Proportional mean residual life model for right-censored length-biased data

BY KWUN CHUEN GARY CHAN

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
kcgchan@u.washington.edu

YING QING CHEN AND CHONG-ZHI DI

Program in Biostatistics and Biomathematics, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.
yqchen@fhcrc.org cdi@fhcrc.org

SUMMARY

To study disease association with risk factors in epidemiologic studies, cross-sectional sampling is often more focused and less costly for recruiting study subjects who have already experienced initiating events. For time-to-event outcome, however, such a sampling strategy may be length biased. Coupled with censoring, analysis of length-biased data can be quite challenging, due to induced informative censoring in which the survival time and censoring time are correlated through a common backward recurrence time. We propose to use the proportional mean residual life model of Oakes & Dasu (*Biometrika* **77**, 409–10, 1990) for analysis of censored length-biased survival data. Several nonstandard data structures, including censoring of onset time and cross-sectional data without follow-up, can also be handled by the proposed methodology.

Some key words: Biased sampling; Bivariate survival data; Proportional hazards model; Renewal process.

1. INTRODUCTION

The longitudinal cohort study design has frequently been used in epidemiologic studies to investigate disease association with risk factors. To assemble a group of study participants in such a design, a conceptually straightforward sampling strategy is incident cohort sampling. For example, to study the incubation period between HIV-1 infection and development of an AIDS symptom, a subject would be recruited into the incident cohort as soon as his or her HIV positive serostatus is confirmed, and then followed till an AIDS-defining illness occurs. Although incident cohort sampling is a valid design to study the incubation period, it may need enormous resources and effort for subject accrual and follow-up, particularly when the HIV-infection rate is typically low even among high-risk populations because active antiretroviral therapy is highly effective for disease prevention.

Cross-sectional prevalent cohort sampling may offer a quick alternative way to assemble the cohort when the prevalence of initiating events is reasonably high. That is, only those who have experienced the initiating events would be sampled. However, when the disease outcome is the event time between initiating and failure events, cross-sectional sampling is likely to oversample those who have longer event times. One particular type of such cross-sectional sampling is so-called intercept or length-biased sampling (Canfield, 1941; Cox, 1969), when the sampling probability is proportional to the length of event times. For instance, when disease incidence follows a stationary Poisson process over time, cross-sectional sampling yields length-biased event times (Wang, 1991; Asgharian et al., 2002). Coupled with potential censoring on the event times, induced informative censorship may further complicate the analysis of such length-biased

event times because the survival time and censoring time are correlated through a common backward recurrence time.

Biased sampling has been studied at least as far back as [Wicksell \(1925\)](#) for the corpuscle problem. Prevalent cohort bias in cross-sectional sampling was first discussed by [Neyman \(1955\)](#). [Cox \(1969\)](#) studied biased sampling in quality control, and [Zelen & Feinleib \(1969\)](#) identified biased sampling in screening for chronic diseases. [Vardi \(1982\)](#) studied nonparametric estimation for length-biased data in the absence of censoring. For censored length-biased outcomes, nonparametric estimation of survival has been discussed by [Vardi \(1989\)](#), [Asgharian et al. \(2002\)](#) and [Huang & Qin \(2011\)](#). Regression methods based on the Cox models were developed for length-biased data ([Wang, 1996](#); [Qin & Shen, 2010](#)); methods based on the accelerated failure time model were also recently proposed ([Bergeron et al., 2008](#); [Shen et al., 2009](#); [Chen, 2010](#); [Mandel & Ritov, 2010](#)).

In this article, we focus instead on a new regression model for length-biased data, which relies on modelling mean residual life functions. This takes advantage of a natural relationship between the mean residual life function and a length-biased outcome. It allows an efficient use of data without resorting to artificial pseudolikelihood sampling or inverse probability of censoring weighting, as used in existing methodologies ([Wang, 1996](#); [Shen et al., 2009](#)).

2. A NEW REGRESSION METHOD

In a cross-sectional study, subjects are eligible to be sampled if, at the sampling time, they have experienced initiating events but not failure events. The event time, T say, for a typical subject has a natural decomposition $T = U + V$, where U is the time between the initiating event and when the subject is sampled, and V is the time from recruitment to the failure event. When T is considered to be a recurrence time of a renewal process, U and V are also called backward and forward recurrence times ([Cox, 1962](#)).

