University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2002 *Paper* 113

Current Status Data: Review, Recent Developments and Open Problems

Nicholas P. Jewell[*] Mark J. van der Laan[†]

[*] Division of Biostatistics, School of Public Health, University of California, Berkeley, jewell@uclink.berkeley.edu

[†] Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/ucbbiostat/paper113

Copyright ©2002 by the authors.

Current Status Data: Review, Recent Developments and Open Problems

Nicholas P. Jewell and Mark J. van der Laan

Abstract

Researchers working with survival data are by now adept at handling issues associated with incomplete data, particular those associated with various forms of censoring. An extreme form of interval censoring, known as current status observation, refers to situations where the only available information on a survival random variable T is whether or not T exceeds a random independent monitoring time C. This article contains a brief review of the extensive literature on the analysis of current status data, discussing the implications of response-based sampling on these methods. The majority of the paper introduces some recent extensions of these ideas to more complex forms of survival data including, competing risks, multivariate survival data, and general counting processes. Our comments are largely focused on nonparametric techniques where the form of the distribution function, or survival curve, associated with T, is left unspecified. Modern theory of efficient estimation in semiparametric models has allowed substantial progress on many questions regarding estimation based on current status data in these extended formats; we also highlight remaining open questions of interest.

1 Introduction

In some survival analysis applications, observation of the lifetime random variable T is restricted to knowledge of whether or not T exceeds a random monitoring time C. This structure is widely known as current status data, and sometimes referred to as interval censoring, case I (Groeneboom & Wellner, 1992). Section 2 briefly notes several generic examples where current status data is encountered frequently.

Let T have a distribution function F, with associated survival distribution $S = 1 - F$. We assume that interest focuses on estimation and inference on F, but recognize throughout that, in most applications, the goal will be estimation of a variety of functionals of F. In many cases, the regression relationship between T and a set of covariates $\mathbf{Z}$ will be of primary concern. In some situations, parametric forms of F may be useful, although we pay most attention to the nonparametric problem where the form of F is unspecified. In the regression model, semiparametric models for the conditional distribution of T, given $\mathbf{Z}$, are appealing and heavily used.

The monitoring time C is often taken to be random, following a distribution function G, almost always assumed independent of T. However, most techniques are based on the conditional distribution of T, given C, and so work equally well for fixed non-random C. In the random case, we assume, for the most part, that the data arise from a simple random sample from the joint distribution of T and C; in the non-random case, we assume that simple random samples, often of size 1, are selected for each fixed choice of C. When C is random, the data can thus be represented by n observations from the joint distribution of (T, C); however, only $\{(Y_i, C_i : i = 1, \ldots, n\}$ is observed where $Y = I(T \leq C)$. In Section

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

6, we make some brief remarks about the intriguing possibility of dependence between C and T, particularly when such dependence is introduced by design.

In Section 4.1, we discuss an important variant to simple random sampling, namely the analysis of case-control samples. Here, two separate random samples are obtained, the first an i.i.d. random sample of size n_0 from those for whom $T > C$ (controls), the second an i.i.d. random sample of size n_1 from individuals for whom $T \leq C$ (cases). Section 4.2 covers the situation where observation of the origin of T is also subject to censoring, thereby yielding doubly censored current status data.

Section 5 extends the notion of current status observation to more complex forms of survival data. These include competing risks, multivariate survival variables $\mathbf{T} = (\mathbf{T_1}, \ldots, \mathbf{T_p})$, and special cases of the latter, for example, when $T_p \geq T_{p-1} \geq \cdots \geq T_1$. This leads naturally to consideration of the scenario, in Section 5.4, where observation at time C is on a general counting process, rather than the case of a single jump from count '0' to '1' as occurs with a simple survival random variable.

2 Motivating Examples

Before discussing estimation techniques designed for current status data, it will be helpful to have some motivating examples at the back of our minds as we proceed. Early examples arose in demographic applications, with a common version occurring in studies of the distribution of the age at weaning in various settings (Diamond, McDonald & Shah, 1986; Diamond & McDonald, 1991; Grummer-Strawn, 1993). Here, T represents the age of a child at weaning and C the age at observation. Inaccuracy and bias surrounding exact

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

measurement of T, even when $T < C$, led to use of solely current status data on T at C for the purpose of understanding F.

Another kind of example arises naturally in the study of infectious diseases, particularly when infection is an unobserved event, that is, one with often no or few clinical indications. The prototypical example is infection with the Human Immunodeficiency Virus (HIV), in particular, partner studies of HIV infection (Jewell and Shiboski, 1990; Shiboski, 1998a). The most straightforward partner study occurs when HIV infection data is collected on both partners in a long-term sexual relationship. These partnerships are assumed to include a primary infected individual (index case) who has been infected via some external source, and a susceptible partner who has no other means of infection other than contact with the index case. Suppose T denotes the time (or number of infectious contacts) from infection of the index case to infection of the susceptible partner, and that the partnership is evaluated at a single time C after infection of the index case; then, the infection status of the susceptible partner provides current status data on T at time C. Since partnerships are often recruited retrospectively so that the event of the susceptible partner's infection has occurred (or not) at the time of recruitment, some form of case-control design may be used; in this case the methods of Section 4.1 are appropriate.

Our next area of application is in carcinogenecity testing when a tumor under investigation is occult (see Gart et al, 1986). In this example, for each experimental animal, T is the time from exposure to a potential carcinogen until occurrence of the tumor, and C is the time, on the same scale, of sacrifice. Upon sacrifice, the presence or absence of the occult tumor can be determined providing current status information on T.

Finally, a common source of current status data is estimation of the distribution of age

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

at incidence of a non-fatal human disease for which the exact incidence time is usually unknown although accurate diagnostic tests for prevalent disease are available. If a cross-sectional sample of a given population receives such a diagnostic test, then the presence or absence of disease in an individual of age C yields current status information on the age, T, at disease incidence. Keiding (1991) describes the nonparametric maximum likelihood estimator of the distribution of the age at incidence of Hepatitis A infection, based on cross-sectional data obtained by K. Dietz. A case study of the application of current status techniques to estimation of age-specific immunization rates is given in Keiding et al (1996). For rare diseases, this approach to age incidence is only viable if a case-control sampling scheme is used. For example, with Alzheimer's disease, it is feasible to obtain a random sample of prevalent Alzheimer's patients, measuring their age at sampling, and then subsequently sample population controls. However the data are obtained, modification to current status methods are required if presence of the disease substantially modifies the risk of death, thereby reducing the probability of being sampled. This is an issue that deserves further study.

Note that, in econometrics, there is a parallel terminology and literature that has developed on similar topics to those discussed below.

3 Simple Current Status Data

Recall that the binary random variable Y is defined to be 1 if $T \leq C$ and 0 if $T > C$. Thus, $E(Y|C = c) = P(T \leq C|C = c) = F(c)$, and so estimation of F can be viewed in terms of estimation of the conditional expectation of Y for all c, with a monotonicity constraint imposed on the regression function.

