Biostatistics (2016), **0**, 0, *pp.* 1–10
doi:10.1093/biostatistics/biosts-14295-File002

Supplementary materials to "Accelerated failure time model under general biased sampling scheme"

JANE PAIK KIM
Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford CA 94305
janepkim@stanford.edu

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
zying@stat.columbia.edu

Introduction

This set of supplementary materials describes our simulation study in details in Section A. In particular, we present our numerical performances as comparisons with other existing methods. It also covers, in Sections B and C, proofs for some technical results that are used in the main text.

A. Simulations

Simulation studies were conducted to assess the effectiveness of the proposed method. In our simulations, we considered the linear regression model (2.1) of $\log T = -\beta' Z + \epsilon$ where the random variable ϵ was assumed to follow a standard Normal distribution with the density function $(2\pi)^{-1/2}\exp\{-x^2/2\}$. Covariates Z_1 and Z_2 were generated from uniform (0,1) that are independent of each other. The parameters

*To whom correspondence should be addressed.

$\beta_0 = (\beta_1, \beta_2)$ were chosen to be -1.0 and 1.0 respectively. The censoring time was generated by $e^{a+0.5U}$, where U is a standard uniform variable. Values of a were set to attain the desired censoring proportion.

The four biased sampling designs that are under the scope of the proposed framework include (i) length-biased sampling, (ii) case-cohort design, (iii) generalised case-cohort design and (iv) a combo biased sampling: case-cohort analysis on length-biased data. In all of the following sets of simulations, 500 resamplings were preformed in order to obtain the estimated standard error of the estimates $\hat{\beta}$. Log-rank and Gehan weights were chosen for illustration purposes.

For length-biased sampling, given the data generated by $q(\tilde{T}, \Delta)$, we resampled those units with $U_i \leqslant \tilde{T}_i/\gamma$, where U_i follows the uniform distribution and γ is constant which is larger than $\tilde{T}_i$ for all $i = 1, \ldots, n$. Computation was conducted on the resampled individuals of sizes 100 and 200. Simulations were based on 500 replications. Results are presented in Table 1. We also compared our performance with that suggested in Mandel and Ritov (2010) for observations. It can be observed from our numerical results that Mandel and Ritov (2010)'s approach does not significantly outperform the proposed method and the edge diminishes as the sample size grows. When the censoring rate increases, the biases of the estimator given by Mandel and Ritov (2010) inflate; this agrees with our intuition because their method is designed for handling life-time data with no censoring.

For the case-cohort design, a full cohort of sample size 3,000 was generated and then case-cohort samples were selected from each full cohort by selecting from cases with a probability of p such that about two thirds of the selected samples in the subcohort are controls. The average sample size of a subcohort is 1,000 with censoring rates 0.8 or 0.9, which mimics a rare-disease study. Estimates computed were based on 500 simulations. The numerical results are summarised in Table 2. For comparison, we applied the methodology of Nan, Kalbfleisch and Yu (2009) to the same simulated data. We randomly drew 10% of the samples to form a subsample regardless of individuals' censoring status. That corresponds to the predictable weight as discussed in Nan, Kalbfleisch and Yu (2009). It can be seen from Table 2 that the proposed method is comparatively more efficient especially with high censoring cases.

For generalised case-cohort design (see Kim et al., 2013), a full cohort of sample size 3,000 was generated and then case-cohort samples were selected from each full cohort by selecting from cases with a probability of $p_i = 1 - \{1 + \exp(1 + \tilde{T}_i)\}^{-1}$ and controls with a probability of $p_i = 1 - \{1 + \exp\{-3 + 2\tilde{T}_i\}\}^{-1}$. The average size for a subcohort is 1,000, with one third of samples are cases. The censoring rates chosen include 0.8 and 0.9. 500 replications were created to assess the performance. Readers are referred to Table 3 for the corresponding numerical performance. Comparisons between our procedure and that of Nan, Kalbfleisch and Yu (2009) with predictable weights for observed failures and censored subjects reveal that our procedure yields more efficient estimates as reflected by the smaller of SE's (and SEE's) values.

For the case-cohort design on length biased data (combo), data were generated in the same way as in length-biased sampling case after which a case-cohort sampling was applied. Same as the previous two studies, one third, on average, of the samples selected were cases. Table 4 tabulates the simulation results. Since Nan, Kalbfleisch and Yu (2009) method is not decided to handle this type of biased samples, we can see from the results that their method leads to large biases as well as poor estimates for the standard errors and empirical coverage probabilities.

