UW Biostatistics Working Paper Series

7-18-2005

Linear Regression of Censored Length-biased Lifetimes

Ying Qing Chen
Fred Hutchinson Cancer Research Center, yqchen@u.washington.edu

Yan Wang
Division of Biostatistics, School of Public Health, University of California, Berkeley, yanw@stat.berkeley.edu

UW Biostatistics Working Paper Series

Year 2005 *Paper* 258

Linear Regression of Censored Length-biased Lifetimes

Ying Qing Chen[*] Yan Wang[†]

[*] Fred Hutchinson Cancer Research Center

[†] Division of Biostatistics, School of Public Health, University of California, Berkeley

Linear Regression of Censored Length-biased Lifetimes

Abstract

Length-biased lifetimes may be collected in observational studies or sample surveys due to biased sampling scheme. In this article, we use a linear regression model, namely, the accelerated failure time model, for the population lifetime distributions in regression analysis of the length-biased lifetimes. It is discovered that the associated regression parameters are invariant under the length-biased sampling scheme. According to this discovery, we propose the quasi partial score estimating equations to estimate the population regression parameters. The proposed methodologies are evaluated and demonstrated by simulation studies and an application to actual data set.

1 Introduction

Length-biased sampling schemes are usually involved in observational studies and sample surveys due to their convenience or cost-effectiveness. When these biased sampling schemes are used, length-biased lifetimes are observed and collected. Several examples of the length-biased lifetimes are discussed, for example, in a recent article of Asghrian, et al. (2002). To be specific, suppose that in a target population $\widetilde{T}$ is a positive lifetime random variable of p-vector covariate $Z \in \mathcal{Z}$. Its density function is $\widetilde{f}(\cdot \mid Z)$. Under the length-biased sampling scheme, the induced density function, $f(\cdot \mid Z)$, of the actually collected lifetime at $T = t$ is proportional to the magnitude of t, i.e.,

$$f(t \mid Z) = \frac{t\widetilde{f}(t \mid Z)}{\widetilde{\mu}(Z)}, \tag{1}$$

where $\widetilde{\mu}(Z) = E(\widetilde{T} \mid Z) = \int_0^\infty u\widetilde{f}(t \mid Z)du < \infty$. A comprehensive review on the distributional properties and historical research with $f(\cdot)$ can be found in Wang (1998).

Although the phenomenon of length bias has been noted and studied in statistical literature for decades, most of the methodology development has been focused on the one-sample estimation of population distributions. In many real applications, practical interest may be in the regression analysis of the population lifetimes to study their association with the covariates. As an effort, Wang (1996) proposed the widely used proportional hazards model (Cox, 1972) in such regression analysis:

$$\widetilde{\lambda}(t \mid Z) = \widetilde{\lambda}_0(t)\exp\left(\alpha^{\mathrm{T}} Z\right), \tag{2}$$

where $\alpha \in \mathcal{B} \subset \mathbf{R}^p$ is the regression parameter. Here, $\widetilde{\lambda}_0(\cdot)$ is some unspecified baseline hazard function, and $\widetilde{\lambda}(\cdot \mid Z)$ is the hazard function of Z, respectively. To estimate the parameters in model (2), however, the usual conditional probability argument on the length-biased samples of (T, Z) does not lead to a clear decomposition of the partial likelihood. In fact, Wang (1996) noticed this embedded difficulty with the proportional hazards model and proposed a pseudo-likelihood approach of riskset sampling to estimate α and $\widetilde{\lambda}_0(\cdot)$.

One reason for such difficulty is that the proportionality in (2) does not hold in the induced hazard functions of the length-biased lifetimes, as shown later. In this article, we instead consider an important alternative linear regression model, also known as the accelerated failure time model in

$$\log \widetilde{T} = -\beta^{\mathrm{T}} Z + \epsilon, \tag{3}$$

where $\beta \in \mathcal{B} \subset \mathbf{R}^p$ is the regression parameter. Here, ϵ are the random variables with unspecified distribution functions. The book of Kalbfleisch & Prentice (2002, Ch. 7) provides detailed discussion on the implication, estimation and application of this regression model for the lifetimes without length bias.

In the sections to follow, we first derive the formula of the induced hazard function of the length-biased T. This formula sheds new light on the advantage of using the linear regression model (3) for the population lifetimes, due to the invariance of regression parameter in the length bias sampling. We further propose the quasi partial score estimating equations and a Riccati ordinary differential equation to estimate the population parameters of β and the baseline function, respectively, based on the actual length-biased samples of (T, Z). The validity and performance of the proposed inference procedures are evaluated by Monte-Carlo simulations. For the demonstration purpose, the data in Wang (1996) are used as well for our proposed methods. This article is mainly methodological, while most of its theory justification can be adapted from the references thereinafter.

