

Locally Efficient Estimation of Regression Parameters Using Current Status Data

Chris Andrews^a, Mark van der Laan^{b,*}, James Robins^c

^a*Oberlin College, OH, USA*

^b*University of California, Berkeley, CA, USA*

^c*Harvard University, Boston, MA, USA*

Abstract

In biostatistics applications interest often focuses on the estimation of the distribution of a time-variable T . If one only observes whether or not T exceeds an observed monitoring time C , then the data structure is called current status data, also known as interval censored data, case I. We consider this data structure extended to allow the presence of both time-independent covariates and time-dependent covariate processes that are observed until the monitoring time. We assume that the monitoring process satisfies coarsening at random.

Our goal is to estimate the regression parameter β of the regression model $T = Z^\top \beta + \epsilon$. The curse of dimensionality implies no globally-efficient nonparametric estimator with good practical performance at moderate sample sizes exists. We present an estimator of the parameter β that attains the semiparametric efficiency bound if we correctly specify (a) a model for the monitoring mechanism and (b) a lower dimensional model for the conditional distribution of T given the covariates. In addition, our estimator is robust to model misspecification. If only (a) is correctly specified, the estimator remains consistent and asymptotically normal. We conclude with a simulation experiment and a data analysis.

Key words: Extended Current Status Data, Asymptotically Linear Estimator, Influence Curve, Efficient, Regression, Coarsening at Random, One-step Estimator

* Corresponding author. Phone: 510-643-9866. Fax: 510-643-5163. Warren Hall 7360, Division of Biostatistics, School of Public Health, University of California, Berkeley, CA 94720-7360 USA

Email address: laan@stat.berkeley.edu (Mark van der Laan).

1 Introduction

1.1 Regression with Current Status Data

Consider a study in which interest lies in the distribution of a random variable, T , that is never observed. Rather, for each individual, we observe at a random monitoring (censoring) time, C , whether T exceeds C . This data structure $(C, \Delta = I(T \leq C))$ is called *current status data*. Our goal is to estimate the parameter vector β of the regression model $T = Z^\top \beta + \epsilon$ where Z is a vector of time-independent covariates. The conditional distribution of the error ϵ given Z has location parameter equal to zero but has an otherwise unrestricted conditional distribution. In addition to Z , time-independent covariates and time-dependent covariate processes up till monitoring time C , denoted by $\bar{L}(C) = \{L(s) : s \leq c\}$, may be available. These covariates explain any dependence between the time T and the monitoring time C and might be used to improve estimation of β . The observed data then is $(C, \Delta = I(T \leq C), Z, \bar{L}(C))$. Our regression model includes the accelerated failure time model because we can transform the chronological variables (e.g., $T = \log(T_*)$ and $C = \log(C_*)$).

Note that we do not specify a parametric family for the error distribution. Furthermore, we do not assume that the error ϵ is independent of Z . Rather, we only assume that the conditional distribution of ϵ given Z has a specified location parameter equal to zero. That is, in order to make β identifiable, we assume

$$E[K(\epsilon) \mid Z] = E[K(T - Z^\top \beta) \mid Z] = 0 \quad (1)$$

where $K(\cdot)$ is a known, monotone function. If $K(\epsilon) = \epsilon$, then equation (1) implies the conditional mean given Z of the error distribution is zero. However, estimation of the mean is quite difficult with current status data because the distribution of the monitoring mechanism must extend as far as the tails of the distribution of T . Thus other measures of center may be advantageous or necessary.

The conditional median model is obtained when $K(\epsilon) = I(\epsilon < 0) - 1/2$. Our estimators require a smoother $K(\cdot)$ than this because the median is not \sqrt{n} -estimable. A convenient family is $K(\cdot) = 2\Phi(\cdot) - 1$ where Φ is a (typically symmetric, mean zero) continuous distribution function. If the mass of Φ is concentrated near zero, we have a “smoothed median”; if Φ has large variance, we have a trimmed mean. We propose to choose a K with compact support $[-\tau, \tau]$ for some user-supplied or data-determined τ .

As an example, consider the following idealized mouse tumorigenicity exper-

iment designed to investigate the relationship between the time, T , until the development of liver adenoma and the dose level, Z , of a suspected tumorigen. Suppose study mice are randomly allocated to dose groups and that liver adenomas are never, in themselves, the primary cause of an animal's death. Therefore, each mouse is sacrificed (monitored) at a random time C . At autopsy it is determined whether a tumor has developed before C . In such studies, it is easy to collect daily measurements of the weight of each mouse prior to sacrifice. Let $L(u)$ be the weight at time u and let $\bar{L} = L(\cdot)$ be the entire weight process. Only the weight process up to time C is observed. Thus for each mouse $Y = (C, \Delta = I(T \leq C), Z, \bar{L}(C))$ is observed, which we consider as a censored observation of the full data $X = (T, Z, \bar{L})$. Because mice with liver adenomas tend to lose weight, $\bar{L}(C)$ and T are associated.