To study the association between T and a p -vector covariate Z , we consider the proportional mean residual life model ([Oakes & Dasu, 1990](#)),

$$m(t; Z) = m_0(t) \exp(\beta^T Z), \quad (1)$$

where $m(t; Z) = E(T - t \mid T > t, Z)$ is the mean residual life function of T at t , $m_0(t)$ is an unspecified baseline mean residual life function and β is a $p \times 1$ regression parameter. The primary goal is then to estimate β using cross-sectional length-biased data. For data arising from incident cohort sampling, methods have been developed, for example, by [Maguluri & Zhang \(1994\)](#) and [Chen & Cheng \(2005\)](#).

The mean residual life function itself is important in reliability, demographics and actuarial sciences, with many other applications. Due to its natural link to the hazard functions of backward and forward recurrence times, we can model mean residual life functions in regression analysis of length-biased data.

Let $f(t, z)$ be the joint density function of T and Z . The sampling distribution of (T, Z) under length-biased sampling is $tf(t, z)/\mu$, where μ is the mean of T . The sampling density of (V, Z) was also shown ([Cox, 1969](#)) to be

$$f_{V,Z}(v, z) = \mu^{-1} \int_v^\infty f(t, z) dt \quad (v > 0).$$

The conditional density of V given Z is therefore

$$f_{V|Z}(v | z) = \frac{f_{V,Z}(v, z)}{\int f_{V,Z}(v, z) dv} = \frac{\bar{F}_T(v | z)}{\mu(z)} \quad (2)$$

where $\bar{F}_T(t | z) = \text{pr}(T > t \mid Z = z)$ is the conditional survival function, and $\mu(z)$ is the conditional mean of T given $Z = z$. Let $\lambda(v; z)$ be the hazard function for the forward recurrence time V with covariate z . It follows from (2) that $\lambda(v; z) = 1/m(v; z)$, and it further follows from (1) that the sampling density of V

satisfies a proportional hazards model

$$\lambda(v; z) = \lambda_0(v) \exp(\gamma^\top z) \quad (3)$$

where $\gamma = -\beta$. See, for example, [Maguluri & Zhang \(1994\)](#) for a related discussion. Under right censoring of survival time, we observe (U, X, Δ, Z) , where $X = \min(V, C)$, $\Delta = I(V \leq C)$ and C is the time from recruitment to censoring. Note that C potentially censors the forward recurrence time and may be called the residual censoring time. Suppose (U, V) and C are conditionally independent given covariates Z . Since V follows a proportional hazards model, the assumptions on the censoring mechanism guarantee the validity of partial likelihood estimation using data (X, Δ, Z) to estimate β . Such an estimate, however, is inefficient, because information from U is not used. It also follows from [Cox \(1969\)](#) that the marginal sampling distribution of (U, Z) is the same as that of (V, Z) , which also follows a proportional hazards model (3). Note that (U, V) are correlated in general, but (U, V) satisfy marginal proportional hazards models. We could treat $\{(U_i, 1), (X_i, \Delta_i) : i = 1, \dots, n\}$ as paired survival data, and methods for multivariate survival data can be applied to estimate β . In particular, we consider an estimating function for bivariate survival data that is equivalent to a partial likelihood score function under an independent working correlation between the backward and forward times,

$$S_{\text{PL}}(\gamma) = \sum_{i=1}^n \left\{ Z_i - \frac{S^{(1)}(\gamma, U_i)}{S^{(0)}(\gamma, U_i)} \right\} + \Delta_i \left\{ Z_i - \frac{S^{(1)}(\gamma, X_i)}{S^{(0)}(\gamma, X_i)} \right\} \quad (4)$$

where $S^{(k)}(\gamma, t) = n^{-1} \sum_{i=1}^n Z^{\otimes k} \{I(U_i \geq t) + I(X_i \geq t)\} \exp(\gamma^\top Z_i)$ ($k = 0, 1, 2$) and for any vector a , $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^\top$. Let $\hat{\gamma}$ be a solution to the equation $S_{\text{PL}}(\gamma) = 0$ and $\hat{\beta} = -\hat{\gamma}$. The idea is similar to [Lee et al. \(1992\)](#) and [Spiekerman & Lin \(1998\)](#) for clustered survival data, and the validity of the method is based on the fact that $S^{(k)}(\gamma, t)$ converges to the same limit $s^{(k)}(\gamma, t) = E[Z^{\otimes k} \{I(U \geq t) + I(X \geq t)\} \exp(\gamma^\top Z)]$ regardless of whether or not (U, V) are independent. Since the problem is transformed to a bivariate survival model, asymptotic properties of $\hat{\beta}$ follow from [Spiekerman & Lin \(1998\)](#) and its asymptotic variance can be estimated by a sandwich-type estimate, $\hat{V} = \hat{A}^{-1} \hat{B} \hat{A}^{-1}$, where