The results presented in Tables 1-4 reveal that the proposed estimators of the regression parameters are virtually unbiased for all the cases. Furthermore, the standard error estimators depict well the true variability of the parameter estimators. Both 90% and 95% empirical coverage probabilities are close to the nominal levels.

B. Monotonicity of U_n for length-biased sampling

Recall that, as discussed in Fygenson and Ritov (1994), a function $W(\beta) : \mathbb{R}^d \to \mathbb{R}^d$ is called a monotone non-decreasing field if, for any $\beta, \xi \in \mathbb{R}^d, \xi' W(\beta + x\xi)$ is a monotone non-decreasing function of the real variable x. For any monotone non-decreasing field, $W_n(\beta)$, all the generalised solutions of $W_n(\beta) = 0$ belongs to a convex set whose diameter is $\mathcal{O}(n^{-1})$. In other words, due to the monotonicity of the

estimating equation, the set of its generalised solutions is convex, and it is relatively easy to locate an estimator and to establish its properties. In fact, the estimator is $\sqrt{n}$-consistent and asymptotically normal under certain regularity conditions. If we define the right-hand side of (2.12) as $U_n(\beta)$, we can write

$$\begin{aligned}
&\xi' U_n(\beta + x\xi) \\
&= \xi' \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (Z_i - Z_j) \left\{ \frac{\tilde{T}_i}{\tilde{T}_j} e^{(\beta + x\xi)'(Z_i - Z_j)} \right\} \right. \\
&\quad \left. \times I \left\{ \log \tilde{T}_i + (\beta + x\xi)' Z_i \leqslant \log \tilde{T}_j + (\beta + x\xi)' Z_j \right\} \right] \\
&:= \xi' \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (Z_i - Z_j) \left\{ \frac{\tilde{T}_i}{\tilde{T}_j} e^{(\beta + x\xi)'(Z_i - Z_j)} \right\} I\{e_i(\beta + x\xi) \leqslant e_j(\beta + x\xi)\} \right].
\end{aligned}$$

With a slight abuse of notation, we consider

$$\begin{aligned}
&\frac{\partial}{\partial x} \xi' U_n(\beta + x\xi) \\
&= \xi' \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (Z_i - Z_j) \frac{\tilde{T}_i}{\tilde{T}_j} \frac{\partial}{\partial x} \left[e^{(\beta + x\xi)'(Z_i - Z_j)} I \left\{ e_i(\beta + x\xi) \leqslant e_j(\beta + x\xi) \right\} \right] \right) \\
&\quad + \xi' \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i (Z_i - Z_j) \left\{ \frac{\tilde{T}_i}{\tilde{T}_j} e^{(\beta + x\xi)'(Z_i - Z_j)} \right\} \frac{\partial}{\partial x} I \left\{ e_i(\beta + x\xi) \leqslant e_j(\beta + x\xi) \right\} \right] \\
&= I + II, \quad \text{say.}
\end{aligned}$$

It can be observed that I is non-negative and so is II as shown in Fygenson and Ritov (1994).

C. Derivation of asymptotic results

We assume the following regularity conditions that are similar to those in Ying (1993) and Kim et al. (2013):

1. The covariates are uniformly bounded, and without loss of generality, we may assume that $\sup_i \|Z_i\| \leqslant 1$.

2. The error density f_ϵ and its derivative f'_ϵ are bounded, satisfying that $\int (f'_\epsilon(t)/f(t))^2 f(t) dt < \infty$.

3. The matrix A_G is non-singular.

4. $E\left[\int_{-\infty}^{\infty} S_{\omega}^{(0)}(\hat{\beta}_G;t)\{Z_i - \bar{Z}_{\omega}(\hat{\beta}_G;t)dN_i(\hat{\beta}_G;t)(\xi_i - 1)\}\right]^2 < \infty.$

Conditions 1 and 2 correspond to those imposed in Ying (1993) so as to ensure the asymptotic linearity of the weighted log-rank estimating function. Condition 3 can be easily satisfied if the vector of covariates does not lie in a lower dimensional hyperplane that leads to singularity. Condition 4 is a mild assumption on the weight function $\omega(\cdot)$ that allows the convergences in distribution of U_G and U_G^* due to the central limit theorem.