2 Length-biased Hazard Functions

Let $\widetilde{S}(t \mid Z) = \text{pr}\{\widetilde{T} \geq t \mid Z\}$ be the survival function, and $\widetilde{m}(t \mid Z) = E(\widetilde{T} - t \mid \widetilde{T} \geq t, Z)$ be the mean residual life function, respectively, for the population lifetime $\widetilde{T}$. According to the inversion formula in Cox (1962, p. 128),

$$\int_t^\infty \widetilde{S}(u \mid Z)du = \widetilde{S}(t \mid Z) \cdot \widetilde{m}(t \mid Z).$$

An integration by parts further leads to $\int_t^\infty u\widetilde{f}(u \mid Z)du = \widetilde{S}(t \mid Z)\{\widetilde{m}(t \mid Z) + t\}$, for $t > 0$, given that $\widetilde{\mu}(Z) = E(\widetilde{T} \mid Z) < \infty$. The survival function of the length-biased T is hence

$$S(t \mid Z) = \int_t^\infty f(u \mid Z)du = \frac{1}{\widetilde{\mu}(Z)} \int_t^\infty u\widetilde{f}(u \mid Z)du = \frac{1}{\widetilde{\mu}(Z)}\widetilde{S}(t \mid Z)\{\widetilde{m}(t \mid Z) + t\}.$$

As a result, the induced hazard function of the length-biased T is

$$\lambda(t \mid Z) = \frac{f(t \mid Z)}{S(t \mid Z)} = \frac{t\widetilde{f}(t \mid Z)}{\widetilde{\mu}(Z)} \cdot \frac{\widetilde{\mu}(Z)}{\widetilde{S}(t \mid Z)\{\widetilde{m}(t \mid Z) + t\}} = \frac{\widetilde{\lambda}(t \mid Z)t}{\widetilde{m}(t \mid Z) + t}, \tag{4}$$

respectively. It is clear that $\lambda(t \mid Z) \leq \widetilde{\lambda}(t \mid Z)$ given $\widetilde{m}(t \mid Z) \geq 0$, for any $t > 0$.

When the proportional hazards models are assumed for the population lifetimes as in (2), their length-biased hazard functions would follow

$$\lambda(t \mid Z) = \frac{\widetilde{m}_0(t) + t}{\widetilde{m}(t \mid Z) + t} \cdot \lambda_0(t) \exp(\alpha^{\mathrm{T}} Z).$$

Here, $\lambda_0(t) = \lambda(t \mid Z = 0)$ and $\widetilde{m}_0(t) = \widetilde{m}(t \mid Z = 0)$, respectively. Since $\widetilde{m}(t \mid Z) + t$ is usually unknown function of both t and Z under (2), the proportionality between the length-biased hazard functions generally does not hold. Thus any naive application of the proportional hazards models to the length-biased lifetimes would cause biased estimation on the population regression parameters of α.

Consider instead the accelerated failure time models assumed for the population lifetimes as in (3). We denote the hazard function of $\exp(\epsilon)$ in the model (3) by $\widetilde{\lambda}_0(\cdot)$ as well. Then under the accelerated failure time model, $\widetilde{S}(t \mid Z) = \widetilde{S}_0\{t \exp(\beta^{\mathrm{T}} Z)\}$ and $\widetilde{\lambda}(t \mid Z) = \widetilde{\lambda}_0\{t \exp(\beta^{\mathrm{T}} Z)\} \exp(\beta^{\mathrm{T}} Z)$, which leads to

$$\widetilde{m}(t \mid Z) = \frac{\exp(-\beta^{\mathrm{T}} Z)}{\widetilde{S}_0\{t \exp(\beta^{\mathrm{T}} Z)\}} \int_{t \exp(\beta^{\mathrm{T}} Z)}^{\infty} \widetilde{S}_0(u) du = \exp(-\beta^{\mathrm{T}} Z) \widetilde{m}_0 \left\{t \exp(\beta^{\mathrm{T}} Z)\right\},$$

and

$$\lambda(t \mid Z) = \frac{\widetilde{m}_0'\{t \exp(\beta^{\mathrm{T}} Z)\} + 1}{\widetilde{m}_0\{t \exp(\beta^{\mathrm{T}} Z)\} \exp(-\beta^{\mathrm{T}} Z) + t} \cdot \frac{t \exp(\beta^{\mathrm{T}} Z)}{\widetilde{m}_0\{t \exp(\beta^{\mathrm{T}} Z)\}}, \tag{5}$$

respectively, due to the fact that $\widetilde{\lambda}_0(t) = \{\widetilde{m}_0'(t) + 1\}/\widetilde{m}_0(t)$. Let $\lambda_0(t) = \widetilde{\lambda}_0(t) t / \{\widetilde{m}_0(t) + t\}$. By (5), we thus obtain that

$$\lambda \left\{t \exp(-\beta^{\mathrm{T}} Z) \mid Z\right\} = \lambda_0(t) \exp(\beta^{\mathrm{T}} Z). \tag{6}$$

Thus the length-biased lifetimes follow the accelerated failure time model with the same regression parameters of β as in their population models. That is, the regression parameters are invariant under the accelerated regression model. This fact would yield much advantage over the usual proportional hazards model in making inferences on the regression parameters, as shown in the later sections.