COBRA
Collection of Biostatistics
Research Archive

Now, suppose an i.i.d. random sample of the population is obtained with observed data thereby given by $\{(y_i, c_i) : i = 1, \ldots, n\}$. The likelihood of this data is thus given by

$$L = \prod_{i=0}^{n} F(c_i)^{y_i}(1 - F(c_i))^{1-y_i} dG(c_i). \tag{1}$$

Assuming that the monitoring time C is independent of survival, estimation of F can then be based on the conditional likelihood of Y, given C, namely,

$$CL = \prod_{i=0}^{n} F(c_i)^{y_i}(1 - F(c_i))^{1-y_i}. \tag{2}$$

This conditional likelihood is immediately applicable also in the case of fixed non-random selection of the monitoring times, assuming that such selection is again independent of T.

If F belongs to a finite-dimensional parametric family, $\{F = F_\theta : \theta \in \Theta\}$, then estimation and inference regarding θ and thus F_θ, can be obtained by standard maximum likelihood techniques based on (2). On the other hand, nonparametric maximum likelihood estimation of F requires maximization of (2) over the space of all distribution functions. This nonparametric maximization problem has been much studied—Ayer et al. (1955) provided a fast and effective approach, the ubiquitous pool-adjacent-violators algorithm, to compute the nonparametric maximum likelihood estimator, $\hat{F}$. The connection to convex minorants is extensively discussed in Barlow et al. (1972) and Groeneboom & Wellner (1992). The estimator $\hat{F}$ converges to F as n tends to infinity, but at rate $n^{-1/3}$, unlike the empirical ditribution function, or the Kaplan-Meier estimator, both of which converge at the more familiar $n^{-1/2}$ rate. The limiting distribution of $\hat{F}$ is not Gaussian, but a more complex distribution associated with two-sided Brownian motion (Groenboom & Wellner, 1992). The estimator $\hat{F}$ is a step function, jumping only at a subset of the observed monitoring times $c_1, \ldots, c_n$. In fact, the data only identifies the value of F at $c_1, \ldots, c_n$ and at

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

no other value of t. Identification of the entire distriburion function F as n tends to infinity depends therefore on the support of F being contained within the support of G. Finally, a smoothing technique can be incorporated into the pool-adjacent-violators algorithm to produce smoother estimates of F across the c_i's—see Mammen (1991) and Mukerjee (1988).

Despite the unusual and slow rate of convergence of $\hat{F}$ to F, Huang & Wellner (1995) show that estimates of smooth fuctionals of F, based on $\hat{F}$, converge at rate $n^{-1/2}$ and are asymptotically efficient at many data generating distributions. These authors also supply the influence curve for such smooth functional estimators, thereby facilitating straightforward calculations for (asymptotic) confidence intervals.

3.1 Epidemiological Applications–Calculation of the Relative Risk

In some simple epidemiological studies, interest focuses on the calculation and comparison of the cumulative incidence rate for a specific disease over a pre-determined period of risk and for differing levels of exposure to some risk factor. In many investigations, the risk interval is common to all individuals under study, and calculation of the cumulative risk thereby corresponds to current status estimation of F at a single monitoring time corresponding to the length of the interval, C. Of course, standard 'survival' follow-up of the study participants yields exact incidence times, albeit right censored at C. If risk intervals vary in length across individuals the nonparametric maximum likelihood estimator, $\hat{F}$, discussed above provides an estimate of the cumulative risk, at any observed value of C, that is based only on whether incident disease occurs in the observed risk interval or not. Again, estimates of cumulative risk can again be computed from follow-up data using the Kaplan-Meier estimator for right censored data.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Typically, follow-up measurement of the exact time of disease incidence is considerably more expensive than mere (current status) assessment of incidence at some point during the risk interval. If F is parametrically specified, the efficiency of current status estimates of cumulative incidence, as compared to use of more complete incidence times arising from full follow-up, can be calculated directly. The simpler current status measurements are often surprisingly efficient, except in situations where the monitoring tiomes are all either very small or very large in terms of the location of the support of F. Of more relevance, similar efficiency comparisons can be made when the parameter of interest is a comparative measure of the cumulative incidence rates across exposure groups, often leading to similar conclusions regarding the effectiveness of current status observations. In a study design, the relative costs of continuous follow-up versus a single current status assessment must be fully considered, and, of course, the latter allows investigation of more complex incidence properties. Consideration of the role of more complex measurements of exposures and other factors associated with incidence lead naturally to the development of regression models and their estimation from current status data.

3.2 Regression Models

In Section 3.1, we touched on the two-group situation where the difference in survival properties across exposure groups is of fundamental concern rather than the shape of the underlying survival distributions. Clearly, many applications include more general and higher dimensional covariates in situations where the relationships between the latter and survival time are key. A substantial literature has developed for regression models of this kind for survival outcomes, potentially subject to right censoring. Much recent work has extended the application of these models to current status data.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

There is an immediate and valuable correspondence between the regression models that link T, the survival random variable, and Y, the current status version of T, to a k-dimensional covariate vector $\mathbf{Z}$. Doksum & Gasko (1990) had previously considered this association between survival and binary regression models in the context of censored survival data. This is extremely useful since estimates of parameters in the regression model for the observed Y can then be interpreted in terms of the parameters in the regression model for the unobserved T. For example, suppose that survival times follow a proportional hazards model (Cox, 1972)

$$S(t|\mathbf{Z}=\mathbf{z}) = [S_0(t)]^{e^{\beta\mathbf{z}}} \tag{3}$$

where S_0 is an arbitrary survival function for the sub-population for whom $\mathbf{Z}=\mathbf{0}$, and β is a k-dimensional vector of regression coefficients. Each component of β gives the relative hazard associated with a unit increase in the corresponding component of $\mathbf{Z}$, holding all other components fixed. Then, if we write $p(\mathbf{z}|c) = E(Y|C=c, \mathbf{Z}=\mathbf{z})$, the current status random variable Y is related to $\mathbf{Z}$ through

$$\log - \log(1 - p(\mathbf{z}|c)) = \log - \log[S_0(c)] + \beta\mathbf{z}. \tag{4}$$

This is a particular case of a generalized linear model for Y with complementary log-log link and offset given by an arbitrary increasing function of the observed 'covariate' C (that is, $\log - \log[S_0(C)]$). The regression coefficients, β, here are thus exactly the relative hazards from the regression model for T.