The terms $n^{-1}L_G$ and $n^{-1}L_G^*$, for both bias-sampling settings with or without time component involved, are convex functions. Due to the strong law of large numbers, both of them converge almost surely to the same limiting function. Assuming that its second derivative at β_0, the true value of β, is A_G is non-singular, the limiting function has a unique minimiser β_0. It follows that almost surely $\hat{\beta}_G \to \beta_0$ and $\hat{\beta}_G^* \to 0$ as $n \to \infty$. By applying similar arguments of Theorem 2 of Ying (1993), we can write

$$U_G(\hat{\beta}_G) = U_G(\beta_0) + nA_G(\hat{\beta}_G - \beta_0) + o(n^{\frac{1}{2}} + n\|\hat{\beta}_G - \beta_0\|), \quad a.s. \tag{0.1}$$

and

$$U_G^*(\hat{\beta}_G^*) = U_G^*(\beta_0) + nA_G(\hat{\beta}_G^* - \beta_0) + o(n^{\frac{1}{2}} + n\|\hat{\beta}_G^* - \beta_0\|), \quad a.s.. \tag{0.2}$$

Both functions U_G^* and U_G have the same asymptotic slope matrix A_G in (4.1) and (4.2) where the latter follows from the argument presented in Jin et al. (2006). The estimators $\hat{\beta}_G$ and $\hat{\beta}_G^*$ are consistent. Denote $\mathcal{F}$ the σ-field generated by the original data $(\tilde{T}_i, \Delta_i, Z_i, \omega_i)_{i=1,\ldots,n}$. Both $n^{-\frac{1}{2}}U_G(\hat{\beta}_G)$ and $n^{-\frac{1}{2}}U_G^*(\hat{\beta}_G^*)$, conditional on $\mathcal{F}$, are normalised sums of independent zero-mean random vectors. The multivariate central limit theorem implies that $n^{-\frac{1}{2}}U_G^*(\hat{\beta}_G^*)$ converges in distribution to $\mathcal{N}(0, B_G)$. It then follows from (0.2) that the conditional distribution of $n^{-\frac{1}{2}}U_G^*(\hat{\beta}_G^*)$ given $\mathcal{F}$ converges almost surely tto $\mathcal{N}(0, A_G^{-1}B_GA_G^{-1})$, which is the limiting distribution of $n^{\frac{1}{2}}(\hat{\beta}_G - \beta_0)$. This completes the proof.

REFERENCES

FYGENSON, M. AND RITOV, Y. (1994). Monotone estimating equations for censored data. *Ann. Statist.* **22**, 732–46.

JIN, Z., LIN, D. Y. AND YING, Z. (2006). Rank regression analysis of multivariate failure time data based on marginal linear models. *Scand. J. Stat.* **33**, 1–23.

KIM, J. P., LU, W., SIT, T. AND YING, Z. (2013). A unified approach to semiparametric transformation models under generalized biased sampling schemes. *J. Am. Statist. Assoc.* **108**, 217-227.

MANDEL, M. AND RITOV, Y. (2010). The accelerated failure time model under biased sampling. *Biometrics* **66**, 1306–8.

NAN, B., KALBFLEISCH, J.D. AND YU, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Ann. Statist.* **37**, 2351–76.

YING, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21**, 76–99.