3 Inferences on Population Parameters

Suppose that the observed data consist of n iid copies, (X_i, Δ_i, Z_i), $i = 1, 2, \ldots, n$, of (X, Δ, Z), which are the length-biased samples of model (3). Here, $X_i = \min(T_i, C_i)$ and

$\Delta_i = I(T_i \leq C_i)$, respectively, where C_i are the potential censoring times. Assume that (T_i, C_i) are independent given Z. Let $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ and $Y_i(t) = I(X_i \geq t)$. Given the observed $\{(X_i, \Delta_i, Z_i), i = 1, 2, \ldots, n\}$, the log likelihood function is then $l = \sum_{\mathrm{i}}\{\Delta_i \log \lambda(X_i \mid Z_i) + \log S(X_i \mid Z_i)\}$. As a result, the score function for β is

$$\frac{\partial l}{\partial \beta} = -\sum_{i=1}^{n} \int_0^\infty \left\{ \frac{\partial \log \lambda(t \mid Z_i)}{\partial \beta} \right\} Z_i \left\{ dN_i(t) - Y_i(t)\lambda(t \mid Z_i)dt \right\}. \tag{7}$$

If $\widetilde{m}_0(\cdot)$ is known, for example, to be exponential such as $\widetilde{m}_0(t) = \theta$, where $\theta > 0$ is constant, $\partial \log \lambda(t \mid Z)/\partial \beta$ can be calculated as $Z\{1 + 2\theta \exp(-\beta^{\mathrm{T}} Z) + 1\}/\{1 + \theta \exp(-\beta^{\mathrm{T}} Z)\}$. By solving $\partial l/\partial \beta = 0$ and $\partial l/\partial \theta = 0$ simultaneously, the maximum likelihood estimators of β and θ are obtained.

When $\widetilde{m}_0(\cdot)$ is unknown, although the baseline hazard function of the length-biased lifetimes are modified by the factor of $0 < t/\{\widetilde{m}_0(t) + t\} \leq 1$, the regression parameter of β in the population model (3) is invariant under the length-biased sample scheme as shown in (6). This fact would greatly simplify the estimation of the population parameters in model (3). Specifically to estimate β, we propose a quasi partial score estimating equation approach similar to that in Chen & Jewell (2001). Let β_* and $\Lambda_*(\cdot)$ be the true value of β and $\Lambda_0(\cdot)$, respectively, in model (6), where $\Lambda(\cdot)$ are the cumulative hazard functions. Then according to (6), $E\{dM_i(t) \mid \mathcal{F}_{t-}; \beta_*, \Lambda_*(\cdot)\} = 0$, where

$$dM_i(t; \beta, \widetilde{m}) = dN_i\left\{t \exp(-\beta^{\mathrm{T}} Z_i)\right\} - Y_i\left\{t \exp(-\beta^{\mathrm{T}} Z_i)\right\} d\Lambda_0(t),$$

for $i = 1, 2, \ldots, n$. Here, the filtration of $\mathcal{F}_{\mathrm{t}}$ is defined by

$$\sigma\left\{N_i\{u \exp(-\beta_*^{\mathrm{T}} Z_i)\}, Y_i\left\{u \exp(-\beta_*^{\mathrm{T}} Z_i)\right\}, Z_i; 0 \leq u \leq t, i = 1, 2, \ldots, n\right\}.$$

Thus, we consider the following quasi partial score estimating equations to solve for β and $\Lambda_0(\cdot)$ simultaneously,

$$\sum_{i=1}^{n} dM_i(t; \beta, \Lambda_0) = 0, \text{ and} \tag{8}$$

$$\sum_{i=1}^{n} \int_0^\tau Z_i dM_i(t; \beta, \Lambda_0) = 0, \tag{9}$$

where $\tau > 0$ is some constant such that $\liminf_n n^{-1} \sum_i \mathrm{pr}\{X_i \exp(\beta^{\mathrm{T}} Z_i) \geq \tau + \xi\} > 0$ for some $\xi > 0$ as in Tsiatis (1990). Denote $\widehat{\beta}$ and $\widehat{\Lambda}_0(\cdot)$ the solutions of β and $\Lambda_0(\cdot)$ in the above estimating equations, respectively. Straightforward algebra on (8) leads to that

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{\sum_i dN_i\{t\exp(-\beta^{\mathrm{T}} Z_i)\}}{\sum_i Y_i\{t\exp(-\beta^{\mathrm{T}} Z_i)\}}.$$