As another example, suppose T follows the proportional odds regression model (Bennett, 1983) defined by

$$1 - S(t|\mathbf{Z}=\mathbf{z}) = \frac{1}{1+e^{-\alpha(t)-\beta\mathbf{z}}},$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

where $S_0(t) = \frac{1}{1+e^{\alpha(t)}}$. Here, Y is associated with $\mathbf{Z}$ via the logit link:

$$\log \frac{p(\mathbf{z}|c)}{(1-p(\mathbf{z}|c))} = \alpha(c) + \beta \mathbf{z}. \tag{5}$$

Again, the 'intercept' term, $\alpha(C) = \log \frac{(1-S_0(C))}{S_0(C)}$ is an increasing function of C.

If the baseline survival function S_0 is assumed to follow a particular parametric form, the corresponding binary regression model will often simplify to a familiar generalized linear model, so that standard software can be used to estimate both S_0 and the regression parameters β. As an example, suppose that S_0 is assumed to be a Weibull distribution with hazard function $e^a b t^{b-1}$, and that the proportional hazards model (3) holds for T. Then, the binary regression model for Y, given by (4) simplifies to a straightforward generalized linear model with complementary log-log link:

$$\log - \log(1 - p(\mathbf{z}, c)) = a + b\log(c) + \beta \mathbf{z}.$$

On the other hand, if S_0 is left arbitrary, semiparametric methods can be used to tackle inference on β, treating S_0 as a nuisance parameter. Shiboski (1998b) provides an excellent review of these methods for current status data, discussing versions of a backfitting algorithm to compute estimates of β while fully acknowledging the monotonicity constraints in the intercept terms of the kind illustrated in (4) and (5). In the semiparametric regression model, dependence between C and the covariates $\mathbf{Z}$ can introduce some bias in estimation of β. Shiboski (1998b) also describes some simulations that compare the relative performance of coefficient estimates based on parametric or nonparametric assumptions on S_0. Asymptotic results regarding coefficient estimates within a semiparametric model (S_0 left unspecified), necessary for inference, are discussed in Rabinowitz, Tsiatis & Aragon (1995), Huang (1996) and Rossini & Tsiatis (1996) for the accelerated failure time, proportional

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

hazards and proportional odds regession models, respectively, for T. Andrews, van der Laan & Robins (2002) give locally efficient estimates for regression coefficient estimates in a broad class of models that (i) includes the accelerated failure time model, and (ii) allows for time-dependent covariates.

4 Different Sampling Schemes

In Section 3, and in the construction of (1) and (2) in particular, we have assumed that an i.i.d. random sample of observations of (Y, C) are available, noting that, with the assumption of independence between T and C, the use of (2) allows the methods to apply directly to designs where the monitoring times are pre-determined. Often, the failures of interest are rare in the population so that such random samples provide very few observations where failure has occurred at the observed monitoring time, whether the latter is random or fixed. In these contexts, it is natural to consider a case-control strategy where separate samples of individuals to whom an event has already occurred (cases), and those for whom the event has not yet occurred (controls), are obtained. Section 4.1 briefly discusses the extension of the results of Section 3 to case-control designs.

In some applications, the survival time, T, refers to the time between two events in chronological time, for example, the time between infection with HIV and the moment when an infected individual becomes infectious through a specified mechanism (see Jewell, Malani & Vittinghoff, 1994). Current status monitoring of an individual at a single point in chronological time then yields current status observation of T with the random variable C being defined by the difference in chronological time between the 'origin' of T and the monitoring time. Measurement of C assumes that the chronological time of this

COBRA
Collection of Biostatistics
Research Archive

origin is known for all sampled individuals. Situations where this is not known leads to doubly censored current status data which is briefly described in Section 4.2. Some other modifications to standard current status data have also been studied; for example, Shiboski & Jewell (1992) allow for the possibility of a form of staggered entry in an observational study setting.

4.1 Case-Control Sampling

As noted above, it is often useful to consider a case-control sampling scheme. Here, cases refer to a random sample of n_1 observations on C from the sub-population where $T \leq C$, and controls to a random sample of n_0 observations from the sub-population where $T > C$.

Even when the support of T is contained within the support of C, there is an additional identifiability problem that arises in nonparametric estimation of F from case-control samples. Jewell & van der Laan (2002) show that case-control data only identify the odds function associated with F, namely $\log\left[\frac{F(t)}{1-F(t)}\right]$, up to a constant. While this may be sufficient to identify F in an assumed parametric family, it is insufficient nonparametrically. However, additional data regarding the population distribution of cases and controls can be used to identify a specific F with a given odds function that is compatible with the population information.

In particular, suppose that N individuals are sampled from the joint distribution of (Y, C), and that only the numbers of individuals for whom $Y = i, (i = 0, 1)$, say N_0 and N_1, respectively, are observed. Subsequently, case-control data comprised of fixed samples of size $n_0(\leq N_0)$ and $n_1(\leq N_0)$ are selected, by simple random sampling, seperately from the two groups, with $Y = 0$ and $Y = 1$, in the original sample of N. The random variable

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

C is then measured for each of the $n_0 + n_1$ sampled individuals at this stage. In practice, the sampling rates, at this second stage, that is (n_0/N_0) and (n_1/N_1) will usually be quite different.

The supplemented data is thus $\{(y_{ij}, c_{ij}) : i = 0, 1; j = 0, \ldots, n_i; N_0, N_1\}$. Assuming that the sample sizes, n_0 and n_1, are non-informative, a simple consistent nonparametric estimator of F is immediately available by weighting observations inversely proportional to their probability of selection, and using the estimator for standard current status data (Section 3) on this weighted data. Specifically, the weights are (N_0/n_0) for controls and (N_1/n_1) for cases. Jewell & van der Laan (2002) show that this simple estimator is, in fact, the nonparametric maximum likelihood estimator based on case-control data supplemented by knowledge of N_0 and N_1.

This nonparametric estimator assumes knowledge of the population totals N_0 and N_1 (in fact only the ratio N_1/N_0 need be known). Without such information, we can hypothesize a value for N_1/N_0, compute the nonparametric maximum likelihood estimator, and then vary the assumed N_1/N_0 as a sensitivity parameter over a range of plausible values. If N_1/N_0 is allowed to take on all values the corresponding nonparametric maximum likelihood estimators trace out the population odds family associated with any particular choice of N_1/N_0.

For parametric models for F, the situation is not as straightforward, even with knowledge of the supplementary population totals N_0 and N_1, as the weighted and maximum likelihood estimators need not coincide. However, Scott & Wild (1997) provide an elegant iterative algorithm to compute the maximum likelihood estimator of F using data on N_0 and N_1. Their approach is based on the regression model induced for $\Pr(Y = 1|C = c)$,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

and the proposed algorithm is particularly simple when this regression model can be easily fit for randomly sampled (i.e. prospective) data. For example, if F is assumed to follow a Weibull distribution , with hazard $e^a bt^{b-1}$, then $\log - \log[\Pr(Y = 1|C = c)] = a + b\log(c)$, as noted in Section 3.2, that is, a standard generalized linear model with complementary log-log link; the iterative steps in fitting a Weibull distribution to case-control current status data are therefore straightforward since there is standard software that accomodates this form of prospective generalized linear model.

