*VI Congreso Galego de Estatística e Investigación de Operacións*
*Vigo 5–7 de Novembro de 2003*

# LENGTH-BIAS WITH COMPETING RISKS: A GENERAL MODEL

Jacobo de Uña-Álvarez[1]

[1]Department of Statistics and OR
University of Vigo

## ABSTRACT

When analyzing times, one has to cope with several sampling problems, such as censoring and truncation. When the (left-)truncation variable is uniformly distributed, a length-bias model follows. We extend previous models and methods for length-biased, censored data, by distinguishing two types of censoring. We show that, in general, the single-censoring approach is inconsistent.

**Key words**: censoring, competing risks, length-bias, nonparametric likelihood, stationarity, truncation

## 1. INTRODUCTION

When analyzing duration times, such as those encountered in survival analysis, reliability and econometrics, sampling problems referred as truncation and censoring typically emerge. Accounting for these issues in estimation is crucial for consistency purposes. In the last years, some nonparametric models for left-truncation of length-bias type have been introduced (Asgharian *et al.* (2002), de Uña-Álvarez (2003a, b)). These models are suitable whenever the left-truncation variable is uniformly distributed on some time interval which contains the support of the time under investigation. In survival analysis and epidemiology, this modelling allows for efficient estimation of the survival curve (and related parameters) from the so-called prevalent cases, for diseases with stationary incidence (Wang (1991), Asgharian *et al.* (2002)). There exists some important backgroud coming from renewal processes theory too; in this framework, it is known that the length-bias model appears as a limit case when sampling times by cross-section (see Winter and Földes (1988) for access to literature).

The general situation of left-truncated, right-censored data sampled by cross-section can be described as follows. We observe $(T_1, X_1, \gamma_1), ..., (T_n, X_n, \gamma_n)$ independent and identically distributed random vectors; $(T_1, X_1, \gamma_1)$ follows the distribution of the conditional variable $(T, X, \gamma)$ given that $T \leq X$. Here, $X$ is the (possibly right-censored) time of interest; $\gamma$ is the censoring indicator; and $T$ is the left-truncation time, defined as time elapsed from the initial event (*e.g.* diagnosis) to the cross-section date. By "cross-section" we mean that the recruited spells are those in progress at a single point in time (the cross-section date), see Wang (1991) for further details. General nonparametric estimation methods in this setup were investigated in many papers, including Tsai *et al.* (1987), Wang (1991), Lai and Ying (1991), Gijbels and Wang (1993), Zhou (1996), and Zhou and Yip (1999). In all these papers, the observed truncation times $T_1, ..., T_n$ play a crucial role in the construction of the estimates.

In the length-biased situation, it is known that the truncation variable follows a uniform distribution, and more efficient methods (which no longer depend on the observed $T_i$) become available. Asgharian *et al.* (2002) investigated this problem under the model assumption

$$C - T \text{ and } (T, Y - T) \text{ are independent conditionally on } T \leq X, \tag{1}$$

where $C$ denotes the right-censoring variable, and $Y$ the time of ultimate interest (and hence $X = \min(Y, C)$, $\gamma = 1_{\{Y \leq C\}}$ in this case). The proposed estimators are suitable when the censoring risk is restricted to the following-up period after interception (*i.e.* cross-section), that is, under the condition $P(C \geq T) = 1$. However, in many practical cases, censoring risks which are inherent to the population under investigation will appear. This type of censoring is not induced by issues related to the following-up (such as termination of the following-up period), rather being a consequence of other types of failure which may be experienced before (or after) the cross-section time. See de Uña-Álvarez (2003a) for further illustration. In this competing risks setup, a model more general than (1) is needed.

In this work we introduce a general model in order to account for both types of censored observations in the length-biased situation. As indicated in Section 2, the $\gamma$ indicator will inform on which type of censoring has taken place (if so). The nonparametric likelihood function of the $(T_i, X_i, \gamma_i)$ is derived. This likelihood is the key for introducing general (nonparametric) estimators, and allows for (semi-)parametric inference too, whenever a (semi-)parametric model is specified. Since the truncation distribution is known, the $T_i$ will turn out to be irrelevant for the construction of the estimates. We illustrate the new model by considering several important examples. We also give some outline for asymptotic analysis in both nonparametric and (semi-)parametric setups. Previous methods for length-biased, censored data as those discussed in Asgharian *et al.* (2002) and de Uña-Álvarez (2003a, b) turn out to be particular cases of the general estimators introduced here. Importantly, it is shown that ignoring one of the two types of censoring may lead to an underestimation (resp. overestimation) of survival.