Replacing that in (9), we obtain the estimating equations for β:

$$Q(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{Z_i - \bar{Z}(t;\beta)\right\} dN_i\left\{t\exp(-\beta^{\mathrm{T}} Z_i)\right\} = 0, \qquad (10)$$

where $\bar{Z}(t) = \mathcal{E}^{(1)}(t;\beta)/\mathcal{E}^{(0)}(t;\beta)$ with $\mathcal{E}^{(k)}(t;\beta) = n^{-1}\sum_i Z_i^{\otimes k} Y_i\{t\exp(-\beta^{\mathrm{T}} Z_i)\}$, $k = 0, 1, 2$. Assume that $\lim_n \mathcal{E}^{(k)}(t;\beta) = e^{(k)}(t;\beta)$. Since $Q(\beta)$ is discontinuous in β, an estimator of β is defined by the $\widehat{\beta}$ that minimises of $\|Q(\beta)\|^2$ as in Wei, et al. (1990). Following the same conditions and arguments for the estimating equations of the log-rank type in Ying (1993), there exists a neighbourhood $U(\beta_*)$ of β_* such that $\widehat{\beta} = \operatorname{argmin}_{\beta\,\in\,U(\beta_*)}\|Q(\beta)\|^2$ is strongly consistent. Let

$$\begin{aligned} D &= \int_0^{\tau} \{e^{(2)}(t) - e^{(1)}(t)^{\otimes 2}/e^{(0)}(t)\} E Y_1\{t\exp(-\beta_*^{\mathrm{T}} Z_1)\}\lambda_0'(t)dt, \text{ and} \\ V &= \int_0^{\tau} \{e^{(2)}(t) - e^{(1)}(t)^{\otimes 2}/e^{(0)}(t)\} E Y_1\{t\exp(-\beta_*^{\mathrm{T}} Z_1)\}\lambda_0(t)dt, \end{aligned}$$

respectively. Since D is nonnegative definite, $n^{1/2}(\widehat{\beta} - \beta_*) \xrightarrow{\mathcal{L}} N(0, D^{-1}VD^{-1})$, as $n \to \infty$. In addition, assume that there exist possibly data-dependent weight functions of $W(t;\beta)$ such that $W(t;\beta_*) \to w(t)$ almost surely. The weighted estimating equations for β can be also considered:

$$Q_W(\beta) = \sum_{i=1}^{n} \int_0^{\tau} W(t;\beta)\left\{Z_i - \bar{Z}(t;\beta)\right\} dN_i\left\{t\exp(-\beta^{\mathrm{T}} Z_i)\right\} = 0.$$

Denote $\widehat{\beta}_W$ the solution to the above equations. In general, the optimal $W(\cdot)$ that minimises the asymptotic variance of $\widehat{\beta}_W$ should be proportional to $\lambda_0'(t)/\lambda_0(t)$ by an application of the Cauchy-Schwarz inequality, as noted in Tsiatis (1990). Moreover, adaptive $W(\cdot)$ can be similarly constructed as in Lai & Ying (1992) to stabilise the integral tail and hence extend the finite τ to ∞.

To estimate the population baseline functions, consider that $\lambda(t) = \{1 + \widetilde{m}_0'(t)\}t/[\{\widetilde{m}_0(t) + t\}\widetilde{m}_0(t)]$. This leads to a Riccati ordinary differential equation (Reid, 1972),

$$\widetilde{m}_0'(t) - \frac{\lambda(t)}{t}\widetilde{m}_0(t)^2 - \lambda_0(t)\widetilde{m}_0(t) + 1 = 0. \qquad (11)$$

Nominally, $\widetilde{m}_0(t) = -t$ is a solution to this equation. According to Polyanin & Zaitsev (2003), this equation has a closed form of solution

$$\widetilde{m}_0(t) = \frac{\mu_0 \exp\left\{-\int_0^t d\Lambda_0(u)\right\}}{1-\mu_0\int_0^t u^{-1}\exp\left\{-\int_0^u d\Lambda_0(s)\right\})d\Lambda_0(u)} - t,$$

given the initial condition of $\widetilde{m}_0(0) = \int_0^\infty \exp\{-\int_0^t d\Lambda_0(u)\}dt = \mu_0$. Therefore, although the population baseline hazard function is generally not straightforward to be estimated, a natural estimator of the population baseline mean residual life function is

$$\widehat{\widetilde{m}}_0(t;\widehat{\beta}) = \frac{\widehat{\mu}_0 \exp\left\{-\int_0^t d\widehat{\Lambda}_0(u;\widehat{\beta})\right\}}{1-\widehat{\mu}_0\int_0^t u^{-1}\exp\left\{-\int_0^u d\widehat{\Lambda}_0(s)\right\} d\widehat{\Lambda}_0(u;\widehat{\beta})} - t, \tag{12}$$