4.2 Doubly Censored Current Status Data

Suppose that the survival variable T measures the length of time between two successive events in chronological time. We refer to these as the initiating and subsequent events, and assume that their occurrence times are given by the random variables I and J, respectively, so that $T = J - I$. We assume that T is independent of I. Now, consider a single monitoring occasion whose chronological time is given by B, independent of I and J, at which point current status information is available on the subsequent event J; that is, we observe whether $J \leq B$ or not. For a random sample of individuals for whom $I \leq B$, such an observation scheme yields current status observations of T, assuming that the random variable I is known for all observations. In particular, we observe the random variable Y which takes the value 1 if $T \leq B - I$, and 0, otherwise. In this case, the induced monitoring time for T is $C = B - I$, so that its distribution is determined by that of I.

An additional complication is introduced when the random variable I is unknown or unobserved. Now, at chronological time B, we merely observe whether either or both of the initiating and subsequent events have occurred by time B, but not the times of either

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

event. Without loss of information on F, we assume that only individuals for whom $I \leq B$ are included in the sample. The observed data is thus reduced to Y^* where $Y^* = 1$ if $I \leq J \leq B$ and $Y^* = 0$ if $I \leq B < J$.

In order for F to be identifiable from such data, we assume that the conditional distribution of I, given that $I \leq B$, is known (Jewell, Malani & Vittinghoff, 1994), although it is allowable that this distribution varies from individual to individual. For convenience, for the ith sampled individual, suppose that the known conditional distribution of I, given that $I \leq B$, is labeled by H_i, and has finite support on some interval (A_i, B_i). Then, we have

$$P_i = Pr(Y_i^* = 1) = \int_0^{C_i} H_i(B_i - T)dF(T), \tag{6}$$

where now $C_i = B_i - A_i$. Further, the conditional likelihood of n observations of this kind is then

$$CL = \prod_{i=1}^{n} P_i^{Y_i^*}(1 - P_i)^{(1-Y_i^*)}. \tag{7}$$

This data is referred to as doubly censored current status data by Rabinowitz & Jewell (1996) since it is a special case of doubly censored survival data as described by DeGruttola & Lagakos (1989). Two applications to data on HIV are given in Jewell, Malani & Vittinghoff (1994). Parametric estimation of F, based on the likelihood (7) is again straightforward in principal.

Nonparametric maximum likelihood estimation of F can be approached by viewing the model as a nonparametric mixture estimation problem (Jewell, Malani & Vittinghoff, 1994). An important special case occurs when H_i is assumed to be Uniform on $[A_i, B_i]$ in

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

which case (6) reduces to

$$P_i \equiv P(C_i) = \frac{1}{C_i}\int_0^{C_i} F(T)dT. \tag{8}$$

Here, P is a distribution function that only depends on C_i and so doubly censored current status data in this case is a sub-model of current status data. Estimation of F, with this assumption on each H_i, is examined in Jewell, Malani & Vittinghoff (1994), van der Laan, Bickel & Jewell (1997), and van der Laan & Jewell (2001). The latter paper shows that the nonparametric maximum likelihood estimator of F is uniformly consistent, and further that the distribution function $P(C)$, defined by (8), is nonparametrically estimated a rate $n^{-2/5}$, indicating the value of the additional structure given in (8) as compared to standard current status data. On the other hand, it is conjectured that F itself can only be estimated at rate $n^{-1/5}$ (see van der Laan, Bickel & Jewell, 1997), although this result and the limiting distribution of the nonparametric maximum likelihood estimator of F remain to be established. Despite the very slow rate of convergence of the nonparametric maximum likelihood estimator, many smooth functionals can still be efficiently estimated, at rate $n^{-1/2}$, using the appropriate functionals of the nonparametric maximum likelihood estimator. An alternative iterative weighted pool-adjacent-violators algorithm is also given for computation of the nonparametric maximum likelihood estimator.

Rabinowitz & Jewell (1996) extend the results of Rabinowitz, Tsiatis & Aragon (1995), for estimation of regression parameters in the accelerated failure time model for T, to doubly censored current status data assuming each H_i to be Uniform. See also van der Laan, Bickel & Jewell (1997).

van der Laan & Andrews (2000) replace the assumption of a Uniform distribution for H_i by a mixture of a point mass and a Uniform, a generalization that arises naturally in

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

partner studies. The presence of a point mass now permits the nonparametric maximum likelihood estimator to converge to F at rate $n^{-1/3}$, as for standard current status data; again smooth functionals can be efficiently estimated based on the nonparametric maximum likelihood estimator at rate $n^{-1/2}$. Some speculation is given there regarding the situation for other forms of H_i.

5 Complex Outcome Processes

It is well-known that the survival random varaible T can be alterrnatively viewed as the time to the 'jump' of a simple 0-1 counting process $X(t)$. In this context, a current status monitoring scheme corresponds with a single cross-sectional observation of the stochastic process $X(t)$. Considering cross-sectional observation of more complex monotone stochastic processes leads to various extensions of simple current status data structures. In particular, current status competing risk data, discussed in Section 5.1, arise when X still only jumps once in each sample path, but now jumps are marked by a discrete set of outcomes, usually the cause of the jump or failure. Section 5.2 investigates the situation where X is now defined by a bivariate pair of binary counting process, (X_1, X_2). In Section 5.3, we return to a univariate X, but now allow for the possibility of two successive jumps—from 0 to 1, and then from 1 to 2. Finally, we briefly examine the case where is X is a general counting process in Section 5.4. For some brief remarks for the case where $X(t)$ is a renewal process, see Jewell & van der Laan (1997).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

5.1 Competing Risk Outcomes

In Section 3, we introduced simple current status data in terms of a single survival random varaible T with an assumed single definition of failure. In some scenarios, failure may be associated with more than one 'cause', leading to the extensive literature on competing risks. For simplicity here, we assume but two competing risks, although all the material readily extends to an arbitrary number of risks.

If J is the random variable that indicates the cause of failure at time T, the two sub-distribution functions of interest are

$$F_j(t) = \mathrm{pr}(T \leq t, J = j),$$

with the overall survival function given by

$$S(t) = 1 - F_1(t) - F_2(t).$$

Jewell, van der Laan & Henneman (2003) consider nonparametric estimation of F_1, F_2 and $F = F_1 + F_2$, when only current status information on survival is available at the monitoring time C, but cause of failure is known whenever failure is seen to have occurred before C. Here, observed data can thus be represented as $Y = (\Delta, \Phi)$ and C, where $\Delta = 1$ if $T \leq C$ with $J = 1$, and $\Phi = 1$ if $T \leq C$ with $J = 2$. This is a special case of competing risk survival data subject to general interval censoring as studied in Hudgens, Satten & Longini (2001). We again assume that C is independent of (T, J), with the implication that we still focus on the conditional likelihood of the data, given C. This is easily seen to be given by

$$CL = \prod_{i=1}^{n} \{F_1(c_i)\}^{\delta_i} \{F_2(c_i)\}^{\phi_i} \{S(c_i)\}^{1-\delta_i-\phi_i}. \qquad (9)$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Ideas for estimation of parametric competing risk models, based on the likelihood (9), apply here much as they do for standard current status data (Jewell, van der Laan & Henneman, 2003).