## 2. THE GENERAL MODEL: NONPARAMETRIC LIKELIHOOD

The general model is defined as follows: $X = \min(Z, D)$ and $Z = \min(Y, C)$, from which $X = \min(Y, C, D)$; the censoring indicator is

$$\gamma = \begin{cases} 1 \text{ if } X = Y \\ 0 \text{ if } X = C \\ -1 \text{ if } X = D \end{cases}$$

Here, $Y$ is the time of interest; $C$ is the censoring time which can be regarded as "independent" of the cross-section issue; and $D$ is the censoring induced by the following-up period (*e.g.* termination of following-up). Under competing risks, the $C$ variable represents a second type of failure (resp. the minimum among the remaining possible failures) which can be experienced by each subject. According to this, natural assumptions are

$$Y \text{ and } C \text{ are independent; and} \qquad T \text{ and } (Y, C) \text{ are independent,} \tag{2}$$

and

$$D - T \text{ and } (T, Z - T, \delta) \text{ are independent conditionally on } T \leq X, \tag{3}$$

where $\delta = 1_{\{Y \leq C\}}$. We set $P(D \geq T) = 1$, which implies $\{T \leq X\} = \{T \leq Z\}$.

Assumption (2) ensures that the distribution function of $Y$, say $F$, can be recovered from that of the conditional vector $(Z,\delta) \mid T \leq Z$, provided that the distribution function of $T$, say $L$, is specified. Indeed, the cumulative hazard rate of $F$ satisfies

$$\Lambda_F(y) \equiv \int_0^y \frac{dF(u)}{1-F(u-)} = \int_0^y \frac{dH^{*1}(u)}{L(u)C_L(u)}, \qquad \text{where} \qquad C_L(y) = \int_y^\infty \frac{dH^*(u)}{L(u)}, \tag{4}$$

and where $H^*(y) = P(Z \leq y \mid T \leq Z)$, $H^{*1}(y) = P(Z \leq y, \delta = 1 \mid T \leq Z)$, see de Uña-Álvarez (2003a). On the other hand, assumption (3) is needed when recovering $H^*$ and $H^{*1}$ from the $(T_i, X_i, \gamma_i)$ (actually, as mentioned above, the $T_i$ will be no longer useful to this aim).

Under (2)-(3), the full likelihood of the available data is as follows (check):

$$\begin{aligned}\mathcal{L}_n &= \prod_{i=1}^n \left\{ dH^{*1}(X_i)^{1\{\gamma_i=1\}} dH^{*0}(X_i)^{1\{\gamma_i=0\}} \left[\int_{X_i}^\infty \frac{dH^*(u)}{L(u)}\right]^{1\{\gamma_i=-1\}} \frac{dL(T_i)}{L(X_i)^{1\{\gamma_i\neq -1\}}} \right\} \times \\ &\quad \times \prod_{i=1}^n \left\{ [1-R^*((X_i-T_i)-)]^{1\{\gamma_i\neq -1\}}\, dR^*(X_i-T_i)^{1\{\gamma_i=-1\}} \right\},\end{aligned}$$

where $H^{*0}(y) = P(Z \leq y, \delta = 0 \mid T \leq Z)$ and $R^*$ denotes the conditional distribution function of $D-T$ given $T \leq X$. Now, for introducing length-bias, we assume that the truncation time is uniformly distributed on some time interval which contains the support of $Z$. Provided that $R^*$ contains no information on $F$, we come up with

$$\mathcal{L}_n \propto \prod_{i=1}^n \left\{ dH^{*1}(X_i)^{1\{\gamma_i=1\}} dH^{*0}(X_i)^{1\{\gamma_i=0\}} \left[\int_{X_i}^\infty \frac{dH^*(u)}{u}\right]^{1\{\gamma_i=-1\}} \right\}. \tag{5}$$

