

CENSORING AND WEIGHTING IN SURVIVAL ESTIMATION FROM SURVEY DATA

J.F.Lawless¹

ABSTRACT

The estimation of survival distributions from complex survey data requires that we account for the sampling design, and this is typically done by including design weights in estimators or estimating functions. This article examines weighting for nonparametric (Kaplan-Meier) estimation of a survivor function, when survival times are subject to censoring. It is shown that, in general, weights related to both censoring and the sampling plan are needed for consistent estimation. Some areas for further study are identified.

KEY WORDS: Lifetime data; Longitudinal surveys; Survivor functions; Weighted Kaplan-Meier estimates.

RÉSUMÉ

Le plan d'échantillonnage doit être pris en considération lorsque l'on estime des distributions de survie à partir de données d'enquête complexe. Habituellement, cela est fait en introduisant les poids d'échantillonnage dans les estimateurs ou les fonctions d'estimation. Cet article examine cette pondération pour l'estimation non-paramétrique (Kaplan-Meier) d'une fonction de survie à partir de données censurées. Il est montré qu'en général, il est nécessaire d'utiliser des poids tenant compte à la fois de la censure et du plan d'échantillonnage afin d'obtenir une estimation convergente. De futurs domaines de recherche sont également identifiés.

MOTS CLÉS: Données de survie; enquêtes longitudinales; estimateurs pondérés de Kaplan-Meier; fonctions de survie.

1. INTRODUCTION

In longitudinal surveys where individuals are followed over time, data on survival or lifetime variables T are often collected; examples are the duration of a jobless spell or the age at which a child reaches some developmental milestone. Design weights, possibly adjusted for nonresponse, are typically used for estimation of finite population or superpopulation parameters from survey data. However, as we discuss here, the estimation of survival distributions and related quantities is often more complicated, and may require time-varying weights to adjust for nonignorable censoring. This article discusses followup and censoring processes in longitudinal surveys, and the type of weighting needed to provide consistent estimators.

We consider a superpopulation framework in which T represents the survival time for an individual and x is an associated vector of covariates. Attention is directed at the marginal distributions of T_i or, in the case of regression analysis, at the distributions of T_i given x_i for individuals $i = 1, \dots, N$ who are considered a random sample from the superpopulation. The data that are obtained consist of survival (or censoring) times and covariates for a sample s of individuals drawn from the survey population $\{1, \dots, N\}$, typically using a complex sampling design involving stratification and clustering. The T_i 's for different individuals in the finite population and in the sample s are not in general mutually independent either unconditionally or conditional on the covariate values x_i .

¹ J.F. Lawless (jlawless@uwaterloo.ca), Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

As is often done, we focus on estimation of the marginal distributions of T_i or T_i given x_i . Association among survival times is not modelled explicitly, but is accounted for in the estimation procedures. This article deals mainly with the estimation of a common marginal survivor function,

$$S(t) = P(T_i \geq t) \quad (1)$$

with regression models mentioned only in the last section. As indicated, we focus on the issue of weights for estimation, and do not consider other matters such as variance estimation in any detail.

Section 2 describes the framework for estimation of a survivor function (1) from complex survey data, and examines weighted Kaplan-Meier estimates. Section 3 provides a brief illustration in the context of a longitudinal labour survey. Section 4 discusses extensions and identifies some areas for further study.

2. WEIGHTED KAPLAN-MEIER ESTIMATION

Let $R_i = I(i \in \mathcal{E})$ indicate whether individual $i (i = 1, \dots, N)$ is in the sample, and assume that for some vector of observable covariates x_i the sample selection probabilities π_i satisfy

$$\pi_i = P(R_i = 1 | T_i, x_i) = P(R_i = 1 | x_i). \quad (2)$$

In what follows we use the notation $Y \perp X$ to denote that random variables Y and X are independent, and the notation $Y \perp X | Z$ to denote that Y and X are conditionally independent, given the value of the random variable Z . Thus, (2) indicates that $T_i \perp R_i | X_i$, but note that it is not in general true that $T_i \perp R_i$. For example it is often the case that variables in x_i are also related to survival time, so unconditional independence of T_i and R_i does not hold.