Since, by definition, $E(\Delta|C) = F_1(C)$ and $E(\Phi|C) = F_2(C)$, simple nonparametric estimators of F_1 and F_2 can be constructed via separate current status estimators based on $(\delta_i, c_i : i = 1, \ldots, n)$ and $(\phi_i, c_i : i = 1, \ldots, n)$, respectively, using the methods of Section 3. A disadvantage of this naive approach is that there is no guarantee that $\hat{F}_1 + \hat{F}_2$ is a distribution function, so that the derived estimator of the overall survival function $\hat{S}(t) = 1 - \hat{F}_1(t) - \hat{F}_2(t)$ may be negative for large t.

An alternative ad hoc approach is developed by Jewell, van der Laan & Henneman (2003) as follows. First, reparameterise F_1 and F_2 in terms of F and F_1. An immediate estimator of F is available from the data (γ_i, c_i), where $\gamma_i = \delta_i + \phi_i$; since $E(, \ = \Delta + \Phi|C) = F(C)$, we can again use the current status methods of Section 3 to produce $\hat{F}$ as an estimator of F. Now, restrict attention to the data where $\Delta + \Phi = 1$, and define a constructed variable Z by:

$$Z = F(C)\Delta.$$

Note that $E(Z|C, \Delta + \Phi = 1) = F(C) \times \Pr(\Delta = 1|C, \Delta + \Phi = 1) = F_1(C)$. This suggests an isotonic regression estimator of the constructed data, $\hat{F}\delta_i$, against c_i, using only observations where $\delta_i + \phi_i = 1$, yielding an estimator $\hat{F}_{1p}$. Similarly, the isotonic regression of $\hat{F}\phi_i$ against c_i, will provide the analogous estimator $\hat{F}_{2p}$ for F_2. Again, $\hat{F}_{1p}(t) + \hat{F}_{2p}(t)$ may exceed one for large t, although this may be less likely than for the naive approach since the isotonic regressions are here based on $\hat{F}(\cdot)\Delta$ and $\hat{F}(\cdot)\Phi$, both smaller than the respective dependent variables, Δ and Φ, for the previous estimators.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Neither of these approaches yields the nonparametric maximum likelihood estimator in general. The difference between the second approach and the nonparametric maximum likelihood estimators, say F_{1n} and F_{2n}, hinges on variation in the support of F_{1n} and F_{2n}; that is, the nonparametric maximum likelihood estimator uses the fact that F_{2n} may be non-constant between support points of F_{1n}. However, Jewell, van der Laan & Henneman (2003) show that smooth functionals of either F_1 or F_2 are efficiently estimated using the appropriate functionals of either of the two simpler estimators of F_1 and F_2, respectively. Simulations show that the naive current status estimator (which ignores cause of failure data) and the full NPMLE of F have very similar performances in general; this is to be expected as there can be no value in knowing the cause of failure if one is solely interested in estimating the overall survival distribution.

The general EM algorithm can be used to compute the nonparametric maximum likelihood estimators of F_1 and F_2. However, Jewell & Kalbfleisch (2002) provide a much faster algorithm that generalizes pool-adjacent-violators. Their approach can most easily be described by restating the problem as follows: let (A_i, B_i, D_i) be a trinomial variate with index n_i and probabilities p_i, q_i, $1 - p_i - q_i$, independently for $i = 1, ..., k$. We wish to maximize the log likelihood function

$$\ell(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k} \{a_i \log p_i + b_i \log q_i + d_i \log[1 - p_i - q_i]\}, \qquad (10)$$

where $\mathbf{p} = (p_1, ..., p_k)$ and $\mathbf{q} = (q_1, ..., q_k)$. The parameter space,

$$\Theta = \{(\mathbf{p}, \mathbf{q}) : 0 \leq p_1 \leq ... \leq p_k;\ 0 \leq q_1 \leq ... \leq q_k; \mathbf{1} - \mathbf{p} - \mathbf{q} \geq 0\},$$

is a compact convex set in $\mathcal{R}^{2k}$. Equivalence to maximization of the conditional likelihood given in (9) is easily seen by ordering and grouping observations according to the size of the

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

c_is; then, for each distinct c_i, let A_i be the number of observations with monitoring time c_i for which $\delta_i = 1$, with a similar definition for B_i and D_i; in previous notation, $p_i = F_1(c_i)$ snd $q_i = F_2(c_i)$.

An iterative algorithm to find the estimator of $(\mathbf{p}, \mathbf{q})$ that maximizes (10) is given in Jewell & Kalbfleisch (2002) using the strategy of maximizing over the vector $\mathbf{p}$, holding $\mathbf{q}$ fixed, and vice-versa. These maximizations are achieved using a variation on the pool-adjacent-violators algorithm where pooling now involves solution of a polynomial equation rather than simple averaging. Care is needed with regard to estimates of the vectors $\mathbf{p}, \mathbf{q}$ for both the first and last set of entries. Further work is required to establish the limiting distribution of the nonparametric maximum likelihood estimator or other techniques that may be used to provide confidence limits for specific values of F_1 or F_2; one approach is to approximate such 'parameters' by smooth functionals of F_1 and F_2.

Jewell, van der Laan & Henneman (2003) and Jewell & Kalbfleisch (2002) illustrate the application of the nonparametric estimators discussed in this section to an example on womens' age at menopause, where the outcome of interest (menopause) is associated with two competing causes, natural and operative menopause. Jewell, van der Laan & Henneman (2003) also consider the situation where failure times for one risk are observed exactly whenever failure due to that cause occurs prior to the monitoring time.