One reasonable monitoring scheme is to increase the hazard of monitoring shortly after a mouse begins to lose weight. If the time of sacrifice can be made closer to the time of tumor onset then the variance of the estimator is lower. This monitoring scheme introduces dependence between C and T . Estimators that ignore this dependence will be biased. Collecting information on a surrogate process and allowing the censoring time to depend on it is a superior design to carcinogenicity experiments that require independent censoring.

In the mouse experiment the dependence between C and T is only through the observed covariates. That is, the hazard of censoring at time t , given the full (unobserved) data $X = (T, Z, \bar{L})$, is only a function of Z and the observed portion of the covariate process, $\bar{L}(t)$:

$$\lambda_C(t | X) = \lambda_C(t | Z, \bar{L}(t)). \quad (2)$$

This implies $G(\cdot | X)$, the conditional distribution function of C , satisfies coarsening at random [1]. Coarsening at random (CAR) was originally formulated by Heitjan and Rubin [2] and generalized by Jacobsen and Keiding [3] and Gill, et al. [4].

Our proposed estimator of β is consistent and asymptotically normal if we succeed in consistently estimating $\lambda_C(\cdot | X)$ at a suitable rate under the assumption (2). One such case is the idealized experiment described above where $\lambda_C(t | Z, \bar{L}(t))$ is known by design because it is under the control of the investigator (so estimation of $\lambda_C(t | Z, \bar{L}(t))$ is not even necessary). In general, a correctly specified semiparametric model that admits a consistent estimator for $\lambda_C(t | Z, \bar{L}(t))$ can be used. In this paper, we emphasize modelling $\lambda_C(t | Z, \bar{L}(t))$ by a time-dependent Cox proportional hazards model:

$$\lambda_C(t | Z, \bar{L}(t)) = \lambda_0(t) \exp(\eta^\top V(t)), \quad (3)$$

where $V(t)$ is a function of $(Z, \bar{L}(t))$. van der Laan and Robins [5] explain

why modelling the monitoring mechanism under CAR is a sensible approach to fight the curse of dimensionality in high dimensional models. Our model for the observed data distribution is now specified since the observed data distribution $P_{F_X, G}$ of Y is indexed by the full data distribution F_X , which needs to satisfy the regression model (1), and the conditional distribution $G(\cdot | X)$, which needs to satisfy a semiparametric model such as (3).

To have identifiability of β , we need to assume the conditional density function $g(\cdot | X)$ of the monitoring process is located correctly relative to the support of T and the location parameter K . A sufficient condition is that $g(c | X)$ must be bounded away from zero when both $K'(c - Z^\top \beta)$ and $1 - F(c | Z, \bar{L}(c))$ are non-zero. If T is unbounded, K' must have finite support. Minimal conditions are provided in the Appendix.

Our estimator also uses an estimator of $F(t | Z, \bar{L}(u)) = P(T \leq t | Z, \bar{L}(u))$ for various u and t . By the curse of dimensionality, one will need to specify a lower dimensional working model for this conditional distribution and estimate it accordingly. The resulting estimator is locally efficient in the sense that it is asymptotically efficient if the working model contains the truth and it remains consistent and asymptotically normal otherwise. Thus our estimator uses time-dependent covariate information, such as the weight history of the mouse up till time u , to predict the time T till onset thereby recovering information lost due to censoring. To illustrate the potential gain possible, if the weight process perfectly predicts T , then our estimator is asymptotically equivalent with the Kaplan-Meier estimator if we specify a correct model for $F(t | Z, \bar{L}(u))$.

Current practice is to sacrifice the mice at one point in time. Since our methodology shows that sophisticated mouse experiments can be nicely analyzed, we hope that experiments of the type above will be carried out in the future. In section 5 we analyze a cross sectional study to estimate the time-till-transmission distribution in a previously analyzed HIV-partner study. In this data analysis we estimate the effects of ‘‘History of Sexually Transmitted Disease’’ and ‘‘Condom Use’’ in a model $\log(T) = Z^\top \beta + \epsilon$ (which thus includes the accelerated failure time model as a submodel) while using covariates outside the model to allow for informative censoring and to improve efficiency. It is important to note that such an analysis is not possible with any of the existing methods since these methods assume that there are no relevant covariates outside the regression model.