where $\widehat{\mu}_0 = \int_0^\tau \exp\{-\int_0^t d\widehat{\Lambda}_0(u;\widehat{\beta})\}$. Here, $\widehat{\Lambda}_0(t;\widehat{\beta})$ is the Breslow-type estimator in the form of $\int_0^t \sum_i dN_i\{t\exp(-\widehat{\beta}^{\mathrm{T}}Z_i)\}/\sum_i Y_i\{t\exp(-\widehat{\beta}^{\mathrm{T}}Z_i)\}$. Similar to the decomposition in Tsiatis (1981),

$$n^{1/2}\{\widehat{\widetilde{m}}_0(t;\widehat{\beta}) - \widetilde{m}_0(t)\} = n^{1/2}\{\widehat{\widetilde{m}}_0(t;\widehat{\beta}) - \widehat{\widetilde{m}}_0(t;\beta_*)\} + n^{1/2}\{\widehat{\widetilde{m}}_0(t;\beta_*) - \widetilde{m}_0(t)\}$$

can be thus shown to converge weakly to a zero-mean Gaussian process.

4 Estimate and Variance Calculation

To calculate $\widehat{\beta}$ of the estimating equations (10), in addition to the way of minimising $\|Q(\beta)\|^2$, direct grid search such as the bisection method can be used when Z is of low-dimension. When Z is of moderate dimension, the recursive bisection can be used, i.e., to recursively apply the univariate bisection search to the k-dimension problem given the $(k-1)$-dimensional problem solved (Huang, 2002). When Z is of high dimension, random search methods such as the simulated annealing method in Lin & Geyer (1992) can be used. However, the estimating equations (9) in general may have multiple solutions, some of which are not consistent (Fygenson & Ritov, 1994). When $W(\cdot)$ is the Gehan weight function, the weighted estimating functions $Q_W(\cdot)$ is monotone, and solving $Q_W(\beta) = 0$ reduces to a simple linear programming problem. Thus, the method in Jin, et al. (2003) can be used reliably to calculate the estimates.

To estimate the variance of $\widehat{\beta}$ can be challenging, since the baseline hazard function and its derivative are both involved. Tsiatis (1990) proposed to estimate V by

$$\widehat{V} = n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \{\mathcal{E}^{(2)}(t;\widehat{\beta}) - \mathcal{E}^{(1)}(t;\widehat{\beta})^{\otimes 2}/\mathcal{E}^{(0)}(t;\widehat{\beta})\} Y_i\{t \exp(-\widehat{\beta}^{\mathrm{T}} Z_i)\} d\widehat{\Lambda}_0(t;\widehat{\beta}),$$

and D by

$$\widehat{D} = -n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \widehat{\lambda}_0(t;h_n) d[\{\mathcal{E}^{(2)}(t;\widehat{\beta}) - \mathcal{E}^{(1)}(t;\widehat{\beta})^{\otimes 2}/\mathcal{E}^{(0)}(t;\widehat{\beta})\} Y_i\{t \exp(-\widehat{\beta}^{\mathrm{T}} Z_i)],$$

respectively. Here, $\widehat{\lambda}_0(t;h_n)$ is a consistent kernel estimator of the baseline hazard function of bandwidth h_n. Alternatively, a simple numerical differentiation approach in Huang (2002) can be used to calculate the variance. That is, first decompose $\widehat{V} = vv^{\mathrm{T}}$, where $v = (v_1, v_2, \ldots, v_p)^{\mathrm{T}}$ and solve $Q(b_i) = v_i$, $i = 1, 2 \ldots, p$. Then the variance of $n^{1/2}(\widehat{\beta} - \beta_*)$ can be conveniently estimated by $(b - \widehat{\beta})(b - \widehat{\beta})^{\mathrm{T}}$, where $b = (b_1, b_2, \ldots, b_p)^{\mathrm{T}}$.

Several computer-intensive methods can be also used to calculate the variance, in addition to the usual bootstrap method by Efron (1976). One approach is the estimating function bootstrap of Hu & Kalbfleisch (2000), in which the individual terms of the estimating functions are bootstrapped. The other approach is often used in literature, which is by Parzen, et al. (1994) and recently further generalised by Chatterjee & Bose (2005). To implement it, n independent standard normal deviates of $(G_1, G_2, \ldots, G_n)$ are generated and multiplied as in $Q_{G}(\beta) = \sum_i G_i Q_i(\beta)$, where $Q_i(\beta)$ are the individual terms in $Q(\beta)$. Then the variance of $\widehat{\beta}$ can be approximated by the empirical variance of $\widehat{\beta}_{G}$, where $\widehat{\beta}_{G}$ are solutions of $Q_{G}(\beta) = 0$, as shown in Lin, et al. (1998). Moreover, this approach can be easily adapted to estimate the confidence intervals and confidence bands based on the estimator of the population baseline mean residual life function, although the covariance of $\widehat{\widetilde{m}}_0(\cdot)$ itself is rather complex.