In particular, it is seen that $\mathcal{L}_n$ is essentially free of the truncation times. This likelihood collapses to that in Asgharian *et al.* (2002) in the case $P(\gamma = 0) = 0$. Set $\left(H_n^{*1}, H_n^{*0}\right)$ for the maximizer of (5). When $P(\gamma = -1) = 0$, this pair $\left(H_n^{*1}, H_n^{*0}\right)$ is the key for the construction of a Nelson-Aalen type estimator for $\Lambda_F$, see de Uña-Álvarez (2003a). In the general case, similar arguments are possible; use equation (4) and the uniformity on $L$ to introduce the natural empirical hazard

$$\Lambda_{F,n}(y) = \int_0^y \frac{dH_n^{*1}(u)}{uC_n(u)}, \qquad \text{where} \qquad C_n(y) = \int_y^\infty \frac{dH_n^*(u)}{u}, \tag{6}$$

and where $H_n^* = H_n^{*1} + H_n^{*0}$. The (unique) distribution function associated to $\Lambda_{F,n}$ is a natural, fully nonparametric estimator for $F$. Similar arguments lead to an estimator for the cumulative hazard of the censoring time $C$.

In order to illustrate how a single-censoring model may lead to biased estimates, consider the case in which $Y$ and $C$ are exponentially distributed with parameters $\lambda_1$ and $\lambda_2$ respectively. Then, the maximizer of the likelihood (5) as a function of $\lambda_1$ and $\lambda_2$ is (check)

$$\widehat{\lambda}_1 = \frac{n_1(n_1+n_0+n)}{n(n_1+n_0)\overline{X}_n}, \qquad \widehat{\lambda}_2 = \frac{n_0(n_1+n_0+n)}{n(n_1+n_0)\overline{X}_n},$$

where $n_1$ and $n_0$ indicate the number of cases with $X_i = Y_i$ and $X_i = C_i$, respectively, and $\overline{X}_n$ stands for the sample mean of the $X_i$. The quantity $\widehat{\lambda}_1^{-1}$ consistently estimates the expectation of $Y$ under the exponential assumption. If censoring is simply interpreted as

induced by the following-up (Asgharian *et al.* (2002)), new values $n_1' = n_1$, $n_0' = 0$, $n' = n$ arise, for which $\widehat{\lambda}_1' > \widehat{\lambda}_1$ (underestimation of the mean survival time). Similarly, if both censoring variables are identified as independent of the cross-section issue (de Uña-Álvarez (2003a)), the new values $n_1^* = n_1$, $n_0^* = n - n_1$, $n^* = n$ lead to $\widehat{\lambda}_1^* < \widehat{\lambda}_1$ (overestimation of survival). To get a rough idea of what is going on, consider a situation with 33% of uncensored individuals, and 33% of censored spells with $\gamma = 0$. Then, the (consistent) empirical average duration is 1.6 times the duration $\widehat{\lambda}_1'^{-1}$, and 0.8 times the duration $\widehat{\lambda}_1^{*-1}$. Then, correctly accounting for both types of censored individuals may be crucial.

## 3. SEMI-PARAMETRIC ESTIMATION

Introduce now $p(y) = P(\delta = 1 \mid Z = y)$. This function play a central role in the context of informative censoring models, see Dikta (1998). Under (2)-(3) it is seen that

$$p(y) = P(\delta = 1 \mid Z = y, T \leq X) = P(\gamma = 1 \mid \gamma \neq -1, X = y, T \leq X).$$

Hence, this function can be estimated nonparametrically from the available information. Interestingly, the likelihood (5) can be written as

$$\begin{aligned} \mathcal{L}_n &\propto \prod_{i=1}^{n} \left\{ dH^*(X_i)^{1\{\gamma_i \neq -1\}} \left[ \int_{X_i}^{\infty} \frac{dH^*(u)}{u} \right]^{1\{\gamma_i = -1\}} \right\} \times \\ &\quad \times \prod_{i=1}^{n} \left\{ p(X_i)^{1\{\gamma_i = 1\}} \left[1 - p(X_i)\right]^{1\{\gamma_i = 0\}} \right\} \\ &\equiv \mathcal{L}_{n1} \times \mathcal{L}_{n2}. \end{aligned}$$

The factor $\mathcal{L}_{n1}$ formally equals the likelihood in Asgharian *et al.* (2002); the difference is in that $\mathcal{L}_{n1}$ depends, rather on the truncated distribution function of the $Y$, on that of the $Z$. Both $\mathcal{L}_{n1}$ and $\mathcal{L}_{n2}$ can be maximize independently, as functions of $H^*$ and $p$, respectively. Of course, maximization of $\mathcal{L}_{n1}$ leads to the empirical $H_n^*$ introduced in Section 2. Note that each pair $(H^*, p)$ determines a unique conditional distribution for $(Z, \delta) \mid T \leq Z$.