1.2 Previous work and comparison with our results

There is a large literature on estimation of the distribution of T with current status data when covariates are absent: Diamond, et al. [6], Jewell and Shi-

boski [7], Diamond and McDonald [8], Keiding [9], Sun and Kalbfleisch [10], Groeneboom and Wellner [11], Jewell et al. [12], van de Geer [13], Huang and Wellner [14], and several others. van der Laan and Robins [5] consider estimation of the distribution of T with current status data in the presence of time-dependent covariate processes and time-independent covariates, using them to improve efficiency and allow for informative monitoring schemes.

Several authors have investigated estimation of regression parameters using current status data, (C, Δ) , together with a time-independent covariate, Z . Rabinowitz, et al. [15] fit an accelerated failure time model $\log(T) = Z^\top \beta + \epsilon$ that requires error ϵ to be independent of the covariates Z . Huang [16] derives an efficient estimator of the regression parameters of the proportional hazards model. Rossini and Tsiatis [17] assume a semiparametric proportional odds regression model and carry out sieve maximum likelihood estimation. In each case the monitoring time may depend on the covariates of the model, Z , but not on additional covariates. Shen [18] fits a linear regression model with current status data and time-independent covariates. In each of these references all covariates that explain the dependence between C and T must be included in the model for T . Because the models are for time-independent covariates only, no time-dependent covariates can be used to explain the dependence between C and T . None of these limitations apply to our approach. In addition, our approach provides in general a mapping from full-data estimating functions to observed data estimating functions and thus provides the class of all estimators for any well understood full data model.

We would like to stress the implication of our results for the accelerated failure time model as studied by Rabinowitz, et al. [15]. Consider our model with the additional restriction on the regression model that ϵ is independent of Z . Our restricted model generalizes the problem of estimation of β in the accelerated failure time model of Rabinowitz, et al. [15] based on current status data, namely by allowing the presence of additional time-dependent and time-independent covariates. The literature does not provide an estimator in this estimation problem. However, because this restricted model is a submodel of our model our locally efficient estimator (e.g. using as working model the accelerated failure time model) yields a closed-form, consistent, and asymptotically normally distributed estimator of the regression parameters in the accelerated failure time model. This estimator will be efficient in the accelerated failure time model and will remain consistent and asymptotically normal when the monitoring mechanism depends on the additional (time-dependent) covariates. Furthermore, it will still be consistent if the error distribution is not independent of Z , but $E(K(\epsilon) | Z) = 0$.

The next two sections are the heart of the paper. In section 2 we present the locally efficient estimator, details for implementing the estimator, and some ideas of efficiency theory and one-step estimation. In section 3 (and the

Appendix) we prove consistency, asymptotic linearity and local efficiency of our estimator. Two simulations that demonstrate some asymptotic and finite sample properties of the estimators are presented in section 4. An analysis of the California Partners' Study of HIV infectivity is given in section 5 and finally we have some closing remarks.

2 Estimation

We define estimating functions for β as functions of the data Y and parameters, including the parameter of interest β , that are orthogonal (i.e. covariance equal to zero) to all nuisance scores when evaluated at the true parameter values. We will first present the class of all estimating functions of the observed data model (which are orthogonal to the nuisance tangent space) defined by the regression model (1) and CAR (2) on G . This set can be represented as the range of a mapping $D \rightarrow IC(D) \equiv IC_0(D) - IC_{nu}(D)$ from estimating functions of the full data model. The estimating functions for β in the model for the full data (T, Z, \bar{L}) are of the form $D(T, Z) = h(Z)K_\beta(T, Z)$ for some $h(Z)$, where $K_\beta(T, Z) = K(T - Z^\top \beta)$. The first piece of the mapping, IC_0 , is an (inverse probability of censoring weighted) estimating function of β in the model with known censoring density, $g(\cdot | X)$, and is given by

$$IC_0(Y | G, D) \equiv \frac{D'(C, Z)(1 - \Delta)}{g(C | X)} + D(\alpha_W, Z). \quad (4)$$

D' is the derivative with respect to the first argument and α_W is the minimum of the support of $g(\cdot | X)$, which is (by CAR) allowed to be a function of the baseline covariates W . These estimating functions satisfy $E(IC_0(Y | G, D) | X) = D(X)$, under a weak identifiability condition (see Appendix) and are therefore indeed unbiased.

