the proportional hazards model provides a better fit to the dataset.

The second example covers another version of length-biased sampling discussed in Vardi (1989), Wang (1991), and Shen *and others* (2009). It also corresponds to the second type of censoring under the length-bias setting studied in Tsai (2009). This configuration involves censoring of the residual lifetime after the data are sampled with bias. Under this setting, individuals' unbiased failure times can be observed only when they are event-free before the unbiased truncation time A^*, i.e. $T^* \geqslant A^*$. For the observed (biased) data, we define V as the time measured from the initiation time A to failure, which is regarded as the residual survival time. We assume that V is censored by $\tilde{C}$, the residual potential censoring variable, and that $\tilde{C}$ is independent of (A, V) conditional on Z. Then the observed survival and censoring times, T and C, can be expressed, respectively, as $T = A + V$ and $C = A + \tilde{C}$; again $\tilde{T} = \min(T, C)$ and $\Delta = I(T \leqslant C)$.

It can be shown that, conditional on the truncation time A, the weight function ω is

$$\omega_i(t) = \frac{q_{Z_i}(\tilde{T}_i, \Delta_i)}{\tilde{q}_{Z_i}(\tilde{T}_i, \Delta_i)} \times \frac{\tilde{q}_{Z_i}(t, 1)}{q_{Z_i}(t, 1)} = I(A_i \leqslant t).$$

This weight function leads to

$$E_Z\{\mathrm{d}N_i(t) - I(A_i \leqslant t)Y_i(t)\Lambda(t \mid Z_i)\} = 0 \quad \text{and} \tag{3.1}$$

$$E_Z\{\mathrm{d}N_i(t) - \Delta_i I(\tilde{T}_i - A_i \leqslant t)Y_i(t)\Lambda(t \mid Z_i)\} = 0, \tag{3.2}$$

where (3.2) is due to the stationarity assumption, see Vardi (1989). We can, therefore, define a family of subject-specific weight functions by combining (3.1) and (3.2):

$$\omega_i(t) = \pi I(A_i \leqslant t) + (1 - \pi)\Delta_i I(\tilde{T}_i - A_i \leqslant t), \tag{3.3}$$

where $\pi \in [0, 1]$. We recommend the choice of $\pi = 0.5$, which can strike a balance between (3.1) and (3.2). The indicator functions included, however, may not necessarily lead to a monotone estimating equation as discussed at the end of Section 2.2. Our numerical experience, however, shows that this non-monotone structure does not affect the estimation accuracy significantly for 2D cases. To handle this problem for a more general setting, one may adopt the induced smoothing approach for the AFT model proposed in

Table 1. (a) *Estimates and standard errors for regression parameters in Shrub Dataset*; *see* Muttlak and McDonald (1990). (b) *Estimates and standard errors for regression parameters CSHA Dataset.* (c) *Analysis of time from the first employment to the nasal sinus cancer death for the Welsh Nickel Refiners Study*; Breslow and Day (1987)

(a)

Parameter	Gehan	Log-rank	Parameter	Gehan	Log-rank
β_1			β_2		
Est.	0.580	0.624	Est.	-0.182	-0.221
SE	0.213	0.235	SE	0.210	0.231

(b)

Parameter	Gehan	Log-rank	Parameter	Gehan	Log-rank
β_{VD}			β_{PA}		
Est.	-0.043	0.014	Est.	-0.060	-0.050
SE	0.044	0.030	SE	0.050	0.060

(c)

	Full cohort		Case–cohort		Gen. Case–cohort	
Parameter	Gehan	Log-rank	Gehan	Log-rank	Gehan	Log-rank
$\log(\text{AFE} - 10)$						
Est.	0.612	0.598	0.706	0.768	0.604	0.589
SE	0.110	0.103	0.122	0.148	0.126	0.137
$(\text{YFE} - 1915)/10$						
Est.	0.014	0.017	0.168	0.223	0.044	0.041
SE	0.077	0.075	0.101	0.116	0.083	0.086
$(\text{YFE} - 1915)^2/100$						
Est.	-0.173	-0.304	0.019	-0.054	-0.230	-0.322
SE	0.103	0.130	0.129	0.181	0.104	0.126
$\log(\text{EXP} + 1)$						
Est.	0.168	0.210	0.254	0.324	0.147	0.183
SE	0.041	0.042	0.052	0.066	0.044	0.047

Chiou *and others* (2015) to avoid the multiple-root problem. This is, however, beyond the scope of this paper.