$$\begin{aligned} \hat{A} &= \frac{1}{n} \sum_{i=1}^n \left(\frac{S^{(2)}(\hat{\gamma}; U_i)}{S^{(0)}(\hat{\gamma}; U_i)} - \left\{ \frac{S^{(1)}(\hat{\gamma}; U_i)}{S^{(0)}(\hat{\gamma}; U_i)} \right\}^{\otimes 2} \right. \\ &\quad \left. + \Delta_i \left[\frac{S^{(2)}(\hat{\gamma}; X_i)}{S^{(0)}(\hat{\gamma}; X_i)} - \left\{ \frac{S^{(1)}(\hat{\gamma}; X_i)}{S^{(0)}(\hat{\gamma}; X_i)} \right\}^{\otimes 2} \right] \right), \\ \hat{B} &= \frac{1}{n} \sum_{i=1}^n \left[\left\{ Z_i - \frac{S^{(1)}(\hat{\gamma}; U_i)}{S^{(0)}(\hat{\gamma}; U_i)} \right\} + \Delta_i \left\{ Z_i - \frac{S^{(1)}(\hat{\gamma}; X_i)}{S^{(0)}(\hat{\gamma}; X_i)} \right\} \right. \\ &\quad \left. - \sum_{j=1}^n \{h_i(U_j, U_i) + \Delta_j h_i(X_j, U_i) + h_i(U_j, X_i) + \Delta_j h_i(X_j, X_i)\} \right]^{\otimes 2}, \\ h_i(s, t) &= \frac{e^{\hat{\gamma}^\top Z_i} I(s \leq t) \{Z_i - S^{(1)}(\hat{\gamma}; s)/S^{(0)}(\hat{\gamma}; s)\}}{n S^{(0)}(\hat{\gamma}; s)}. \end{aligned}$$

In the above discussion, we assume the observation of standard right-censored length-biased survival data, where censoring events affect only the forward recurrence times, and the backward recurrence times are completely observed. However, since the initial events happen before recruitment, they may not always be recorded in practice. For instance, if the date of disease diagnosis can only be obtained from a disease registry up to two years before recruitment, then every subject with U greater than two years is subject to left administrative censoring. Let C_B be the time from censoring of the initial event to the time of recruitment, $Y = \min(U, C_B)$ and $\Delta_B = I(U \leq C_B)$, and suppose we observe $(Y, \Delta_B, X, \Delta, Z)$. Further we assume that (U, V) and (C_B, C) are conditionally independent given covariates Z . Under this scenario,

Table 1. Comparisons among the proposed estimators under the residual time model $m(t; z) = (At + B) \exp(0.5z_1 + 0.5z_2)$

(a) $A = 0.1, B = 0.5$					
Estimator	Parameter	Bias $\times 10^3$	SSE $\times 10^3$	SEE $\times 10^3$	Coverage (%)
$\hat{\beta}_U$	β_1	2	151	152	94.8
	β_2	-5	260	254	94.7
$\hat{\beta}_V$	β_1	2	163	167	95.8
	β_2	<1	281	284	95.3
$\hat{\beta}_{UV}$	β_1	-2	117	117	94.8
	β_2	-6	202	198	94.3
(b) $A = 0.05, B = 1$					
Estimator	Parameter	Bias $\times 10^3$	SSE $\times 10^3$	SEE $\times 10^3$	Coverage (%)
$\hat{\beta}_U$	β_1	1	160	152	94.0
	β_2	1	267	254	93.5
$\hat{\beta}_V$	β_1	-5	184	183	94.7
	β_2	-9	314	313	94.7
$\hat{\beta}_{UV}$	β_1	-4	124	118	94.2
	β_2	-7	212	199	92.4

SSE, the sampling standard deviation; SEE, the sample mean of standard error estimates; Coverage, the empirical coverage of approximate 95% confidence intervals.

we could modify (4) to estimate β using the paired survival data $\{(Y_i, \Delta_{Bi}), (X_i, \Delta_i) : i = 1, \dots, n\}$, where $\hat{\gamma} = -\hat{\beta}$ is estimated by solving $\tilde{S}_{PL}(\gamma) = 0$, where

$$\tilde{S}_{PL}(\gamma) = \sum_{i=1}^n \Delta_{Bi} \left\{ Z_i - \frac{\tilde{S}^{(1)}(\gamma, Y_i)}{\tilde{S}^{(0)}(\gamma, Y_i)} \right\} + \Delta_i \left\{ Z_i - \frac{\tilde{S}^{(1)}(\gamma, X_i)}{\tilde{S}^{(0)}(\gamma, X_i)} \right\},$$

with $\tilde{S}^{(k)}(\gamma, t) = n^{-1} \sum_{i=1}^n Z_i^{\otimes k} \{I(Y_i \geq t) + I(X_i \geq t)\} \exp(\gamma^T Z_i)$ ($k = 0, 1, 2$).