n	Censoring	Weight/Method	Parameters	Bias	SE	SEE	90% ECP	95% ECP
100	0%	Gehan	β_{01}	0.003	0.419	0.430	0.911	0.961
			β_{02}	0.020	0.369	0.349	0.930	0.957
		Log-rank	β_{01}	0.008	0.301	0.244	0.920	0.952
			β_{02}	0.007	0.311	0.242	0.896	0.928
		Mandel and Ritov (2010)	β_{01}	0.034	0.394	0.350	0.843	0.933
			β_{02}	0.018	0.333	0.350	0.890	0.936
	15%	Gehan	β_{01}	-0.014	0.459	0.490	0.922	0.958
			β_{02}	0.015	0.472	0.490	0.896	0.944
		Log-rank	β_{01}	-0.035	0.507	0.549	0.880	0.920
			β_{02}	0.034	0.485	0.545	0.920	0.934
		Mandel and Ritov (2010)	β_{01}	0.041	0.364	0.352	0.876	0.940
			β_{01}	-0.096	0.302	0.350	0.940	0.972
	25%	Gehan	β_{01}	-0.018	0.506	0.537	0.916	0.946
			β_{02}	0.043	0.510	0.529	0.906	0.946
		Log-rank	β_{01}	0.035	0.541	0.574	0.890	0.938
			β_{02}	0.012	0.556	0.584	0.919	0.952
		Mandel and Ritov (2010)	β_{01}	0.189	0.437	0.363	0.754	0.850
			β_{01}	-0.127	0.360	0.357	0.882	0.928
200	0%	Gehan	β_{01}	0.001	0.259	0.280	0.917	0.956
			β_{02}	-0.047	0.258	0.279	0.879	0.939
		Log-rank	β_{01}	0.011	0.339	0.370	0.924	0.964
			β_{02}	-0.013	0.337	0.366	0.914	0.960
		Mandel and Ritov (2010)	β_{01}	-0.049	0.265	0.245	0.886	0.964
			β_{02}	0.020	0.266	0.245	0.942	0.945
	15%	Gehan	β_{01}	-0.011	0.313	0.332	0.924	0.952
			β_{02}	0.017	0.334	0.331	0.882	0.928
‘		Log-rank	β_{01}	-0.005	0.332	0.367	0.908	0.960
			β_{02}	0.009	0.313	0.336	0.916	0.950
		Mandel and Ritov (2010)	β_{01}	0.104	0.233	0.250	0.872	0.936
			β_{02}	-0.058	0.253	0.249	0.856	0.941
	25%	Gehan	β_{01}	0.004	0.460	0.526	0.928	0.966
			β_{02}	0.025	0.473	0.523	0.918	0.956
		Log-rank	β_{01}	0.015	0.535	0.572	0.902	0.944
			β_{02}	0.037	0.567	0.515	0.914	0.958
		Mandel and Ritov (2010)	β_{01}	0.106	0.227	0.260	0.882	0.931
			β_{02}	-0.122	0.238	0.250	0.864	0.937

Table 1. Estimates and standard errors for regression parameters β based on 500 replications and 500 perturbed resampling on length-biased data. Bias, SE, SEE and $x\%$ECP are defined as the difference between the estimated and the true parameter values, the standard error estimated, the standard error of the resampled estimated parameter values as well as the $x\%$ empirical coverage probability respectively.

n	Censoring	Weight/Method	Parameters	Bias	SE	SEE	90% ECP	95% ECP
100	80%	Gehan	β_{01}	-0.026	0.431	0.442	0.904	0.938
			β_{02}	0.028	0.449	0.444	0.890	0.940
		Log-rank	β_{01}	0.067	0.353	0.359	0.876	0.932
			β_{02}	-0.075	0.364	0.353	0.872	0.932
		Nan et al. (2009)	β_{01}	0.023	0.491	0.489	0.856	0.931
			β_{02}	0.077	0.534	0.498	0.891	0.934
	90%	Gehan	β_{01}	0.045	0.363	0.368	0.884	0.936
			β_{02}	-0.087	0.358	0.359	0.878	0.928
		Log-rank	β_{01}	0.035	0.356	0.366	0.884	0.944
			β_{02}	-0.039	0.342	0.368	0.910	0.946
		Nan et al. (2009)	β_{01}	0.088	0.655	0.608	0.878	0.910
			β_{02}	0.003	0.650	0.683	0.868	0.924
200	80%	Gehan	β_{01}	-0.007	0.290	0.291	0.904	0.948
			β_{02}	-0.012	0.280	0.293	0.932	0.964
		Log-rank	β_{01}	-0.060	0.346	0.338	0.888	0.938
			β_{02}	0.044	0.329	0.335	0.903	0.947
		Nan et al. (2009)	β_{01}	-0.071	0.383	0.364	0.870	0.930
			β_{02}	0.048	0.368	0.359	0.876	0.941
	90%	Gehan	β_{01}	0.001	0.256	0.242	0.876	0.922
			β_{02}	-0.008	0.239	0.244	0.880	0.936
		Log-rank	β_{01}	-0.037	0.240	0.243	0.882	0.926
			β_{02}	-0.032	0.231	0.242	0.888	0.940
		Nan et al. (2009)	β_{01}	-0.039	0.478	0.424	0.875	0.911
			β_{02}	0.019	0.463	0.410	0.852	0.934

Table 2. Estimates and standard errors for regression parameters β based on 500 replications and 500 perturbed resampling on case-cohort data. Bias, SE, SEE and $x\%$ECP are defined as the difference between the estimated and the true parameter values, the standard error estimated, the standard error of the resampled estimated parameter values as well as the $x\%$ empirical coverage probability respectively.

