 DOI: 10.1111/j.1541-0420.2009.01366.x

The Accelerated Failure Time Model Under Biased Sampling

Micha Mandel* **and Ya'akov Ritov****

The Hebrew University, Mount Scopus, Jerusalem, 91905, Israel
**email:* msmic@huji.ac.il
***email:* yaacov@mscc.huji.ac.il

Summary. Chen (2009, *Biometrics*) studies the semi-parametric accelerated failure time model for data that are size biased. Chen considers only the uncensored case and uses hazard-based estimation methods originally developed for censored observations. However, for uncensored data, a simple linear regression on the log scale is more natural and provides better estimators.

Key words: Length bias; Size bias; Survival; Weighted distribution.

1. The Model

Let $\xi > 0$ and Z be independent random variables, and let $X = \{c(Z)\}^{-1}\xi$, where $c(\cdot) > 0$ on the support of Z. Here Z is a covariate (or a vector of covariates) that determines the scale c of a lifetime variable X: $f_{X|Z}(x|z) = c(z)f_\xi\{xc(z)\}$, where f_R denotes the density function of a random variable R. This model is a generalization of the accelerated failure time (AFT) model considered by Chen (2009) in which $c(z) = e^{-\beta z}$ for a scalar β.

Obviously, $\log(X) = -\log\{c(Z)\} + \log(\xi)$ has the form of a standard regression model with an additive noise $\log(\xi)$; such models have been extensively studied under various assumptions on $c(\cdot)$ and f_ξ. For the AFT model, $\log(X) = -\beta Z + \log(\xi)$ has the familiar linear regression form with a constant term $E\{\log(\xi)\}$, for which least squares is the standard inference approach.

Often, survival data are obtained by biased sampling in which observations are over or under represented according to their outcome. For example, Wang (1996) and Ghosh (2008) study the proportional hazard model, and Bergeron, Asgharian, and Wolfson (2008) study parametric models under size biased sampling. A general biased sampling model assumes observations from the density

$$\frac{w(x,z)f_{X,Z}(x,z)}{E\{w(X,Z)\}} = \frac{w(x,z)f_{X|Z}(x|z)f_Z(z)}{E\{w(X,Z)\}}, \qquad (1)$$

where $w > 0$ is a weight function satisfying $E\{w(X,Z)\} < \infty$.

A common example of biased sampling arises in situations like the following: Patients arrive to several hospitals according to independent Poisson processes. A research is planned in order to study the effect of a hospital-specific covariate Z on the total hospitalization duration, ξ. Conditionally on Z, the hospitalization durations are independent of the arrival processes. The design is of a cross-sectional sampling nature, where the sample comprises patients staying in the hospitals at a given fixed time point. We distinguish between two cases: In the first, all patients staying in the hospitals at sampling time are sampled, and in the second, only one patient is taken from each hospital. The likelihood of the first sampling design is a product of terms as (1), but under the second design, the likelihood is a product of terms as

$$f_{X|Z,\mathrm{BS}}(x|z) = \frac{w(x,z)f_{X|Z}(x|z)}{E\{w(X,Z)|Z=z\}} = \frac{w(x,z)c(z)f_\xi\{xc(z)\}}{E\{w(X,Z)|Z=z\}}, \qquad (2)$$

where BS stands for biased sampling. This distinction may be important for inference, but it is often overlooked. In fact, straightforward calculations show that the density of $X|Z$ under (1) is exactly (2), so that the difference between the two designs is in the sampling distribution of the covariate (see further discussion below).

Equation (2) shows that, in general, the conditional distribution of X given the covariate Z under biased sampling does not belong to the same generalized AFT model defined above, and special methods are required for fitting such models. However, consider independent realizations, (X_i^*, Z_i^*), from the weighted density $f_{X^*,Z^*}(x,z) = x^\alpha w(z) f_{X,Z}(x,z)/E(X^\alpha w(Z))$, for some weight function $w(\cdot) > 0$ and constant α. It is easy to see that

$$f_{X^*|Z^*}(x|z) = \frac{x^\alpha c(z) f_\xi\{xc(z)\}}{E(X^\alpha|Z=z)} = \frac{x^\alpha c(z) f_\xi\{xc(z)\}}{[c(z)]^{-\alpha}E(\xi^\alpha)},$$

so that the model after biased sampling remains in the same family with $X^* = [c(Z^*)]^{-1}\xi^*$ and $f_{\xi^*}(t) = t^\alpha f_\xi(t)/E(\xi^\alpha)$, and the same procedures for estimating $c(z)$, used for data from $f_{X,Z}$, can be used when data are obtained by biased sampling.