5.2 Bivariate Current Status Data

Consider a study in which interest focuses on the bivariate distribution F of two random survival variables (T_1, T_2), neither of which can be directly measured. Rather, for each individual, we observe, at a random monitoring time, C, whether T_j exceeds C or not for

COBRA
Collection of Biostatistics
Research Archive

each $j = 1, 2$. That is, on each subject, we observe:

$$(Y_1 \equiv I(T_1 \leq C), Y_2 \equiv I(T_2 \leq C), C).$$

Again, C is assumed independent of (T_1, T_2). Wang and Ding (2000) refer to this data structure as bivariate current status data. Conditional on the observed values of C, the likelihood of a set of n independent observations of this kind is given by

$$CL = \prod_{i=1}^{n} F_3(c_i)^{y_1 y_2}(1+F_3-F_1-F_2)(c_i)^{(1-y_1)(1-y_2)}(F_1-F_3)(c_i)^{y_1(1-y_2)}(F_2-F_3)^{(1-y_1)y_2}, \quad (11)$$

where $F_1(t) = P(T_1 \leq t)$, $F_2(t) = P(T_2 \leq t)$ and $F_3(t) = P(T_1 \leq t, T_2 \leq t)$ are marginal distributions of F along the two axes and the diagonal, respectively. It follows that only these three univariate cdf's F_1, F_2 and F_3 are identified from the data. In particular, the complete bivariate distribution, F, is not identifiable; however, the dependence measure $F_3 - F_1F_2$ is identifiable from the data, so that some assessment of independence of T_1 and T_2 is possible. Wang & Ding (2000) considered a semiparametric copula model for F, parametrized by the marginals, F_1 and F_2, and a single real valued parameter α which represents a measure of dependence between T_1 and T_2.

Note that 'marginal' nonparametric current status estimators of F_j, $j = 1, 2, 3$, are available. With $Y_3 = Y_1Y_2$, $F_j(t)$ can be represented in terms of a monotonic regression of Y_j on C since $F_j(t) = E(Y_j \mid C = t)$, for $j = 1, 2, 3$; we can thus use the current status estimator based on (Y_j, C) to estimate F_j. This estimator is, of course, the nonparametric maximum likelihood estimator based on the reduced data (Y_j, C). From the results of Section 3, it follows that these reduced data nonparametric maximum likelihood estimators are consistent and converge, under appropriate conditions, at rate $n^{-1/3}$, to known asymptotic distributions. In spite of the simplicity of these three reduced data nonparametric maximum likelihood estimators relative to the full nonparametric nonparametric

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

maximum likelihood estimator based on (11), van der Laan & Jewell (2002b) show that, at most data generating distributions, the reduced data nonparametric maximum likelihood estimators yield efficient estimators of smooth functionals of (F_1, F_2, F_3). If interest focuses on the possible dependence of T_1 and T_2, then estimates of appropriately chosen functionals of $F_3 - F_1F_2$ may be examined based on these reduced data nonparametric maximum likelihood estimators.

We can restate the problem on nonparametric maximization of the likelihood (11) in terms of a multinomial random variable as follows: let (A_i, B_i, D_i, E_i) be a four-state multinomial variate with index n_i and probabilities p_i, q_i, r_i, $1 - p_i - q_i - r_i$, independently for $i = 1, ..., k$. Having ordered the observations according to the c_is, maximizing the likelihood (11) is equivalent to maximization of the log likelihood function

$$\ell(\mathbf{p}, \mathbf{q}, \mathbf{r}) = \sum_{i=1}^{k} \{a_i \log p_i + b_i \log(q_i - p_i) + d_i \log(r_i - p_i) + e_i \log(1 + p_i - q_i - r_i)\}, \quad (12)$$

where $\mathbf{p} = (p_1, ..., p_k)$, $\mathbf{q} = (q_1, ..., q_k)$ and $\mathbf{r} = (r_1, ..., r_k)$ with the parameter space defined by $\Theta = \{(\mathbf{p}, \mathbf{q}, \mathbf{r}) : 0 \leq p_1 \leq ... \leq p_k;\ 0 \leq q_1 \leq ... \leq q_k;\ 0 \leq r_1 \leq ... \leq r_k; \mathbf{q} - \mathbf{p} \geq 0; ; \mathbf{r} - \mathbf{p} \geq 0; \mathbf{1} - \mathbf{p} - \mathbf{q} - \mathbf{r} \geq 0\}$. Note that Θ is again a compact convex set in $\mathcal{R}^{3k}$. This formulation is obtained by setting, for each distinct c_i, A_i to be the number of observations with monitoring time c_i for which $y_3 = 1$, B_i the number of observations with monitoring time c_i for which $y_1 = 1 - y_2 = 1$ and D_i the number of observations with monitoring time c_i for which $1 - y_1 = y_2 = 1$. With regard to the parameters, we have $p_i = F_3(c_i), q_i = F_1(c_i)$ and $r_i = F_2(c_i)$. With this respecification of the problem, it would be of considerable value to derive an iterative algorithm akin to the Jewell & Kalbfleisch (2002) approach of Section 5.1; the main issue here is the appropriate handling of the 'edge' effects of the constraints linking $\mathbf{p}, \mathbf{q}$ and $\mathbf{r}$.

COBRA
BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

We have assumed that the monitoring time C is the same for both T_1 and T_2. In some applications, the monitoring times may differ so that current status information on T_i is obtained at time $C_i, i = 1, 2$, where the random or fixed C_1 is not the same as C_2. This is a substantially more complex problem than the case considered here, and, to date, there is little work that has addressed this version of bivariate current status data.

5.3 Outcomes with Intermediate Stage

A special form of bivariate survival data arises from observations on the time to failure where all individuals pass through an intermediate stage prior to failure. In this situation, let T_1 represent the time from the origin until the intermediate event occurs, with T_2 being the time to failure, Here, necessarily, $T_2 \geq T_1$. Current status observation of this process at a monitoring time C reveals whether an individual has failed by time C or not, and in the latter case, whether the intermediate event has occurred by time C or not. As a result, the observed data is then given by the random variable

$$(Y_1 \equiv I(T_1 \leq C), Y_2 \equiv I(T_2 \leq C), C).$$

Unlike arbitrary bivariate current status data, there are only three possible outcomes for $\mathbf{Y} \equiv (Y_1, Y_2)$, namely $(0,0), ((1,0)$, and $((1,1)$. Once more, C is assumed independent of (T_1, T_2). A variant of this data structure where exact information is available on T_2 whenever $T_2 \leq C$, is studied in van der Laan, Jewell & Petersen (1997).

Conditional on the observed values of C, the likelihood of a set of n independent observations of this kind is given by

$$CL = \prod_{i=1}^{n} F_2(c_i)^{y_1 y_2}(1 - F_1)(c_i)^{(1-y_1)(1-y_2)}(F_1 - F_2)(c_i)^{y_1(1-y_2)}, \tag{13}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where $F_1(t) = P(T_1 \leq t)$, $F_2(t) = P(T_2 \leq t)$ are the marginal distributions of T_1 and T_2, respectively. It follows that just the two marginal cdf's F_1 and F_2 are identified. As for bivariate current status data, the complete bivariate distribution of (T_1, T_2) is not identifiable; an unfortunate consequence of this is that the data contains no information on the possibility of dependence between T_1 and $T_2 - T_1$, the recurrence times of the fiirst and second event, respectively. Thus, the relationship between recurrence times can only be investigated via a prior model assumption whose dependence structure cannot be verified nonparametrically from the data.