Survival times are usually subject to right censoring in longitudinal studies. That is, an individual has a potential censoring time $C_i > 0$ such that if $T_i \leq C_i$ the exact value of T_i is observed, but if $T_i > C_i$, then this is all that is known (e.g. Lawless 2003a, Ch.2). In longitudinal surveys, C_i is related to the length of the survey followup period, the time at which an event (e.g. the start of a jobless spell for an individual) triggering the survival time occurred, and whether the individual is lost to followup before the end of the survey period. An example is discussed in Section 3.

The Kaplan-Meier (KM) estimate of a survivor function $S(t)$ from “standard” survival data is defined as follows (e.g. Lawless 2003a, Ch.3). Let $t_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, $i = 1, \dots, n$ represent the observed data from a random sample of individuals with survivor function $S(t)$. For simplicity we take time to be discrete with values $1, 2, \dots$ and define the hazard function of T as

$$\lambda(t) = P(T = t | T \geq t) = \frac{f(t)}{S(t)} \quad t = 1, 2, \dots$$

Also define $d_i(t) = I(T_i = t, \delta_i = 1)$ and $y_i(t) = I(t_i \geq t)$. Then the hazard value $\lambda(t)$ is estimated by

$$\hat{\lambda}(t) = \frac{\sum_{i=1}^n d_i(t)}{\sum_{i=1}^n y_i(t)}, \quad (3)$$

and from the relationship $S(t) = \prod_{u < t} (1 - \lambda(u))$, the KM estimate is obtained as

$$\hat{S}(t) = \prod_{u < t} (1 - \hat{\lambda}(u)) \quad (4)$$

For the estimator (4) to be consistent for $S(t)$, it is effectively necessary that $T_i \perp C_i$, although this can be weakened a little to allow adaptive censoring (see Lawless 2003a, Ch.2).

When survival times are defined for a sample of individuals drawn from a complex survey design satisfying (2), the condition $T_i \perp R_i$, also needed for (4) to be consistent, no longer holds. Moreover, the censoring times C_i are often not

independent of the T_i 's, because of their joint association with design variables or other covariates. Consequently with (2) in mind, we assume that for some set of covariates X_i we have, for $i = 1, \dots, N$,

$$(i) T_i \perp R_i \mid X_i \quad (ii) T_i \perp C_i \mid X_i, R_i = 1 \quad (5)$$

These conditions are often reasonable, given appropriate choice of x -variables and imply that the sample design and the observation process (from which censoring arises) are ignorable, conditional on the X_i 's.

We are now in a position to discuss estimation of $S(t)$ from complex longitudinal survey data. We do this by considering weighted estimating functions

$$U(t) = \sum_{i=1}^N R_i w_i(t) I(t_i \geq t) [I(T_i = t) - \lambda(t)] \quad (6)$$

for $t = 1, 2, \dots$. The equations $U(t) = 0$ give estimates

$$\hat{\lambda}(t) = \sum_{i \in \mathcal{S}} w_i(t) d_i(t) / \sum_{i \in \mathcal{S}} w_i(t) y_i(t) \quad (7)$$

and then an estimate $\hat{S}(t)$ from (4). Two special cases should be noted, (i) $w_i(t) = 1$, which case (7) and $\hat{S}(t)$ are the standard KM estimates, and (ii) $w_i(t) = \pi_i^{-1}$, in which case (7) and $\hat{S}(t)$ are design-weighted KM estimates proposed for use with surveys (e.g. Folsom et al. 1989, Kalton et al. 1992).

Consider the general case of (6), where the $w_i(t)$'s are known but may depend on the X_i 's. The estimators $\hat{\lambda}(t)$ and $\hat{S}(t)$ are quite generally consistent if the estimating functions $U(t)$ in (6) have mean 0 (are "unbiased"), but may fail to be consistent otherwise. To examine this, let $\lambda^*(t)$ represent the values of $\lambda(t)$ such that $E(U(t)) = 0$; this is what $\hat{\lambda}^*(t)$ of (7) converges to in probability as sample size increases (e.g. White 1982). From (5) and (6) we get that

$$\lambda^*(t) = \frac{\sum_{i=1}^N E_{X_i} \{w_i(t) G_i(t \mid X_i) \pi_i(X_i) P(T_i = t \mid X_i)\}}{\sum_{i=1}^N E_{X_i} \{w_i(t) G_i(t \mid X_i) \pi_i(X_i) P(T_i \geq t \mid X_i)\}} \quad (8)$$

where

$$G_i(t \mid X_i) = P(C_i \geq t \mid X_i, R_i = 1), \quad \pi(X_i) = P(R_i = 1 \mid X_i) \quad (9)$$