The invariance property of the order of weighting and scaling is given by Chen (2009) as Property 1 for the special AFT model, $c(z) = e^{\beta z}$, under size biased sampling ($\alpha = 1$ and $w \equiv 1$). The case $\alpha = 1$ arises in cross-sectional sampling under steady state assumptions, see van Es, Klaassen, and Oudshoorn (2000), and it is the most interesting case in

Table 1
Comparison of standard errors (SE) *and coverage probabilities of* 95% *confidence intervals* (Cover) *of least squares* (LS) *and Chen's method under the log-normal and the Weibull models for* ξ

			Log-normal		Weibull	
β	n	Estimate	SE	Cover	SE	Cover
0	50	Chen	0.6533	0.955	0.5084	0.951
		LS	0.4992	0.950	0.2413	0.957
	200	Chen	0.3261	0.952	0.2603	0.950
		LS	0.2461	0.950	0.1190	0.947
	500	Chen	0.2477	0.950	0.1619	0.949
		LS	0.1552	0.950	0.0750	0.950
1	50	Chen	0.5227	0.955	0.4603	0.951
		LS	0.5126	0.950	0.2478	0.954
	200	Chen	0.2603	0.951	0.2443	0.949
		LS	0.2523	0.950	0.1220	0.950
	500	Chen	0.1654	0.951	0.1447	0.951
		LS	0.1591	0.950	0.0769	0.953

practice. The assumption that $w \equiv 1$ is also very reasonable, as one can replace $f_Z(z)$ with $w(z)f_Z(z)/E(w(Z))$ without affecting the conditional distribution of X given Z.

We note that $f_{Z^*}(z) = E(X^\alpha|Z=z)f_Z(z)/E(X^\alpha) = \{c(z)\}^{-\alpha} f_Z(z)/E[\{c(Z)\}^{-\alpha}]$, which in general differs from $f_Z(z)$. Therefore, the likelihood L_2 in Chen (2009) is not a full likelihood, but conditional on the values of Z^*. If f_Z is known, then there may be a loss of information in using the conditional likelihood instead of the full likelihood (Bergeron et al., 2008). However, if f_Z is unknown, there is no loss of information, and the estimator based on the conditional likelihood is efficient.

To estimate β in the AFT model, Chen (2009) uses hazard-based methods originally tailored to censored data. The discussion above shows that $\log(X^*) = -\beta Z^* + \log(\xi^*)$, and therefore, the least squares approach can be used for estimation. Note that $\log(\xi^*)$ and $\log(\xi)$ are not identically distributed and do not have the same mean, and hence the constant term in the regression of $\log(X^*)$ on Z^* (i.e., $E\{\log(\xi^*)\}$) differs from that in the regression of $\log(X)$ on Z (i.e., $E\{\log(\xi)\}$). However, β, which is the parameter of interest, is exactly the same in the two regressions. Finally, the form of the weight function $x^\alpha w(z)$ is essential for our arguments, but knowledge of α or of w is not required for inference.

2. Numerical Study

In this section, we compare the performance of the estimator of Chen (2009) to that of the least square estimator under an AFT model, $X = e^{-\beta Z}\xi$, and size biased sampling ($\alpha = 1$). As in Chen (2009), two models for ξ were examined: a standard log-normal distribution with expectation of $e^{0.5}$ and a Weibull distribution with scale parameter 1 and shape parameter 2 with expectation of $\sqrt{\pi}/2$. Under the log-normal distribution, the error term $\log(\xi^*) \sim N(1,1)$ with $\sigma^2 : \mathrm{Var}\{\log(\xi^*)\} = 1$. For the Weibull distribution, the length-biased density is $f(t) = 4t^2e^{-t^2}/\sqrt{\pi}$ and the corresponding density of the log is $4e^{3t-e^{2t}}/\sqrt{\pi}$ with mean 0.0182 and variance $\sigma^2 = 0.2337$ (numerical integration).

There is one covariate, Z, that has a $\mathrm{U}(0,1)$ distribution. As commented in Section 1, the density of Z^* differs from that of Z, and in the current model it is $f_{Z^*}(z) = \beta e^{\beta z}/(e^\beta - 1)$ for $0 < z < 1$. Following Chen (2009), we consider the model $\beta = 0$ in which Z and X are independent, and the model $\beta = 1$. For the former model, $Z^* \sim \mathrm{U}(0,1)$, but for the latter, Z^* has a tilted uniform distribution with weight function e^z.

It is well known that the least squares estimator is unbiased and that, for a given distribution of Z^*, its performance does not depend on β. However, since the biased sampling tilts the distribution of the covariate, the performance of the least squares estimator does depend on β. Table 1 compares the standard errors (SE) and the coverage probabilities of a 95% confidence interval (Cover) of the least squares method to that of Chen's; the results for the latter method are taken from Chen's paper (2009, Table 1, Semi-G). The standard error of the least squares estimator is $\sigma\sqrt{E\{1/(S_n^*)^2\}}$, where ${S_n^*}^2$ is the sample variance of n independent copies of Z^*, and $\sigma^2 = \mathrm{Var}\{\log(\xi^*)\}$. It is approximately $\sigma/\sqrt{n\mathrm{Var}(Z^*)}$, but it was calculated by 10^6 simulations. The confidence intervals of the log normal model are exact; the coverage probability for the Weibull model was evaluated by 10,000 simulated data sets.