An important semi-parametric submodel is obtained when assuming a special parametric form on $p$. There is some motivation for considering a constant value for $p$, say $p(.) \equiv \theta$ (see Gather and Pawlitschko (1998)). Under continuity, this is equivalent to assuming that $Z$ and $\delta$ are independent random variables. In such a case, it is easily seen that the maximizer of $\mathcal{L}_{n2}$ becomes

$$\theta_n = \frac{n_1}{n_1 + n_0}$$

where $n_1$ and $n_0$ are as in Section 2. Furthermore, the NPMLE of $F$ equals

$$F_n(y) = 1 - (1 - H_n(y))^{\theta_n}, \tag{7}$$

where

$$H_n(y) = \frac{\int_0^y u^{-1} H_n^*(du)}{\int_0^{\infty} u^{-1} H_n^*(du)}.$$

This $F_n$ is an ACL-type estimator, see Dikta (1998), although length-biasing and censoring by $D$ heavily complicates the nature of this empirical. The asymptotic analysis of $H_n$ can be

performed by following the arguments in Asgharian *et al.* (2002). Given the simplicity of $F_n$ as a function of $H_n$ and $\theta_n$, obtaining the results correponding to (7) is then straightforward.

If $p$ is assumed to belong to a parametric family of regression curves, say $\{p(.;\theta)\}$, where $\theta \in \Theta \subseteq \mathbb{R}^q$, semi-parametric estimation is possible too. As above, the $\theta$ parameter is estimated through the maximization of the resulting $\mathcal{L}_{n2}$, that is

$$\prod_{i=1}^{n} \left\{ p(X_i;\theta)^{1_{\{\gamma_i=1\}}} \left[1-p(X_i;\theta)\right]^{1_{\{\gamma_i=0\}}} \right\}.$$

Put $\widehat{\theta}$ for the maximizer of this product. Now, consider the equation

$$\Lambda_F(y) = \int_0^y \frac{p(u)dH^*(u)}{uC(u)}, \qquad \text{where} \qquad C(y) = \int_y^\infty \frac{dH^*(u)}{u},$$

which immediately follows from (4) under length-bias. Define the natural semi-parametric estimator for the cumulative hazard

$$\Lambda_{F,n}^{SP}(y) = \int_0^y \frac{p(u;\widehat{\theta})dH_n^*(u)}{uC_n(u)}, \tag{8}$$

where $C_n$ is defined in (6). Again, the distribution function associated to $\Lambda_{F,n}^{SP}$ is an estimator for $F$. This is an adaptation to length-biasing and further censoring of the methods proposed by Dikta (1998). Semi-parametric methods give estimators more efficient than purely nonparametric empiricals, provided that the parametric family of regression curves is correctly specified.

We mention that an estimator of type (8) with a nonparametric smoother of $p$ in the place of $p(.;\widehat{\theta})$ is possible too. Some benefits of this method when compared to purely nonparametric ones are expected, similarly as in previous works on presmoothing for Kaplan-Meier estimation, see Jácome and Cao (2002) for details.