n	Censoring	Weight/Method	Parameters	Bias	SE	SEE	90% ECP	95% ECP
100	80%	Gehan	β_{01}	0.043	0.455	0.424	0.876	0.920
			β_{02}	-0.026	0.427	0.430	0.888	0.948
		Log-rank	β_{01}	-0.003	0.425	0.439	0.912	0.956
			β_{02}	-0.017	0.431	0.435	0.894	0.944
		Nan et al. (2009)	β_{01}	-0.019	0.422	0.406	0.910	0.961
			β_{02}	-0.003	0.458	0.399	0.893	0.964
	90%	Gehan	β_{01}	0.012	0.367	0.383	0.904	0.968
			β_{02}	0.025	0.401	0.388	0.882	0.944
		Log-rank	β_{01}	0.041	0.382	0.387	0.886	0.938
			β_{02}	0.009	0.382	0.390	0.892	0.936
		Nan et al. (2009)	β_{01}	0.011	0.443	0.416	0.880	0.965
			β_{02}	0.008	0.401	0.408	0.901	0.968
200	80%	Gehan	β_{01}	0.011	0.302	0.307	0.906	0.948
			β_{02}	-0.008	0.293	0.304	0.912	0.960
		Log-rank	β_{01}	0.012	0.304	0.310	0.896	0.950
			β_{02}	-0.014	0.296	0.308	0.892	0.962
		Nan et al. (2009)	β_{01}	0.004	0.325	0.328	0.892	0.919
			β_{02}	0.003	0.312	0.316	0.884	0.932
	90%	Gehan	β_{01}	-0.033	0.259	0.286	0.910	0.955
			β_{02}	0.016	0.284	0.283	0.900	0.945
		Log-rank	β_{01}	0.002	0.304	0.295	0.882	0.938
			β_{02}	-0.024	0.284	0.295	0.900	0.948
		Nan et al. (2009)	β_{01}	0.047	0.328	0.341	0.832	0.939
			β_{02}	0.001	0.342	0.344	0.839	0.881

Table 3. Estimates and standard errors for regression parameters β based on 500 replications and 500 perturbed resampling on generalised case-cohort data. Bias, SE, SEE and $x\%$ECP are defined as the difference between the estimated and the true parameter values, the standard error estimated, the standard error of the resampled estimated parameter values as well as the $x\%$ empirical coverage probability respectively.

n	Censoring	Weight/Method	Parameters	Bias	SE	SEE	90% ECP	95% ECP
100	80%	Gehan	β_{01}	-0.020	0.520	0.556	0.884	0.938
			β_{02}	0.016	0.572	0.551	0.890	0.932
		Log-rank	β_{01}	0.001	0.542	0.577	0.904	0.954
			β_{02}	-0.011	0.547	0.575	0.884	0.942
		Nan et al. (2009)	β_{01}	-0.131	1.085	1.548	0.830	0.920
			β_{02}	0.206	1.059	2.067	0.890	0.940
	90%	Gehan	β_{01}	-0.027	0.487	0.490	0.914	0.950
			β_{02}	0.009	0.458	0.484	0.914	0.958
		Log-rank	β_{01}	-0.025	0.517	0.519	0.906	0.940
			β_{02}	0.008	0.493	0.511	0.908	0.956
		Nan et al. (2009)	β_{01}	-0.355	1.989	3.028	0.880	0.951
			β_{02}	0.079	1.112	2.541	0.840	0.939
200	80%	Gehan	β_{01}	-0.018	0.379	0.385	0.888	0.936
			β_{02}	0.025	0.367	0.390	0.914	0.946
		Log-rank	β_{01}	0.015	0.392	0.404	0.906	0.958
			β_{02}	-0.016	0.390	0.406	0.908	0.954
		Nan et al. (2009)	β_{01}	-0.108	0.728	0.977	0.890	0.915
			β_{02}	0.151	0.715	1.179	0.880	0.931
	90%	Gehan	β_{01}	-0.024	0.341	0.351	0.898	0.944
			β_{02}	0.010	0.335	0.346	0.906	0.960
		Log-rank	β_{01}	-0.001	0.346	0.375	0.892	0.950
			β_{02}	0.017	0.358	0.373	0.910	0.968
		Nan et al. (2009)	β_{01}	-0.157	0.597	0.755	0.899	0.932
			β_{02}	0.156	0.684	0.750	0.909	0.929

Table 4. Estimates and standard errors for regression parameters β based on 500 replications and 500 perturbed resampling on case cohort sampling on length-biased (combo) data. Bias, SE, SEE and $x\%$ECP are defined as the difference between the estimated and the true parameter values, the standard error estimated, the standard error of the resampled estimated parameter values as well as the $x\%$ empirical coverage probability respectively.