Note that the true value of $\lambda(t)$ is

$$\lambda_0(t) = \frac{f(t)}{S(t)} = \frac{E_{X_i} \{P(T_i = t \mid X_i)\}}{E_{X_i} \{P(T_i \geq t \mid X_i)\}} \quad (10)$$

We see that if $w_i(t) = 1$ (the standard KM) then $\lambda^*(t) = \lambda_0(t)$ if either $T_i \perp X_i$ or if $(R_i, C_i) \perp X_i$, but not in general. That is, $\hat{\lambda}(t)$ is consistent for $\lambda(t)$ if $T_i \perp (R_i, C_i)$, or in other words, if both the sampling design and censoring process are independent of T_i , and hence ignorable. This is usually not the case.

If $w_i(t) = \pi(X_i)^{-1}$ (the design-weighted KM) then $\lambda^*(t) = \lambda_0(t)$ if either $T_i \perp X_i$ or $C_i \perp X_i \mid R_i = 1$, but not more generally. That is, design-weighting provides consistent estimation if T_i and C_i are independent. This is sometimes the case, but often censoring and survival times are related to some common covariates; in this case design-weighting does not make $\hat{\lambda}(t)$ consistent.

If $T_i \perp X_i$ then it is clear that $T_i \perp (R_i, C_i)$ and $\lambda^*(t) = \lambda_0(t)$, but this is rather rare. However, we see from (8) that if

$$w_i(t)^{-1} = \pi_i(X_i)G_i(t | X_i) \quad (11)$$

then $\lambda^*(t) = \lambda_0(t)$. Therefore, we can achieve consistent estimation of $\lambda(t)$ and $S(t)$ if we can model the dependence of censoring on the covariates X_i ; the design probabilities $\pi_i(X_i)$ are assumed known. The use of such “inverse probability of censoring” (IPC) weights has been discussed in the survival analysis literature (e.g. Robins 1993, Satten et al. 2001, Scharfstein and Robins 2002) and is often used to deal with nonignorable censoring or losses to followup in longitudinal studies.

In practice it is necessary to estimate $G_i(t | X_i)$ in (9). Assuming that this is done consistently, then $\hat{\lambda}(t)$ and $\hat{S}(t)$ will remain consistent estimators (e.g. Satten et al. 2001). However, the development of variance estimates for $\hat{S}(t)$ is a major challenge in the longitudinal survey context, where lifetimes cannot be assumed independent across all individuals. As a result, methods developed in the standard survival analysis setting (e.g. Satten et al. 2001) cannot be employed; see Lawless (2004) and Section 4 below.

3. A BRIEF ILLUSTRATION

Labour force surveys such as Statistics Canada’s Survey of Labour and Income Dynamics (SLID) provide information on the durations of jobless spells for individuals who lose their jobs. In SLID, individuals in the first panel were selected in January 1993 and followed (barring loss to followup) from then until December 1998 (see e.g. Statistics Canada 1997). Subsequent panels are selected every three years and similarly followed for six years. To illustrate the issues discussed in Section 2, suppose that we wish to examine the marginal distribution $S(t)$ of jobless spell durations T_i for individuals losing their jobs in, say, the calendar year 1997. One way to do this would be to fit regression models for T_i and then to average these over the population (e.g. Korn and Graubard 1999, Lawless 2004). A second option is to use the methods of Section 2.

In SLID, data on the previous calendar year are collected each January. Thus, the start date for a jobless spell starting in 1997 would be obtained in January 1998. Information about the date at which the spell ended, and thus the value of T_i , could be obtained in January 1998 or January 1999, if this date were in the calendar years 1997 or 1998, respectively.

However, the value of T_i could also be right-censored through either of the following situations:

- (a) the person is still jobless on December 31, 1998 or
- (b) the person is still jobless on December 31, 1997 (as ascertained in January 1998) but is lost to followup as of January 1999.

In either case there is a censoring time C_i for which it is known that $T_i \geq C_i$. Note that the value of C_i depends on the start date of the jobless spell.