The target parameter β can also be estimated when either backward or forward recurrence times are being collected. The former type of data can be collected in cross-sectional surveys without further follow-up (Allison, 1985). The latter type of data is collected when time of disease incidence cannot be ascertained retrospectively. The target parameter β from the mean residual lifetime model can be estimated by partial likelihood methods using only backward or forward recurrence times.

3. NUMERICAL STUDIES

We examined finite sample properties of the proposed estimator by simulation. For each simulation scenario, 5000 independent datasets were generated. For each dataset, we independently generated a Bernoulli variable Z_1 with $p = 0.5$ and a $U(0, 1)$ variable Z_2 . Conditioning on Z_1 and Z_2 , T followed a mean residual life model $m_T(t) = (At + B) \exp(\beta_1 Z_1 + \beta_2 Z_2)$. To obtain a length-biased sample, we generated random disease incidence time U^0 from a $U(0, 30)$ distribution; an observation was in the cross-sectional sample if $-U^0 + T \geq 0$. Data were generated until the cross-sectional sample had 100 observations. We generated residual censoring time C from a $U(0, 5)$ distribution. We considered three estimators of $\beta = (\beta_1, \beta_2)^T$. The estimator $\hat{\beta}_U$ maximized the partial likelihood using only backward recurrence times. The $\hat{\beta}_V$ maximized the partial likelihood using censored forward recurrence times. The estimator $\hat{\beta}_{UV}$ was the solution of $S_{PL}(-\beta) = 0$. Table 1 shows that all the estimators have small bias and that using both backward and forward recurrence times greatly improves efficiency.

When $A = 0$, the proportional mean residual life model is identical to a proportional hazards model with a target parameter of the same magnitude but a different sign. Table 2 compares the proposed estimator $\hat{\beta}_{UV}$ with the partial likelihood estimator for left-truncated and right-censored data, $\hat{\beta}_{PL}$. The proposed estimator

Table 2. Comparison of the proposed estimator $\hat{\beta}_{UV}$ with the partial likelihood estimator $\hat{\beta}_{PL}$

(a) $U(0, 5)$ censoring					
Estimator	Parameter	Bias $\times 10^3$	SSE $\times 10^3$	SEE $\times 10^3$	Coverage (%)
β_{UV}	β_1	7	113	113	95.2
	β_2	-2	194	191	94.8
β_{PL}	β_1	13	181	178	95.0
	β_2	1	291	300	95.6
(b) No censoring					
Estimator	Parameter	Bias $\times 10^3$	SSE $\times 10^3$	SEE $\times 10^3$	Coverage (%)
β_{UV}	β_1	2	107	107	94.8
	β_2	-9	181	178	94.9
β_{PL}	β_1	7	160	156	94.3
	β_2	-5	251	258	95.4

SSE, the sampling standard deviation; SEE, the sample mean of standard error estimates; Coverage, the empirical coverage of approximate 95% confidence intervals.

has better efficiency than the usual partial likelihood estimator under length-biased sampling because it takes advantage of the length-biased structure; the usual risk-set-based methods for left-truncated and right-censored data do not. In the presence of right censoring, the sampling variability of $\hat{\beta}_{UV}$ is 39% and 44% of the sampling variability of $\hat{\beta}_{PL}$ for β_1 and β_2 respectively. The results echo similar findings for nonparametric estimation in Wang (1989) and Asgharian et al. (2002). A reviewer pointed out that U and V are independent when T is exponentially distributed and that the asymptotic variance of $\hat{\beta}_{UV}$ should be half of the asymptotic variance of $\hat{\beta}_{PL}$ if there is no censoring. In small sample simulations without censoring, the sampling variances of $\hat{\beta}_{UV}$ are 45% and 52% of the sampling variances of $\hat{\beta}_{PL}$ for β_1 and β_2 , respectively, close to the theoretical prediction.

4. DISCUSSION

For the usual right-censored data collected from an incident population, semiparametric models on the mean residual life function are identifiable only under a strong assumption that the right tail of the survival distribution is fully identified. This assumption is violated when administrative censoring is present after a limited study period. Cross-sectional sampling alleviates this common identifiability problem in residual time regression models. With the induced proportional hazards structure, the mean residual life ratio is identifiable for length-biased survival data when follow-up is limited. In fact, the parameter β in model (1) can even be estimated by partial likelihood methods when we collect only backward or forward recurrence times. The proposed methodology is applicable when one adopts a retrospective study based on a cross-sectional sample and to cases where onset of disease cannot be ascertained retrospectively. For prevalent cohort data, Wang et al. (1993) advocated the use of proportional hazards models for population survival time over proportional hazards models for forward recurrence time, because the former has a direct interpretation for the incident study population. Under length-biased sampling, the relative hazard for forward recurrence time indeed has a useful interpretation as the inverse of the relative mean residual life for the underlying population.