5 Examples and Numerical Studies

Consider that $\exp(\epsilon)$ in the population linear regression model (3) are exponential with the constant mean of $\mu > 0$. Then the population baseline hazard function is $\widetilde{\lambda}_0(t) = 1/\mu$. Assume that $\beta = -\log 2$ for $Z = 1$ against $Z = 0$. In Fig (1), both population and length-biased hazard functions are plotted in $Z = 0$ and 1, respectively. As shown in the plot, the length-biased hazard functions are smaller than their population ones, respectively. The difference is more prominent in the early period of time. As the time progresses, the length-biased hazard functions move closer to the population hazard functions, respectively. The population hazard functions are indeed the asymptotes of the length-biased hazard functions, as t approaches ∞.

Since the underlying population distribution functions are exponential, the population hazard functions also follow the proportional hazards model (2) with $\alpha = \log 2$. In Fig (1), the hazard ratios are plotted for the population hazard functions and the length-biased hazard functions, respectively. As shown in the plot, although the hazard ratio for the population hazard functions remains constant of $1/2$, that of the length-biased hazard functions is not constant and indeed monotonically increasing from $t = 0$. This echoes the earlier discussion that any naive application of the proportional hazards model to the length-biased lifetimes would result in biased estimates without considering the unsatisfied proportionality.

[Fig (1) about here]

To contrast with the results in Wang (1996), moderate simulation studies are conducted under the similar settings. Assume that the population accelerated failure time model holds for $Z = 0$ and 1 in (3) with the null hypothesis of $\beta = 0$ and the alternative hypothesis of $\beta = 1$, respectively, where the distribution of $\exp(\epsilon)$ is Weibull with its hazard function being $2t$. When $\beta = 1$, the population survival functions are $\widetilde{S}(t \mid Z = 0) = \exp(-t^2)$ and $\widetilde{S}(t \mid Z = 1) = \exp(-e^2 t^2)$, respectively. As a result, the population hazard functions are $\widetilde{\lambda}(t \mid Z = 0) = 2t$ and $\widetilde{\lambda}(t \mid Z = 1) = 2e^2 t$, respectively. Hence the population lifetimes also follows the proportional hazards model (2) with $\alpha = 2$.

In the actual simulations, a total of $n = 50$, 200 and 500 length-biased lifetimes are simulated, respectively, representing relatively small, moderate and large sample sizes. The assignment of $Z = 0$ or $Z = 1$ is Bernoulli(0.5). Both of the accelerated failure time models and the proportional hazards models are fitted to the simulated length-biased lifetimes. The simulation results are summarised in Table (1). Each cell in the table is computed according to 1000 simulated samples. The biases are calculated as the average absolute differences between the 1000 estimates and the true population parameters. The coverage probabilities are calculated as the percentage of the 1000 95% nominal confidence intervals containing the ture population parameters. The sample standard error of the estimates and the mean of the estimator standard errors are also calculated, respectively.

[Table (1) about here]

As shown in the table, the estimators based on the accelerated lifetimes models are virtually unbiased and yield appropriate coverage probabilities of the 95% confidence intervals under both the null and the alternative hypotheses, even when the sample sizes are fairly small. Their mean standard errors and the standard errors of the estimates are also comparable. The naive application of the Cox proportional hazards models yields reasonable estimates under the null hypothesis. The estimates, however, tends to be biased with incorrect coverage probabilities under the alternative hypothesis.

We further use the shrub data example in Wang (1996) to compare the different models. The original shrub data can be found in Muttlak & McDonald (1990, Table 3). Similarly, we consider the lifetime proxy outcome of the shrub width Y, and the two covariates of $X_1 = I(y$ belongs to transect I) and $X_2 = I(y$ belongs to transsect II), respectively, where $I(.)$ is indicator function. The proportional hazards model is first applied to the length-biased Y's. The regression coefficients of X_1 and X_2 are estimated as 1.0659 ($s.e. = 0.482$, $p = 0.027$) and 0.0827 ($s.e. = 0.469$, $p = 0.860$), respectively. Compared with those in Table 2 of Wang (1996) under the population proportional hazards model, the standard errors are comparable. The point estimates are however different. Unlike in that shown in Fig. (1), the hazard ratios appear to be larger for the length-biased outcomes than their population counterparts in the shrub data.