This data structure is a special case of current status observation on a counting process which we discuss in more detail in Section 5.4. Here, we point out that, as in Sections 5.1 and 5.2, we can restate the problem on nonparametric maximization of the likelihood (13), now in terms of a trinomial random variable. Let (A_i, B_i, D_i) be a trinomial variate with index n_i and probabilities p_i, q_i, $1 - p_i - q_i$, independently for $i = 1, ..., k$. Nonparametric maximization of (13) is equivalent to maximization of the log likelihood function

$$\ell(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k} \{a_i \log p_i + b_i \log q_i + d_i \log[1 - p_i - q_i]\}, \qquad (14)$$

where $\mathbf{p} = (p_1, ..., p_k)$ and $\mathbf{q} = (q_1, ..., q_k)$, with the parameter space defined by $\Theta = \{(\mathbf{p}, \mathbf{q}) : 0 \leq p_1 \leq ... \leq p_k;\ 0 \leq q_1 \leq ... \leq q_k; \mathbf{1} - \mathbf{p} - \mathbf{q} \geq 0\}$, a compact convex set in $\mathcal{R}^{2k}$. This equivalence is achieved as before by setting A_i to be the number of observations with $y_1 = (1 - y_2) = 1$ and monitoring time a distinct c_i from amongst the ordered monitoring times; similarly, B_i is the number of observations with $y_1 = y_2 = 1$ and monitoring time c_i. The parameters $p_i = F_2(c_i)$ and $q_i = F_1(c_i)$. With this formulation at hand, it would be of interest to describe an appropriate version of the Jewell & Kalbfleisch (2002) iterative algorithm, with again the edge effects being important.

COBRA
Collection of Biostatistics
Research Archive

5.4 Counting Processes

We now consider current status monitoring of a counting process $X(t) = \sum_{j=1}^{k} I(T_j \leq t)$, where, for $j = 1, \ldots, k$, T_j is the random variable which measures the time at which X jumps from $j-1$ to j. Necessarily $T_1 \leq T_2 \leq \cdots \leq T_k$. Now assume that data arises from a sample of n current status observations of the process X, where the monitoring times are described by the random variable C, assumed independent of X. Note this corresponds to simple cross-sectional observation of X. Jewell & van der Laan (1995) describe several possible applications where this data structure arises naturally. Note that allowing the marginal distributions, F_j of T_j, $j = 1, \ldots, k$, to each have a possible point mass at infinity accommodates data structures where individuals may "stop" after one jump, or two, etc. Further, individuals are not therfore required all to pass through the exace same number of stages or jumps. Further, choosing the finite number of states to be large enough accommodates any practical application, so that the case of an infinite number of states is only of theoretical import.

The data is thus a sample of indepependent and identically distributed observations on the random variable $(X(C), C)$. As we have seen in previous sections, particularly Section 5.3, it is easy to see that, nonparametrically, the likelihood only depends on the marginal distributions F_j. An unfortunate consequence of this is again that, absent some additional model assumptions, the data tells us nothing about the interesting possibility of dependence among the recurrence times $T_1, T_2 - T_1, \ldots, T_j - T_{j-1}, \ldots$. Nonparametric maximum likelihood estimation of $F_1, \ldots, F_k$ requires some form of iterative algorithm—see Section 5.3. However, as we observed in Section 5.1 and 5.2, direct estimation of any single F_j is possible using the standard current status observations, $(Y = I(T_j \leq$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$C), C)$, and estimates of smooth functionals of F_j can be based on this simple estimator, enjoying all the asymptotic properties outlined in Section 3. Note that this estimator ignores apparently useful information given in $X(C)$ beyond the simple fact of whether $X(C) \leq j$ or not. Nevertheless, van der Laan & Jewell (2002a) show that, at many data generating distributions, the simple standard current status estimators of F_j yield efficient estimators of smooth functionals. These simple current status estimators are not the full nonparametric maximum likelihood estimators, and van der Laan & Jewell (2002a) discuss in detail the differences between the two approaches, thereby giving insight into why the nonparametric maximum likelihood estimatorshows no asymptotic gain for such functional estimation.

In the above, we have focused on estimation of F_j, the marginal distribution of T_j, for $j = 1, \ldots, k$. In some applications, particularly when the number of states, k, is large there may be little interest in each individual marginal distribution. In such cases, a simple function of the marginal distributions, namely the so-called mean function, $\Lambda(t) = E(X(t))$, may however be of considerable importance. It is easy to see that

$$\Lambda(t) = \sum F_j(t), \tag{15}$$

a description that is applicable even if the number of jumps can be arbitrarily large so that the above sum has an infinite number of terms. The mean function may be particularly useful as a method to summarize the effects of covariates on $X(t)$. Sun & Kalbfleisch (1993) consider estimation of Λ, discuss regression models that allow this mean function to vary across covariate groups, and consider application of the ideas to multiple tumor data from a tumorgenicity experiment. Note that, for current status observation on $X(t)$ at random monitoring times C with no covariates, the mean function is isotonic in the observed Cs, so

COBRA
Collection of Biostatistics
Research Archive

that many of the ideas of Section 3 can be immediately applied to estimation of Λ including the pool-adjacent-violators algorithm.

6 Conclusion

This paper has reviewed recent advances in the understanding of nonparametric estimation based on various forms of current status data. Throughout a key assumption has been independence between the monitoring time variable C and the survival random variable, T, or counting process, X, of interest. An important future area of study with current status data concerns the relaxation of this assumption. For example, suppose, for a survival random variable T and random monitoring time C, we observe the data structure $Y = (I(T \leq C), C, \bar{L}(C))$ that includes observation of covariate processes L up to time C. The assumption of independence between T and C can now be assumed *conditional* on the observed $\bar{L}(C)$. This therefore allows dependence between the monitoring time C and T that arises solely through $\bar{L}(C)$. To illustrate the importance of this extension, consider an animal tumorgenicity experiment designed to estimate the distribution of time to development of an occult tumor. Suppose that $L(u)$ includes the weight of the experimental animal at time u, and that $Y = (I(T \leq C), C, \bar{L}(C))$ is observed. A reasonable alternative to choosing monitoring times completely at random is to increase the 'hazard' of monitoring shortly after an animal begins to lose weight as reflected in measurements of L; this is likely to improve efficiency in estimation if the monitoring time is thereby closer to the time of tumor onset (i.e. T). This monitoring scheme introduces dependence between C and T, and estimators, discussed in Section 3, that ignore this dependence will be biased. For the extended current status data structure $Y = (I(T \leq C), C, \bar{L}(C))$, van

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

der Laan & Robins (1998) develop locally efficient estimators for smooth functionals of F, the distribution function of T. An important open problem of interest involves the use of these results in choosing optimal, or close to optimal, designs for the dynamic selection of monitoring times C that depend on concurrent observation of key covariates within L.