The second piece of the mapping is the projection of IC_0 on the tangent space of the monitoring process only assuming CAR (2). It is given by

$$IC_{nu}(Y | F, G, D) \equiv \int_0^\infty \left(\frac{D'(u, Z)\bar{F}(u | Z, \bar{L}(u))}{g(u | X)} - \frac{1}{\bar{G}(u | X)} \int_u^\infty D'(t, Z)\bar{F}(t | Z, \bar{L}(u))dt \right) dM(u), \quad (5)$$

where $F(\cdot | Z, \bar{L}(u))$ is the conditional cumulative distribution of T given $(Z, \bar{L}(u))$ and $dM(u) = I(C \in du) - \Lambda_C(du | X)I(C \geq u)$. For a given cumulative distribution F we define $\bar{F} = 1 - F$. For convenience, in $IC_{nu}(Y |$

F, G, D) we use shorthand F to represent $F(\cdot | Z, \bar{L}(u))$ for various u . We define $IC(Y | F, G, D) \equiv IC_0(Y | G, D) - IC_{nu}(Y | F, G, D)$.

If \bar{L} is time independent we denote it by W , in which case

$$IC_{nu}(Y | F, G, D) = \frac{D'(C, Z)\bar{F}(C | Z, W)}{g(C | X)} - \int_0^\infty D'(u, Z)\bar{F}(u | Z, W)du \quad (6)$$

and

$$IC(Y | F, G, D) = \frac{D'(C, Z)}{g(C | X)}(F(C | Z, W) - \Delta) + E[D(T, Z) | Z, W]. \quad (7)$$

An estimator β_n of β is asymptotically linear at the observed data distribution $P_{F_X, G}$ with influence curve $IC(Y | F_X, G)$ if $\beta_n - \beta = n^{-1} \sum_{i=1}^n IC(Y_i | F_X, G) + o_P(n^{-1/2})$. A regular estimator attains the semiparametric information bound at $P_{F_X, G}$ if its influence curve at $P_{F_X, G}$ is the so called efficient influence curve, $\ell_{\text{eff}}^*(Y | F_X, G)$, which can be defined as a standardized efficient score.

The optimal estimating function should equal this efficient score (up till a standardization matrix) when evaluated at the true parameter values. Let $D_{h_{\text{opt}}}(X | \beta) = h_{\text{opt}}(Z)K(T - Z^\top \beta)$ be the full data estimating function that is mapped into the optimal estimating function in the observed data model. Theorem 3 in the Appendix gives the explicit form of $D_{h_{\text{opt}}}$ in several data models. The most general setting has

$$\begin{aligned} IC(Y | F, G, D_{h_{\text{opt}}}) &= h_{\text{opt}}(Z)IC(Y | F, G, K_\beta) \\ &= \frac{ZE(K'_\beta | Z)}{\phi(Z)}IC(Y | F, G, K_\beta) \end{aligned} \quad (8)$$

where $\phi(Z) = E(IC(Y | F, G, K_\beta)^2 | Z)$. The efficient influence curve is

$$\ell_{\text{eff}}^*(Y | F, G, h_{\text{opt}}, c_{\text{opt}}, \beta) = c_{\text{opt}}^{-1} h_{\text{opt}}(Z)IC(Y | F, G, K_\beta), \quad (9)$$

where $c_{\text{opt}} = \left[E \left(Zh_{\text{opt}}(Z)IC_0(K'_\beta) \right) \right]$.

Given estimators F_n, G_n , and h_n of F, G , and the optimal index h_{opt} , an efficient estimate can be found by solving

$$0 = \sum_{i=1}^n IC(Y_i | F_n, G_n, D_{h_n}(\cdot | \beta)) \quad (10)$$

for β .

We must solve (10) iteratively for β . If our initial value, β_n^0 , for β is a \sqrt{n} -consistent estimator of β , then a single iteration of the Newton-Raphson algorithm for solving (10) is just the classical one-step estimator as defined in Bickel, et al. [19] (p. 395):

$$\beta_n^1 \equiv \beta_n^0 + \frac{1}{n} \sum_{i=1}^n \widehat{\ell}_{\text{eff}}^*(Y_i | \beta_n^0). \quad (11)$$

Here $\ell_{\text{eff}}^*(Y | F, G, h_{\text{opt}}, c_{\text{opt}}, \beta)$ is estimated by substitution of estimators $F_n, G_n, h_n, c_n, \beta_n^0$ for $F, G, h_{\text{opt}}, c_{\text{opt}}, \beta$, respectively.

We propose to use $IC_0(Y | G, ZK_\beta)$ as an estimating function to compute an initial estimator, β_n^0 . Then the initial estimator of the k -vector regression parameter β is the solution of the system of equations

$$\sum_{i=1}^n Z_i \left(\frac{K'_{\beta_n^0}(C_i, Z_i)(1 - \Delta_i)}{g_n(C_i | X_i)} + K_{\beta_n^0}(\alpha_{W_i}, Z_i) \right) = 0. \quad (12)$$

Formal conditions for existence and \sqrt{n} -consistency of β_n^0 are given in our technical report.