The accelerated failure time model of $\log Y = -\beta_1 X_1 - \beta_2 X_2 + \epsilon$ is also applied to the length-biased Y's, where β_1 and β_2 are the parameters. We estimate $\widehat{\beta}_1 = 0.7705$ ($s.e. = 0.2550$) and $\widehat{\beta}_2 = 0.2091$ ($s.e. = 0.2492$) with the p-values of 0.004 and 0.4068, respectively. Compared with the population proportional hazards model, where both of the population covariates are not statistically significant ($p > 0.05$), the covariate X_1 is indeed a significant population predictor of Y in the accelerated failure time model. That is, whether or not the shrub belongs to transect I is associated with a significant average change in its width by 0.7705 less in log-scale. This association is identical for both the population outcomes and the collected length-biased outcomes.

6 Discussion

In general the assumption of proportionality in the proportional hazards model tend to be strong yet critical to its success in application. When the model is true, it shows robustness as evident in the simulation studies of Wang (1996), in which both of the point estimates and their variances tend to be irrelevant of the choice of L's. This however does not appear to be true in analysis of the shrub data with the proportional hazards model. One explanation, as pointed out in Wang (1996), may be related to the relatively large size of risksets in simulation studies. It is yet interesting to observe that the variances tends to be larger with larger L, although larger L is supposed to increase the use of more observations. This paradoxical discrepancy may be due to the inadequacy of the proportional hazards model, which is assumed in the simulation studies but however not verifiable in the actual shrub data analysis. Further investigation in the proportional hazards model's adequacy needs to be done for its impact on analysis of the length-biased outcomes.

On the other hand, the linear regression models, such as the accelerated failure time model, directly models the outcomes themselves. Their parameter interpretation is usually not limited to the hazard functions. This yields feasibility and flexibility for the linear regression model to be used for the outcomes other than the lifetimes. Especially in the biased sampling such as the length-biased sampling, the parameter invariance of the accelerated failure time model establishes its role similar to that of the logistic regression model for the binary outcomes in both prospective and retrospective study designs. Coupled with its relatively less stringent assumptions, the linear regression models may therefore well serve as an appealing alternative to the proportional hazards model.

References

Asgharian, M., M'Lan, C. R. & Wolfson (2002) Length-biased sampling with right censoring: an unconditional approach. *J. Amer. Statist. Asso.* **97**, 201-209.

Chatterjee, S. & Bose, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.* **33**, 414-436.

Chen, Y. Q. & Jewell, N. P. (2001). On a general class of hazards regression models. *Biometrika* **88**, 687-702.

Cox, D. R. (1962). *Renewal Theory.* Wiley: New York.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B **34**, 187-220.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Fygenson, M. & Ritov, Y. (1994). Monotone estimating equations for censored data. *Ann. Statist.* **22**, 732-746.

Hu, F. & Kalbfleisch, J. D. (2000). The estimating function bootstrap (with discussion). *Canad. J. Statist.* **28**, 449-499.

Huang, Y. (2002). Calibration regression of censored lifetime medical cost. *J. Amer. Statist. Asso.* **97**, 318-327.

Jin, Z., Lin, D. Y., Wei, L. J. & Ying, Z. (2003) Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341-353.

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

Lai, T. L. & Ying, Z. (1992). Linear rank statistics in regression analysis with censored or truncated data. *J. Multivariate Anal.* **40**, 12-45.

Lin, D. Y. & Geyer, C. J. (1992). Computational methods for semiparametric linear regression with censored data. *J. Comp. Graph. Statist.* **1**, 77-90.

Lin, D. Y., Wei, L. J. & Ying, Z. (1998). Accelerated failure time model for counting processes. *Biometrika* **85**, 605-618.

Parzen, M. I., Wei, L. J. & Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341-350.

Polyanin, A. D. & Zaitsev, V. F. (2003). *Handbook of Exact Solutions for Ordinary Differential Equations, 2nd Ed.*, p. 83. Chapman & Hall/CRC: Boca Raton.

Reid, W. T. (1972). *Riccati Differential Equations.* Academic Press: New York.

Tsiatis, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9**, 93-108.

Tsiatis, A. A. (1990). Estimating regression parameter using linear rank tests for censored data. *Ann. Statist.* **18**, 354-372.

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

Wang, M.-C. (1998). Length-biased sample. *Encyclopedia of Biostatistics*, Ed. P. Armitage and T. Colton, pp. 2223-2226. Wiley: New York.

Wei, L. J., Ying, Z. & 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. *Ann. Statist.* **21**, 76-99.