## 4. FULLY PARAMETRIC ANALYSIS

Consider now that $F$ is assumed to belong to a parametric family of distribution functions $\{F_\theta(.)\}$, where $\theta \in \Theta \subseteq \mathbb{R}^s$, and that the distribution function $G$ of $C$ belongs to the family $\{G_\nu(.)\}$, where $\nu \in \Lambda \subseteq \mathbb{R}^t$. Then, the goal is the estimation of the vector $\theta$ (resp. $\nu$) from the available data. Application of the likelihood (5) gives

$$\begin{aligned}\mathcal{L}_n \quad \propto \quad & \frac{1}{\mu(\theta,\nu)^n} \prod_{i=1}^{n} \Big\{ \left[f_\theta(X_i)(1-G_\nu(X_i))\right]^{1_{\{\gamma_i=1\}}} \left[(1-F_\theta(X_i))g_\nu(X_i)\right]^{1_{\{\gamma_i=0\}}} \times \\ & \times \left[ \int_{X_i}^\infty (1-G_\nu(u))f_\theta(u)du + \int_{X_i}^\infty (1-F_\theta(u))g_\nu(u)du \right]^{1_{\{\gamma_i=-1\}}} \Big\},\end{aligned}$$

where $f_\theta(.)$ and $g_\nu(.)$ denote the probability density functions associated to $F_\theta(.)$ and $G_\nu(.)$, respectively, and where

$$\mu(\theta,\nu) = \int_0^\infty u(1-G_\nu(u))f_\theta(u)du + \int_0^\infty u(1-F_\theta(u))g_\nu(u)du$$

stands for the expectation of $Z$. Typically, estimation of $(\theta,\nu)$ will be carried out by solving the system of $s+t$ equations

$$\frac{\partial \ln \mathcal{L}_n}{\partial \theta} = 0, \qquad \frac{\partial \ln \mathcal{L}_n}{\partial \nu} = 0.$$

Asymptotic analysis of the resulting estimators can be performed in the usual way, under a number of technical conditions on the involved functions and parametric spaces. An example of this fully parametric approach has been considered in Section 2 (exponentially distributed times).

## 5. CONCLUSIONS

A new general model for left-truncated, right-censored data has been proposed. The model is suitable for data sampled by cross-section. The nonparametric likelihood function has been obtained, and efficient estimation of survival has been introduced under a length-bias assumption. The model allows for several types of censoring, such as that induced by following-up or that coming from the presence of competing risks. Distinguishing among this censoring types is crucial in order to avoid inconsistencies. Nonparametric, semi-parametric, and fully parametric analysis have been considered. The introduced model and estimation methods extend previous proposals under censoring and length-bias.

## 6. ACKNOWLEDGEMENTS

Work supported by the Grants PGIDIT02PXIA30003PR and BFM2002-03213. This work is dedicated to my beloved children Marcos and Paula.

## 7. 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* 97, 201-209.

Dikta, G. (1998). "On semiparametric random censorship models". *Journal of Statistical Planning and Inference* 66, 253-279.

Gather, U. and Pawlitschko, J. (1998). "Estimating the survival function under a generalized Koziol-Green model with partially informative censoring". *Metrika* 48, 189-207.

Gijbels, I. and Wang, J. L. (1993). "Strong representation of the survival function estimator for truncated and censored data with applications". *Journal of Multivariate Analysis* 47, 210-229.

Jácome, M. A. and Cao, R. (2002). "Presmoothed kernel density estimator for censored data". Unpublished manuscript.

Lai, T. L. and Ying, Z. (1991). "Estimating a distribution function with truncated and censored data". *The Annals of Statistics* 19, 417-442.

Tsai, W. Y., Jewell, N. P. and Wang, M. C. (1987). "A note on the product-limit estimator under right censoring and left truncation". *Biometrika* 74, 883-886.

de Uña-Álvarez, J. (2003a). "Nelson-Aalen and product-limit estimation in selection bias models for censored populations". Unpublished manuscript.

de Uña-Álvarez, J. (2003b). "Empirical estimation under length-bias and Type I censoring". *Reports in Statistics and OR*, 03-04, University of Santiago de Compostela.

Wang, M.-C. (1991). "Nonparametric estimation from cross-sectional survival data". *Journal of the American Statistical Association* 86, 130-143.

Winter, B. B. and Földes, A. (1988). "A product-limit estimator for use with length-biased data". *Canadian Journal of Statistics* 16, 337-355.

Zhou, Y. (1996). "A note on the TJW product-limit estimator for truncated and censored data". *Statistics & Probability Letters* 26, 381-387.

Zhou, Y. and Yip, P. S. F. (1999). "A strong representation of the product-limit estimator for left truncated and right censored data". *Journal of Multivariate Analysis* 69, 261-280.