To obtain the initial estimate β_n^0 using equation (12), it is necessary to $n^{-1/4}$ -consistently estimate the conditional density of the censoring mechanism, $g(\cdot | X)$, from the data. We elected to use a time-dependent Cox proportional hazards model (3). For this model to estimate consistently the censoring density, the usual step-function estimate of the baseline hazard must be smoothed: e.g., as in Andersen, et al. [20]. If we believe the censoring mechanism is independent of all covariates we can use a kernel smoother to estimate $g(\cdot | X) \equiv g(\cdot)$. In any case, after g has been estimated, the initial estimate β_n^0 then quickly can be found by numerical methods (e.g., Newton-Raphson) using equation (12).

Estimation of the efficient influence curve (9) involves estimation of $(F, G, h_{\text{opt}}, c_{\text{opt}}, \beta)$. With an initial estimate of β and the estimate of the censoring density g in hand, we now discuss estimation of each of the other three parameters and computation of the one-step estimator β_n^1 . This general method can always be used but, as illustrated in Example 2, more specific information about the structure of the model can improve efficiency for finite samples.

F_n for time-independent case. If $\bar{L} = W$ is time-independent, then IC_{nu} is given by equation (6) and the following identity can be used to estimate

$F(\cdot | Z, W)$:

$$F(t | Z, W) = E[I(T \leq t) | Z, W] = E[\Delta | C = t, Z, W]. \quad (13)$$

The second equality follows from CAR.

The proposed submodel can be chosen to be a highly parametric model or a flexible semiparametric model. The former leads to a fully efficient estimator in fewer circumstances. Nonetheless, the finite sample performance of a parametric model is comparable if not superior to a semiparametric model because it recognizes the main effects of the covariates and is more stable where the data are sparse. This comparison is made in the second example in the simulation section.

One possible semiparametric model for $F(\cdot | Z, W)$ is a logistic generalized additive model.

$$\begin{aligned} F(t | Z, W) &= E[\Delta | C = t, Z = (Z_1, \dots, Z_k), W = (W_1, \dots, W_l)] \\ &= \frac{\exp(f_C(t) + f_{Z_1}(Z_1) + \dots + f_{W_l}(W_l))}{1 + \exp(f_C(t) + f_{Z_1}(Z_1) + \dots + f_{W_l}(W_l))}. \end{aligned} \quad (14)$$

The Splus function `gam` with `family=binomial(link=logit)` produces an F_n based on the observed data $\{Y_i\}_{i=1}^n$. Furthermore, some or all of the general functions $f_C, f_{Z_1}, \dots, f_{W_l}$, can be replaced by more parametric polynomials.

The factor $IC(Y_i | F, G, K_\beta)$ in equation (9) can now be estimated for each Y_i using the expressions for $IC_0(Y_i | G_n, K_{\beta_n^0})$ and $IC_{nu}(Y_i | F_n, G_n, K_{\beta_n^0})$ given in (4) and (6).

F_n for time-dependent case. If \bar{L} is time-dependent, IC_{nu} must be estimated directly from equation (5). It is necessary to estimate $F(t | Z, \bar{L}(u))$ for a given (t, u) with $t \geq u$. First consider the case where the density of C depends only on the time-independent covariates (even though $F(t | Z, \bar{L}(u))$ may depend on the time-dependent covariates). Then we proceed using the CAR-identity

$$F(t | Z, \bar{L}(u)) = E(\Delta | C = t, Z, \bar{L}(u), C \geq u). \quad (15)$$

To avoid the curse of dimensionality, for each u we replace $\bar{L}(u)$ by a vector of summary measures, $W_u(\bar{L}(u))$, that hopefully captures the most relevant information for predicting T . Now, for each u , we can estimate $F(\cdot | Z, \bar{L}(u)) \approx F(\cdot | Z, W_u)$ by the GAM in equation (14). The model is fit using data Y_i for which $C_i \geq u$ (i.e., individuals for which $\bar{L}(u)$ is observed).