The dataset collected in the Canadian Study of Health and Aging (CSHA) study identified 1132 seniors aged 65 or above having the disease. All the dementia patients had undergone a follow-up procedure up to 1996. Entries with missing attributes are removed, resulting in the final dataset of size 818. Each individual can be categorized into three groups, namely (i) probable Alzheimer's disease (393 patients), (ii) possible Alzheimer's disease (252 patients), and (iii) vascular dementia (173 patients). The censoring rate is slightly over 22%. Table 1(b) reports the estimated values of β_{VD} and β_{PA} based on (3.3). They can be interpreted as the marginal effect on the log-survival times of vascular disease and possible Alzheimer patients, respectively. Neither of these indicator variables is statistically significant. In other words, there is no significant difference in survival times amongst the three groups of patients. Our finding agrees with the conclusion made in Shen *and others* (2009).

In addition to length-/size-biased data, we also analyze a case–cohort sampling-related sample called the South Welsh Nickel Refiners Study dataset. It contains altogether 679 subjects employed in a nickel refinery. The records of these workers are registered in Appendix VIII of Breslow and Day (1987). The study continued the follow-up until 1981 when 56 deaths were observed from cancer of the nasal sinus. The corresponding censoring rate is ~92%. Previous studies discover three significant risk factors, which include AFE (age at first employment), YFE (year at first employment), and EXP (exposure level). In Kim *and others* (2013), specifically, they set up the following four variables as regression covariates: $\log(\texttt{AFE} - 10)$, log of the age of the first employment minus 10 years, $(\texttt{YFE} - 1915)/10$, $(\texttt{YFE} - 1915)^2/100$, two transformed versions of number of years working in the refinery since 1915 and $\log(\texttt{EXP} + 1)$, the log exposure level; some of the subjects had zero exposure and hence $\texttt{EXP} + 1$ is considered so that its logged value is non-negative and well-defined.

In Table 1(c), the first column presents the estimated parameter values obtained from the full cohort dataset. In this case, $p = 1$ for all observations. The second column displays the results from fitting the same model to data obtained from a randomly drawn, hypothetical subcohort. Such a subcohort contains all the observed failures and some censored subjects that make up two-thirds of the size of the subcohort. We also performed an analysis on another hypothetical subcohort which was drawn from the generalized case–cohort sampling scheme. We used selection probability $p(\tilde{T}) = 1 - \{1 + \exp(1 + \tilde{T}\gamma)\}^{-1}$, where $\gamma = -0.010$ and -0.050 for $p_1(t)$ and $p_2(t)$, respectively. The estimated values of $\boldsymbol{\beta}$ and their standard deviations are summarized in the third column of Table 1(c). All of these studies indicate that the covariates $\log(\texttt{AFE} - 10)$ and $\log(\texttt{EXP} + 1)$ are statistically significant. Compared with the full cohort study, the estimated standard deviation of $\hat{\boldsymbol{\beta}}$ presented for the two case–cohort studies are slightly inflated. The estimates obtained from this generalized case–cohort sampling scheme are closed to the corresponding values obtained by using a full cohort. Under the generalized case–cohort setting, however, only 65% of the cases were included in the subcohort with the sample size only about 30% of that of the original samples.

4. Conclusion

In this paper, we propose an inference procedure for the regression parameter and the distribution function of the error term in the AFT model under general biased sampling schemes. This is a parallel of Kim *and others* (2013) to the AFT counterpart. This methodology again covers special cases of general biased sampling schemes in which the sampling probability depends on the outcome variable $(\tilde{T}, \Delta)$ under one unified framework. The weight function ω introduced that adjust the bias due to various sampling schemes plays a major role. As we have demonstrated, the corresponding weight functions for many common sampling schemes are readily available. This adds versatility and usefulness to the proposed method.

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors would like to thank anonymous referees for their valuable comments which have significantly improved the paper, in particular for our treatments for the length-biased sampling case. Acknowledgement should also be dedicated to the CSHA for permission for usage of part of the data collected for the dementia study. *Conflict of Interest*: None declared.