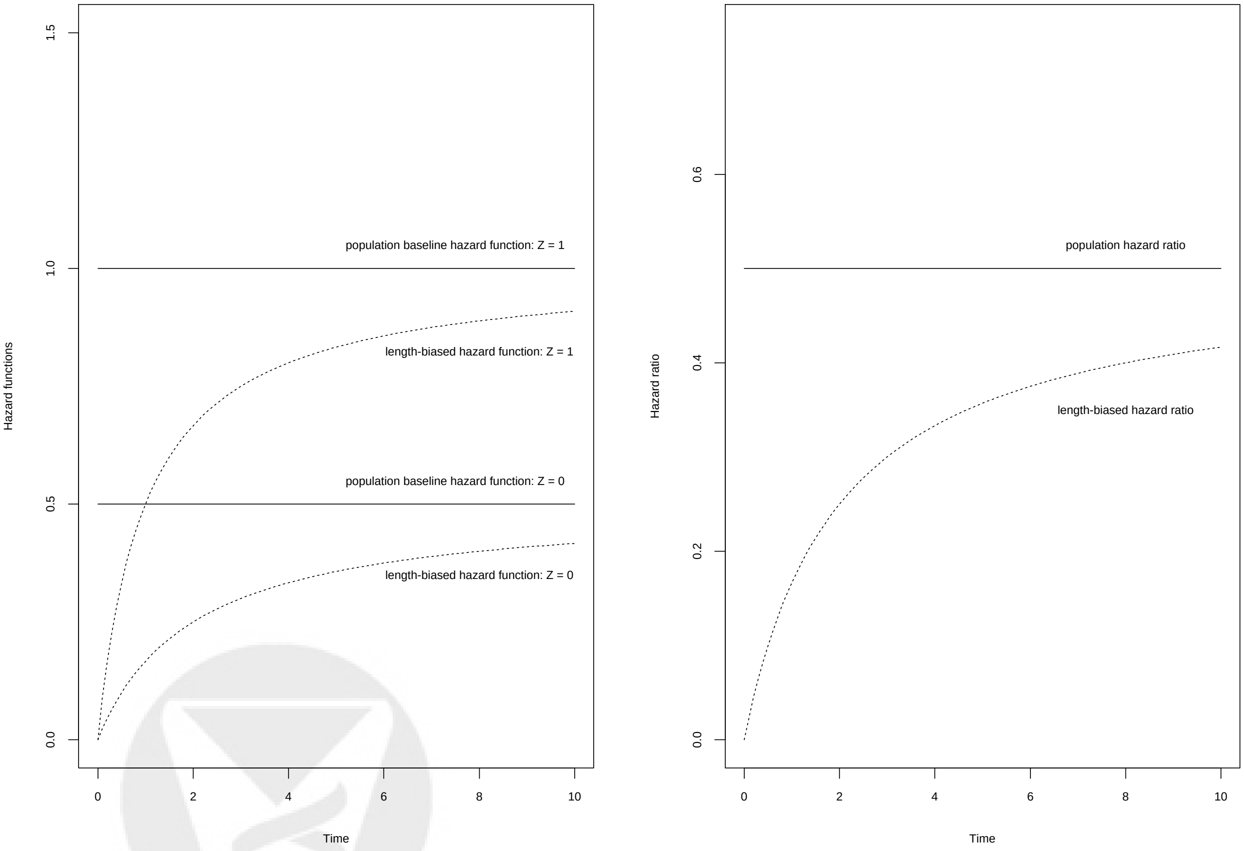
Table 1: Summary of Simulation Studies under the population models of the accelerated failure time model of $\log \widetilde{T} = -\beta Z + \epsilon$ and the proportional hazards model $\widetilde{\lambda}(t \mid Z) = \widetilde{\lambda}_0(t) \exp(\alpha Z)$.

	$\beta = 0$				$\alpha = 0$			
n	Bias	Cov. Prob.	SE	Mean SE	Bias	Cov. Prob.	SE	Mean SE
50	0.1088	0.950	0.1376	0.1367	0.2457	0.938	0.3138	0.2953
200	0.0551	0.955	0.0683	0.0684	0.1108	0.956	0.1396	0.1433
500	0.0337	0.956	0.0422	0.0431	0.0700	0.959	0.0879	0.0900

	$\beta = 1$				$\alpha = 2$			
n	Bias	Cov. Prob.	SE	Mean SE	Bias	Cov. Prob.	SE	Mean SE
50	0.1097	0.954	0.1362	0.1373	0.7161	0.853	1.1583	14.7820
200	0.0559	0.953	0.0696	0.0689	0.5409	0.327	0.2340	0.2291
500	0.0345	0.952	0.0426	0.0431	0.5315	0.017	0.1405	0.1422

Bias, absolute difference between 1000 estimates and the true value; Cov. Prob., percentage of 1000 95% nominal confidence intervals containing the true value; SE, sample standard error of 1000 estimates; Mean SE, average of 1000 estimator standard errors

Figure 1: Population and length-biased hazard functions under the linear regression model (3): solid lines are of the population exponential hazard functions and dotted lines are of their respective length-biased hazard functions, respectively.


Hazard functions
population baseline hazard function: Z = 1
length-biased hazard function: Z = 1
population baseline hazard function: Z = 0
length-biased hazard function: Z = 0
Time
Hazard ratio
population hazard ratio
length-biased hazard ratio
Time