For the general case where the censoring mechanism also depends on the time-dependent covariates, the identity (15) is not guaranteed by CAR. We proceed in estimating $F(t | Z, \bar{L}(u))$ in two stages by using the following CAR-relationship given in van der Laan and Robins [5]:

$$F(t | Z, \bar{L}(u)) = E(\xi(u, t) | Z, \bar{L}(u), C \geq u) \quad (16)$$

where

$$\xi(u, t) = \frac{\bar{G}(u | X)}{\bar{G}(t | X)} I(C \geq t) F(t | Z, \bar{L}(t), C \geq t). \quad (17)$$

We can estimate $F(\cdot | Z, \bar{L}(t), C \geq t)$ for each t from those individuals for which $C_i \geq t$ just as above (see equation (15), but now t plays the role of u). Note that from the fitted model we only need the function evaluated at the left endpoint, $F_n(t | Z, \bar{L}(t), C \geq t)$. From F_n and G_n we can calculate $\hat{\xi}_i(u, t)$ for each individual from equation (17). Now for each u , regress $\hat{\xi}_i(u, t)$ on t , Z , and $W_u(\bar{L}(u))$ using individuals for which $C_i \geq u$. The Splines function `gam` with `family=quasi(link=logit, variance=constant)` can be used to fit a logistic GAM model. This is our estimate $F_n(t | Z, \bar{L}(u))$, the final piece needed to estimate IC_{nu} in the most general case (equation (5)).

h_{opt} . The vector-valued function $h_{\text{opt}}(Z)$ is proportional to Z . The constant of proportionality is the ratio of

$$E(K'_\beta | Z) = E\left(\frac{K''(C - Z^\top \beta)(1 - \Delta)}{g(C | X)} | Z\right) + K'(\alpha_W - \beta^\top Z) \quad (18)$$

and $\phi(Z)$. Using β_n^0 and $g_n(\cdot | X)$ to obtain an observed outcome, the expression (18) can be estimated by regressing an observed outcome on Z . The function $\phi(Z)$ can be estimated in several ways depending on the number and type of covariates are available. In general, $\phi(Z)$ is the conditional expectation given Z of $IC^2(Y | \beta, F, G, K_\beta)$. An estimate of IC has already been computed and its square can be regressed on Z in some parametric or semiparametric method (e.g., splines, gam, running medians).

Although this regression method can always be used, in some cases $\phi(Z)$ has other expressions with more structure that can be exploited. In particular, if there are no covariates other than Z , then ϕ is given by equation (A.7) and can be estimated by substitution of an estimator of $F(t | Z) = E(\Delta | C = t, Z)$. If $\bar{L} = W$ is time independent, $\phi(Z)$ is given by equation (A.8), which can thus be estimated by substitution of an estimator of $F(t | Z, W) = E(\Delta | C = t, Z, W)$. Equation (A.8) will be more accurate but potentially more computationally intensive. In Example 2 the assumptions on $T | Z$ and

$W | T, Z$ imply a distribution on $W | Z$ that can be exploited in computing $\phi(Z)$.

c_{opt} . An estimate, c_n , of the normalizing matrix c_{opt} is

$$c_n = -\frac{1}{n} \sum_{i=1}^n h_n(Z_i) Z_i^\top U_{G_n}(K'_{\beta_n^0})(Y_i). \quad (19)$$

The expectation has been estimated by the empirical mean. Each factor inside the expectation has already been estimated to obtain the estimate of h_{opt} .

3 Properties of One-Step Estimator

Theorem 1 shows that if our models for $g(C|X)$ and $F(t | Z, \bar{L}(u))$ are correctly specified, then the one-step estimator β_n^1 is indeed asymptotically linear with influence curve ℓ_{eff}^* and thus asymptotically efficient. Moreover, β_n^1 has the additional feature that it remains a consistent and asymptotically normal estimator of β even when the model for $F(t | Z, \bar{L}(u))$ is misspecified (i.e., when $F_n \rightarrow F^\dagger \neq F$).

This protection from model misspecification of F follows from the general representation of ℓ_{eff}^* developed by Robins and Rotnitzky [21] and further developed in van der Laan and Robins [22]. For further details about computing this representation we refer to the Appendix and technical report.

Theorem 1 *Under regularity conditions provided in the Appendix $\beta_n^1 \equiv \beta_n^0 + n^{-1} \sum_{i=1}^n \ell_{\text{eff}}^*(Y_i | F_n, G_n, h_n, c_n, \beta_n^0)$ is asymptotically linear with influence curve*

$$\Pi \left[\ell_{\text{eff}}^*(Y | F^\dagger, G, h, c, \beta) \middle| T_2^\perp(P_{F_X, G}) \right], \quad (20)$$

where $T_2(P_{F_X, G})$ is the tangent space for the chosen CAR-model containing the true G . Furthermore, if $h = h_{\text{opt}}$ and $F^\dagger = F$, then β_n^1 is asymptotically efficient.