Funding

T.S. acknowledges the financial support of the Research Grants Council of Hong Kong ECS 24300514. Z.Y. research is supported in part by the National Science Foundation DMS-1308566 and the National Institutes of Health R37GM047845.

References

Asgharian, M., M'Lan, C. E. and Wolfson, D. B. (2002). Length-biased sampling with right censoring: an unconditional approach. *Journal of the American Statistical Association* **95**, 888–902.

Asgharian, M. and Wolfson, D. B. (2005). Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics* **33**, 2109–2131.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000). Exposure stratified case–cohort designs. *Lifetime Data Analysis* **6**, 39–58.

Breslow, N. E. and Day, N. E. (1987) *Statistical Methods in Cancer Research, Vol. II: The Design and Analysis of Cohort Studies*. Lyon, France: IARC.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66**, 429–436.

Chen, Y. Q. (2010). Semiparametric regression in size-biased sampling. *Biometrics* **66**, 149–158.

Chiou, S., Kang, S. and Yan, J. (2015). Rank-based estimating equations with general weight for the accelerated failure time model: an induced smoothing approach. *Statistics in Medicine* **34**, 1495–1510.

Cox, D. R. (1969). Some sampling problems in technology. In: Johnson and Smith (editors), *New Developments in Survey Sampling*. New York: Wiley.

Cox, D. R. (1972). Regression models and life-tables (with Discussion). *Journal of the Royal Statistical Society. Series B* **34**, 187–220.

Cox, D. R. and Oakes, D. (1984) *Analysis of Survival Data*. London: Chapman and Hall.

Gill, R. D. (1980). *Censoring and Stochastic Integrals*, Mathematical Center Tract 124. Amsterdam: Mathematische Centrum.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 133–143.

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. (2006). Rank regression analysis of multivariate failure time data based on marginal linear models. *Scandinavian Journal of Statistics* **33**, 1–23.

Kalbfleisch, J. D. and Prentice, R. L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd edition. New York: Wiley.

Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of Economic Literature* **26**, 646–679.

Kim, J. P., Lu, W., Sit, T. and Ying, Z. (2013). A unified approach to semiparametric transformation models under generalized biased sampling schemes. *Journal of the American Statistical Association* **108**, 217–227.

Kong, L. and Cai, J. (2009). Case–cohort analysis with accelerated failure time model. *Biometrics* **65**, 135–142.

Lu, W. B. and Tsiatis, A. A. (2006). Semiparametric transformation models for case–cohort study. *Biometrika* **93**, 207–214.

Mandel, M. and Ritov, Y. (2010). The accelerated failure time model under biased sampling. *Biometrics* **66**, 1306–1308.

McFadden, J. A. (1962). On the lengths of intervals in a stationary point process. *Journal of the Royal Statistical Society. Series B* **24**, 364–382.

Muttlak, H. A. and McDonald, L. L. (1990). Ranked set sampling with size-biased probability of selection. *Biometrics* **46**, 435–446.

Nan, B., Kalbfleisch, J. D. and Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *The Annals of Statistics* **37**, 2351–2376.

Nan, B., Yu, M. and Kalbfleisch, J. D. (2006). Censored linear regression for case–cohort studies. *Biometrika* **93**, 747–762.

Prentice, R. L. (1978). Linear rank tests with right-censored data. *Biometrika* **65**, 167–179.

Prentice, R. L. (1986). A Case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics* **18**, 303–328.

Shen, Y., Ning, J. and Qin, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association* **104**, 1192–1202.

Shen, Y., Ning, J. and Qin, J. (2012). Likelihood approaches for the invariant density ratio model with biased-sampling data. *Biometrika* **99**, 363–378.

Tsai, W.-Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* **96**, 601–615.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics* **18**, 354–372.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics* **10**, 616–620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics* **13**, 178–203.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* **76**, 751–761.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association* **86**, 751–761.

Wang, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika* **83**, 343–354.

Wei, L. J., Ying, Z. and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845–851.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics* **21**, 76–99.

Zeng, D. and Lin, D. Y. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association* **102**, 1387–1396.

[*Received November 18, 2014; revised December 21, 2015; accepted for publication January 28, 2016*]