The censoring probabilities $G_i(t | X_i)$ in (9) can be modelled and estimated by relating the probability a person is lost to followup at a given (January) interview date to covariates X_i , chosen so as to make the assumptions (5) plausible. Note that it is assumed the probability an individual becomes lost to followup at a given year does not depend on the survival time T_i , once X_i is given. It is also assumed that individuals who started a jobless spell in 1997 but were lost to followup in January 1998 (and so not observed to have been jobless) can be thought of as having a censoring time $C_i = 0$, and so not contributing to the estimation of $S(t)$. That is, observation of the start of a jobless spell is independent of the duration T_i of the spell, given X_i and that $R_i = 1$.

As indicated in Section 2, when the $w_i(t)$ ’s given by (11) are estimated in this way, the question of how to estimate the variance of $\hat{S}(t)$ remains open. Section 4 comments briefly on variance estimation.

4. CONCLUDING REMARKS

The use of weights that vary across a set of discrete followup times in a longitudinal survey has been discussed in the literature (e.g. Miller et al. 2001). The situation for survival or duration data is more complex because survival times can

originate and end at arbitrary times between followup interviews or data collection points. This creates a need for more complicated time-varying weights in certain settings.

For the general setting of Section 2, estimates $\hat{S}(t)$ are readily obtained. However, when IPC weights are estimated through a model for $G_i(t | X_i)$ of (9), the problem of variance estimation for $\hat{S}(t)$ becomes difficult. The presence of association among the survival times for certain individuals rules out the martingale-based estimation procedures used for independent survival times (e.g. Satten et al. 2001). The provision of variance estimation methodology is thus an area where research is needed.

Empirical studies into the effect of ignoring IPC weighting would also be valuable, since if only design weights are used in (7), variance estimates are readily obtained (e.g. Lawless 2003b, 2004). It appears that in some settings this incurs very little bias.

Similar issues concerning weights arise with regression models. For example, if we are interested in the distribution of T given covariates X_{it} and if (5) does not hold with $X_i = X_{it}$, then weighted estimation may be used. Extensions of the weighting issue discussed here also apply for the analysis of event history processes involving repeated events or transitions among states (e.g. Lawless 2003b). We note as well that the discrete time discussion in this article can be carried over to continuous time models, and that time-varying covariates can also be dealt with (e.g. Satten et al. 2001), though these topics have received virtually no attention in the case of survey data.

Finally, although the point of view here is analytic and superpopulation –based, the estimation of descriptive finite population characteristics associated with survival times is similarly affected. In particular, if losses to followup are related to design variables then time-varying weights are needed to provide design-consistent estimates of characteristics, such as $\frac{1}{N} \sum_{i=1}^N I(T_i \geq t) = S_N(t)$.

ACKNOWLEDGEMENT

This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- Folsom, R., LaVange, L. and Williams, R.L. (1989). A probability sampling perspective on panel data analysis. In *Panel Surveys*, editors D. Kasprzyk, G.J. Duncan, G. Kalton and M.P. Singh, pp. 108-138. New York: John Wiley and Sons.
- Kalton, G., Miller, D.P. and Lepkowski, J.(1992). Analyzing spells of program participation in the SIPP. Technical report, University of Michigan Survey Research Center, Ann Arbor MI.
- Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley and Sons.
- Lawless, J.F. (2003a). *Statistical Models and Methods for Lifetime Data*, 2nd edition. Hoboken NJ: John Wiley and Sons.
- Lawless, J.F. (2003b). Event history analysis and longitudinal surveys. Chapter 15 in *Analysis of Survey Data*, Editors R.L. Chambers and C.J. Skinner. Chichester: John Wiley and Sons.
- Lawless, J.F. (2004). Censoring, Weighting and Survival Estimation from Survey Data. Manuscript.
- Miller, M.E., Ten Have, T.R., Reboussin, B.A., Lohman, K.K. and Rajeski, W.J.(2001). A marginal model for analyzing discrete outcomes from longitudinal surveys with outcome subject to multiple-cause nonresponse. *Journal of the American Statistical Association*, **96**, 844-857

- Robins, J. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the American Statistical Association Biopharmaceutical Section*, 24-33.
- Satten, G.A., Datta, S. and Robins, J. (2001). Estimating the marginal survivor function in the presence of time dependent covariates. *Statistics and Probability Letters*, **53**, 397-403.
- Scharfstein, D.O. and Robins, J.M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, **89**, 617-634.
- Statistics Canada (1997). *Survey of Labour and Income Dynamics Users Guide*. Statistics Canada, Ottawa, Ontario, Catalogue 75M0001GPE.
- White, H. (19982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-25.