As expected, the least squares estimator outperforms the estimator of Chen (2009), with smaller standard errors in all models. The performance of the least squares estimator under the Weibull model is almost as good as that of the maximum likelihood estimator (MLE) reported by Chen (2009). For example, for $\beta = 0$, the standard errors of the MLE under a Weibull model are 0.2297, 0.1096, and 0.0716, for $n = 50$, 200, and 500, respectively, quite close to the figures in Table 1.

3. Concluding Remarks

The AFT model defines a regression on a scale parameter of a lifetime variable, and it remains an AFT model after size-biased sampling. This is an important property as new methodologies are not required for size-biased data. It is of interest to characterize similar properties for regression models on a location parameter. Recalling that the log transformation changes a scale to a location parameter, and noticing that if X^* has a weighted distribution with weight x^α, then $\log(X^*)$ has a weighted distribution with weight $e^{\alpha x}$, we conclude that the same invariance property holds for a location parameter under biased sampling with weights $e^{\alpha x}$.

In summary, linear regression on the log scale is a simple and effective approach for fitting the AFT model to size-biased data. Chen's hazard-based method may be useful when censoring is present, but such an extension is yet to be developed.

Acknowledgements

This research was supported by The Israel Science Foundation.

References

Bergeron, P.-J., Asgharian, M., and Wolfson, D. B. (2008). Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association* **103,** 737–742.

Chen, Y. Q. (2009). Semiparametric regression in size-biased sampling. *Biometrics* doi:10.1111/j.1541-0420.2009.01269.x.

Ghosh, D. (2008). Proportional hazards regression for cancer studies. *Biometrics* **64,** 141–148.

van Es, B., Klaassen, C. A. J., and Oudshoorn, K. (2000). Survival analysis under cross-sectional sampling: Length bias and multiplicative censoring. *Journal of Statistical Planning and Inference* **91,** 295–312.

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

Received July 2009. *Revised July* 2009.
Accepted July 2009.

The author replied as follows:

As Mandel and Ritov (MR; 2009) referred, the proposed hazard-based estimating procedure in Chen (2009) was indeed intended with a potential extension to censored length-biased time-to-event outcomes. Since the censored length-biased time-to-event outcomes arise most likely in cross-sectional studies, however, censoring is usually informative and may create extra complication in statistical modeling and inferences, see the note by Asgharian (2003). Without taking into account such embedded informative censoring, a naive application of the hazard-based estimation procedure, or any other similar ones, may lead to incorrect inferences, although its application to uncensored size-biased outcomes is still valid due to the invariance property, as shown in Chen (2009).

It is nevertheless interesting to see that, for uncensored size-biased outcomes, the least-squares estimates that MR studied appear to outperform those of hazard-based by a noticeable margin, and in some cases, are almost as good as those of the maximum likelihood. In Mandel and Ritov's (2009) Table 1, the cited hazard-based estimates of Chen (2009) are semiparametric. They do not require the knowledge of ξ^*'s moments. The performance of MR's least-squares estimates, presumably not relying on the true value of $\sigma^2 = \mathrm{var}(\log \xi^*)$, either, in real applications. That is, the usual accelerated failure time (AFT) model, equipped with the least-squares estimation, may provide an unbiased and nearly efficient estimation of covariate effect for the uncensored size-biased outcomes. In fact, an early example of using the least squares can be found in Keiding et al. (2005) for the AFT models of backward recurrence times.

For either the hazard-based estimation or the least-squares estimation, however, obstacles remain in dealing with the potential informative censoring arising from the cross-sectional time-to-event outcomes that are subject to length bias. More innovative modeling and estimation strategies are needed to overcome these obstacles.

References

Asgharian, M. (2003). Biased sampling with right censoring: A note on Sun, Cui & Tiwari (2002). *The Canadian Journal of Statistics* **31,** 349–350.

Chen, Y. Q. (2009). Semiparametric regression in size-biased sampling. *Biometrics*, doi:10.1111/j.1541-0420.2009.01269.x.

Keiding, N., Fine, J. P., Carstensen, L., and Slama, R. (2005). Accelerated failure time regression for backward recurrence times and current durations. *National University of Singapore Institute for Mathematical Sciences Preprint Series* 2005–86. Available at: `http://www.ims.nus.edu.sg/publications-pp05.htm`.

Mandel, M. and Ritov, Y. (2009). The accelerated failure time model under biased sampling. *Biometrics*, doi:10.1111/j.1541-0420.2009.0136.x.