Finally, since current status data corresponds with taking a single cross-sectional observation on individual survival processes. it is natural to consider similar questions where multiple cross-sectional observations are availble at differing monitoring times for each individual. Data of this kind are often referred to as panel data. In the context of the single survival random variable T of Section 3, this monitoring scheme leads to interval-censored data, case II (Groenboom & Wellner, 1992). There is a parallel extensive literature on estimation problems of the kind considered here, based on this more informative and general form of interval censored data, that deserves a similar review article of recent advances. For helpful introductions, see Sun (1998) and Huang & Wellner (1997). Panel data has also been considered in the context of counting processes as in Section 5.4 by Sun & Kalbfleisch (1995), Wellner & Zhang (2000) and others.

References

Andrews, C., van der Laan, M. & Robins, J.M. (2002). *University of California Berkeley Division of Biostatistics, Working Paper Series. Working Paper 110. http://www.bepress.com/ucbbiostat/paper110.*

Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641-647.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972) *Statistical Inference under Order Restrictions.* New York: Wiley.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273-277.

Cox, D.R. (1972). Regression models with life tables (with discussion). *J. Royal Statist. Soc. B* **34**, 187-220.

DeGruttola, V. & Lagakos, S.W. (1989). Analysis of doubly-censored survival data with application to AIDS. *Biometrics* **45**, 1-11.

Diamond, I.D., McDonald, J.W. & Shah, I.H. (1986). Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography* **23**, 607-620.

Diamond, I.D. & McDonald, J.W. (1991). The analysis of current status data. In *Demograhic Applications of Event History Analysis* Trussel, J, Hankinson, R. & Tilton, J. eds., Oxford, U.K.: Oxford University Press.

Doksum, K.A. & Gasko, M. (1990). On a correspondence between models in binary regression and in survival anlysis. *Int. Statist. Review* **58**, 243-252.

Gart, J.J., Krewski, D., Lee, P.N., Tarone, R.E. & Wahrendorf, J. (1986). *Statistical Methods in Cancer Research, Volume III,The Design and Analysis of Long-term Animal Experiments* IARC Scientific Publications No. 79. Lyon: International Agency for Research on Cancer.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Groeneboom, P. & Wellner, J.A. (1992) *Nonparametric Maximum Likelihood Estimators for Interval Censoring and Denconvolution.* Boston: Birkhäuser-Boston.

Grummer-Strawn, L.M. (1993). Regression analysis of current status data: An application to breast feeding. *J. Amer. Statist. Assoc.* **88**, 758-765.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540-568.

Huang, J. & Wellner, J.A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I. *Statist. Neerlandica* **49**, 153-163.

Huang, J. & Wellner, J.A. (1997) Interval censored survival data: A review of recent progress. In *Proceedings of First Seattle Conference in Biostatistics* Lin, D-Y. ed., Springer Verlag, 123-169.

Hudgens, M.G., Satten, G.A. & Longini, I.M., Jr. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics* **57**, 74-80.

Jewell, N.P. & Kalbfleisch, J.D. (2002). Maximum likelihood estimation of ordered multinomial parameters. To appear.

Jewell, N.P., Malani, H. & Vittinghoff, E. (1994) Nonparametric estimation for a form of doubly censored data with application to two problems in AIDS. *J. Amer. Statist. Assoc.* **89**, 7-18.

Jewell, N.P., Shiboski, S. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics* **46**, 1133-1150.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Jewell, N.P. & van der Laan, M. (1995). Generalizations of current status data with applications. *Lifetime Data Analysis* **1**, 101-109.

Jewell, N.P. & van der Laan, M. (1997) Singly and doubly censored current status data with extensions to multi-state counting processes. In *Proceedings of First Seattle Conference in Biostatistics* Lin, D-Y. ed., Springer Verlag, 171-184.

Jewell, N.P. & van der Laan, M. (2002). Case-control current status data. To appear.

Jewell, N.P., van der Laan, M. & Henneman, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika*, to appear.

Keiding, N. (1997). Age-specific incidence and prevalence: a statistical perspective (with discussion). *J. Royal Statist. Soc. A* **154**, 371-412.

Keiding, N., Begtrup, K., Scheike, T.H. & Hasibeder, G. (1996). Estimation from current-status data in continuous time. *Lifetime Data Analysis* **2**, 119-129.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19**, 724-740.

Mukerjee, R. (1988). Monotone nonparametric regression. *Ann. Statist.* **16**, 741-750.

Rabinowitz, D. & Jewell, N. P. (1996). Regression with doubly censored current status data. *J. Royal Statist. Soc. B* **58**, 541-550.

Rabinowitz, D., Tsiatis, A. & Aragon, J. (1995). Regression with interval censored data. *Biometrika* **82**, 501-513.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Rossini, A. & Tsiatis, A.A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *J. Amer. Statist. Assoc.* **91**, 713-721.

Scott, A.J., Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57-71.

Shiboski, S.C. (1998a). Partner Studies. In *The Encyclopedia of Biostatistics*, P. Armitage and T. Colton (eds.) 3270-3275.

Shiboski, S.C. (1998b). Generalized additive models for current status data. *Lifetime Data Analysis* **4**, 29-50.

Shiboski, S.C. & Jewell, N.P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *J. Amer. Statist. Assoc.* **87**, 360-372.

Sun, J. (1998). Interval Censoring. In *The Encyclopedia of Biostatistics*, P. Armitage and T. Colton (eds.) 2090-2095.

Sun, J. & Kalbfleisch, J.D. (1993). The analysis of current status data on point processes. *J. Amer. Statist. Assoc.* **88**, 1449-1454.

Sun, J. & Kalbfleisch, J.D. (1995). Estimation of the mean function of point processes based on panel count data. *Statist. Sinica* **5**, 279-290.

van der Laan, M. & Andrews, C. (2000). The nonparametric maximum likelihood estimator in a class of doubly censored current status data models with application to partner studies. *Biometrika* **87**, 61-71.

van der Laan, M.J., Bickel, P.J. & Jewell, N.P. (1997). Singly and doubly censored current status data: Estimation, asymptotics and regression. *Scand. J.*

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Statist. **24**, 289-308.

van der Laan, M. & Jewell, N. P. (2001). The NPMLE in the doubly censored current status data model. *Scand. J. Statist.* **28**, 537-547.

van der Laan, M. & Jewell, N. P. (2002a). Nonparametric efficient estimation with current status data and right-censored data structures when observing a marker at the censoring time. *Annals of Statistics*, to appear.

van der Laan, M. & Jewell, N. P. (2002b). Bivariate current status data. To appear.

van der Laan, M., Jewell, N.P. & Petersen, D. (1997). Efficient estimation of the lifetime and disease onset distribution *Biometrika* **84**, 539-554.

van der Laan, M.J. & Robins, J.M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *J. Amer. Statist. Assoc.*, **93**, 693-701.

Wang, W. & Ding, A.A. (2000). On assessing the association for bivariate current status data. *Biometrika* **87**, 879-893.

Wellner, J.A. & Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28**, 779-814.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press