3.1 Construction of Confidence Intervals

A confidence region for the parameter vector β or individual confidence intervals for each regression parameter can be constructed by estimating the covariance matrix of the efficient influence function, ℓ_{eff}^* . If the model for

$F(t | Z, \bar{L}(u))$ is correctly specified, the vector $\sqrt{n}(\beta_n^1 - \beta)$ is asymptotically distributed $N(0, \text{Cov}(\ell_{\text{eff}}^*))$ because the projection operator in expression (20) is the identity operator in this case. Thus an asymptotic 95% confidence region for β is $\{\beta \in \mathbb{R}^k \mid (\beta_n^1 - \beta)^\top \widehat{\Sigma}^{-1}(\beta_n^1 - \beta) \leq (k/n)F_{0.95, k, \infty}\}$ (e.g., Morrison [23]), where $\widehat{\Sigma}$ is the empirical variance of the estimated efficient influence function,

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \left(\widehat{\ell}_{\text{eff}}^*(Y_i) - \frac{1}{n} \sum_{i'=1}^n \widehat{\ell}_{\text{eff}}^*(Y_{i'}) \right) \left(\widehat{\ell}_{\text{eff}}^*(Y_i) - \frac{1}{n} \sum_{i'=1}^n \widehat{\ell}_{\text{eff}}^*(Y_{i'}) \right)^\top,$$

where we define $\widehat{\ell}_{\text{eff}}^*(Y) \equiv \ell_{\text{eff}}^*(Y | F_n, G_n, h_n, c_n, \beta_n^0)$. An asymptotic 95% confidence interval for a single parameter is $\beta_n^1 \pm 1.96\widehat{\sigma}/\sqrt{n}$, where $\widehat{\sigma}^2$ is the appropriate diagonal element of $\widehat{\Sigma}$.

If the model for $F(t | Z, \bar{L}(u))$ is misspecified, the above confidence intervals are conservative. The true variance of the estimator is given by the variance of expression (20), which is smaller than the variance of ℓ_{eff}^* . We refer to van der Laan and Robins [5,22] for exact expressions when the model for the monitoring process is either Cox proportional hazards or independence. However, unless F is very poorly specified, the conservative intervals will be fairly accurate.

3.2 A Doubly Robust Estimator

Given estimates F_n, G_n of the nuisance parameters F, G and a choice h_n for the full data estimating function, consider the estimator β_n solving

$$0 = \frac{1}{n} \sum_{i=1}^n \ell_{\text{eff}}^*(Y_i | F_n, G_n, h_n, c_n, \beta).$$

Due to the orthogonality of $\ell_{\text{eff}}^*(Y | F, G, h, c, \beta)$ to any nuisance score generated by fluctuations of G , we actually have the following double-robustness property with respect to the nuisance parameters F and G of this estimating function $\ell_{\text{eff}}^*(Y | F, G, h, c, \beta)$ if G is dominated by G_1 :

$$E \left(\ell_{\text{eff}}^*(Y | F_1, G_1, h, c, \beta) \right) = 0 \text{ if either } F_1 = F \text{ or } G_1 = G.$$

This means that, in fact, under regularity conditions, β_n will be consistent and asymptotically linear if either the model for $F(t | Z, \bar{L}(u))$ is correctly specified or the model for G is correctly specified. This double robustness property of β_n implies that, in practice, a minor misspecification of the model for G can

be corrected by doing a good job in modelling $F(t | Z, \bar{L}(u))$ and vice versa. For details on double robustness, see van der Laan and Robins [22].

3.3 Data Adaptive Selection of Location Parameter

The regression parameter β represents the effect of Z on the location parameter identified by K . Thus the choice of location parameter affects immediately the interpretation of β and could therefore just be subject-matter driven. However, one might also decide to choose the location parameter that is best identifiable from the data. Suppose that we choose a location parameter K_τ with compact support $[-\tau, \tau]$ that, e.g., approximates the median for $\tau \rightarrow 0$ and approximates the mean for $\tau \rightarrow \infty$. In that case, we propose to calculate $\hat{\Sigma}$ for a range of τ 's and select the τ that minimizes this estimated variance of the efficient influence curve. This corresponds with choosing the location parameter that results in the smallest confidence bands.

4 Simulations

Two simulation studies are presented to illustrate the applicability and efficiency of these methods. Example 1 demonstrates that the asymptotic properties of the one-step estimator apply to a dataset of moderate size. The superiority of the one-step estimator over the initial estimator is also shown. The effects of an additional time-independent covariate, W , and the submodel selected for $F(\cdot | Z, W)$ are considered in Example 2.

The function K we use in these simulations is a smoothed truncated mean given by

$$K(t) = -\tau I(t < -\tau) + \tau I(t > \tau) + (t + (\tau/\pi) \sin(\pi t/\tau)) I(|t| \leq \tau) \quad (21)$$

with $\tau = 3$. K has two continuous derivatives, both of which are zero outside the interval $(-\tau, \tau)$.