A reviewer mentioned that the joint distribution of (U, V) depends only on $U + V$; see Vardi (1989) who considered nonparametric likelihood estimation of the survival function for length-biased right-censored data. However, under the semiparametric proportional mean residual life model, the likelihood function does not lead to tractable analysis. In fact, it is unclear to the authors how the functional nuisance parameter $m_0(t)$ can be eliminated. Moreover, likelihood estimation for the semiparametric proportional mean residual lifetime model is an open problem even for standard survival data in the absence of length-biased sampling. However, using the proportional hazards structure of U and V , a functional nuisance

parameter can be eliminated using risk-set arguments, and censoring in both directions can also be handled naturally.

ACKNOWLEDGEMENT

The authors thank the editor, an associate editor, two reviewers and Prof. Mary Lou Thompson for their helpful comments and suggestions, which greatly improved this paper.

REFERENCES

- ALLISON, P. D. (1985). Survival analysis of backward recurrence times. *J. Am. Statist. Assoc.* **80**, 315–22.
- ASGHARIAN, M., M'LAN, C. E. & WOLFSON, D. B. (2002). Length-biased sampling with right censoring: An unconditional approach. *J. Am. Statist. Assoc.* **97**, 201–10.
- BERGERON, P. J., ASGHARIAN, M. & WOLFSON, D. B. (2008). Covariate bias induced by length-biased sampling of failure times. *J. Am. Statist. Assoc.* **103**, 737–42.
- CANFIELD, R. H. (1941). Application of the line interception method in sampling range vegetation. *J. Forestry* **39**, 388–94.
- CHEN, Y. Q. (2010). Semiparametric regression in size-biased sampling. *Biometrics* **66**, 149–58.
- CHEN, Y. Q. & CHENG, S. (2005). Semiparametric regression analysis of mean residual life with censored survival data. *Biometrika* **92**, 19–29.
- COX, D. R. (1962). *Renewal Theory*. London: Methuen.
- COX, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*, Ed. N. L. Johnson and H. Smith, pp. 506–27. New York: Wiley.
- HUANG, C.-Y. & QIN, J. (2011). Nonparametric estimation for length-biased and right-censored data. *Biometrika* **98**, 177–86.
- LEE, E. W., WEI, L. J. & AMATO, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, Ed. J. P. Klein and P. K. Goel, pp. 237–47. Dordrecht: Kluwer Academic Publishers.
- MAGULURI, G. & ZHANG, C.-H. (1994). Estimation in the mean residual life regression model. *J. R. Statist. Soc. B* **56**, 477–89.
- MANDEL, M. & RITOV, Y. (2010). The accelerated failure time model under biased sampling. *Biometrics* **66**, 1306–8.
- NEYMAN, J. (1955). Statistics: servant of all science. *Science* **122**, 401–6.
- OAKES, D. & DASU, T. (1990). A note on residual life. *Biometrika* **77**, 409–10.
- QIN, J. & SHEN, Y. (2010). Statistical methods for analyzing right-censored length-biased data under Cox model. *Biometrics* **66**, 382–92.
- SHEN, Y., NING, J. & QIN, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *J. Am. Statist. Assoc.* **104**, 1192–202.
- SPIEKERMAN, C. F. & LIN, D. Y. (1998). Marginal regression models for multivariate failure time data. *J. Am. Statist. Assoc.* **93**, 1164–75.
- VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616–20.
- VARDI, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* **76**, 751–61.
- WANG, M.-C. (1989). A semiparametric model for randomly truncated data. *J. Am. Statist. Assoc.* **84**, 742–8.
- WANG, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *J. Am. Statist. Assoc.* **86**, 130–43.
- WANG, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika* **83**, 343–54.
- WANG, M.-C., BROOKMEYER, R. & JEWELL, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics* **49**, 1–11.
- WICKSELL, S. D. (1925). The corpuscle problem: a mathematical study of a biometric problem. *Biometrika* **17**, 84–99.
- ZELEN, M. & FEINLEIB, M. (1969). On the theory of screening for chronic diseases. *Biometrika* **56**, 601–14.

[Received January 2011. Revised July 2012]