4.1 Example 1: No Unmodelled Covariates

The data generating distribution has $\beta = (\beta_0, \beta_1) = (0, 1)$, $Z_0 \equiv 1$ (intercept), $Z_1 \sim N(0, 1)$, $T | Z \sim N(Z_1, 1)$, and $C | Z \sim N(Z_1, 1)$. The observed data is (C, Δ, Z) . The general method of estimation described in section 2 was used with the following specifics. The censoring distribution was estimated

via linear regression of C on Z with independent normal error. The distribution of $T \mid Z$ was estimated using equation (13) and a generalized linear model with probit link. h_{opt} was computed after approximating the integrals $E(K'_{\beta_n}(T, Z) \mid Z) = -\int K''_{\beta_n}(t, Z)F(t \mid Z)dt$ and the expression (A.7) for $\phi(Z)$ by Simpson's Rule with 20 intervals. The results in Table 1 are based on 1000 repetitions.

The one-step estimator is efficient in this example because the submodel chosen for F is correct. In finite samples we estimate the efficiency by comparing the variance of the estimator with the variance of the efficient influence curve. Similarly we estimate the efficiency of the one-step estimator relative to the initial-estimator. Results for the parameter β_1 at three sample sizes are given in Table 1 (similar patterns are seen for β_0 , not shown). The asymptotic efficiency is evident in both of the larger samples.

Table 1

Comparison of initial and one-step estimators for simple linear regression example. One-step estimator is asymptotically efficient and appears to be fully efficient even for moderate sample sizes.

Estimator	Sample Size	Asymptotic Relative Efficiency	Relative Efficiency (baseline=Initial)
Initial ($\beta_{n,1}^0$)	250	0.53	1
	500	0.76	1
	1000	0.81	1
One-step ($\beta_{n,1}^1$)	250	0.67	1.2
	500	1.05	1.3
	1000	1.05	1.3

4.2 Example 2: Unmodelled Covariate

Suppose in addition to Z , another covariate W has been collected that is associated with T . Our method uses the information contained in the covariate to improve the estimate of β . The strength of the relationship between T and W is one factor that determines how much our one-step estimator can improve the initial estimator, which does not use W . In this example we consider three covariates: $W_1 = T$, $W_2 = T + \text{small error}$, and $W_3 = T + \text{large error}$. The first corresponds to a perfect surrogate for T , the second to a good predictor of T , and the third to a poor predictor of T .

The degree to which we will be able to exploit the information in W also

depends on the submodel we select for $F(\cdot | Z, W)$. It is frequently wise to be optimistic and select a small submodel; for example, a generalized linear model often outperforms a generalized additive model if linearity is at all reasonable. In this example we consider two one-step estimators. The first estimator is the generic method described in section 2. The assumed model for $F(\cdot | Z, W)$ is correct for each of the three covariates. Thus β_n^1 is asymptotically efficient in each case.

The second one-step estimator assumes W is a perfect surrogate for T and thus “estimates” $F(t | W, Z)$ with $I(W \leq t)$. This is correct in the first scenario because $W_1 = T$ but not correct for the second or third case. Under the assumption $W = T$, one could directly estimate β by linear regression: $W = Z^\top \beta + \epsilon$. This direct linear regression method is optimal in case 1 where $W_1 = T$. However, in the other two cases, this estimator is inconsistent. Our one-step estimator is consistent in each of the three cases and is asymptotically equivalent with this direct linear regression method if $W = T$.

The simulation results are presented in Table 2. The initial estimator is exactly the estimator in the previous example. It does not use the information provided by the covariate W and thus is not nearly efficient. If W is very informative, as in the first two cases, the variance bound is less than half the variance of the initial estimator.

The generic one-step estimator is efficient, but for samples with $N = 1000$ the variance bound is about 10% smaller than the variance of the estimator. The special one-step estimator that assumes $W = T$ reaches the efficiency bound (and then some) when W is very informative. When W is a poor predictor of T , the performance of this estimator suffers as should be expected because the assumption $W = T$ is bad. The variance of the special estimator is larger than the generic estimator in this case.

Details. The data generating distribution has $Z_0 \equiv 1$ (intercept), $Z_1 \sim N(0, 1)$, $T | Z \sim N(Z_1, 1)$, $W_1 | T, Z = T$, $W_2 | T, Z \sim N(T, 0.1^2)$, $W_3 | T, Z \sim N(T, 1.0^2)$, and $C | Z, W \sim N(Z_1, 1)$. The general method of estimation described in section 2 was used with the following specifics to compute the generic one-step estimator. The censoring distribution was estimated via linear regression of C on Z with independent normal error. The distribution of $T | Z, W$ was estimated using equation (13) and a generalized linear model with probit link. From the data model it can be shown that $f_{W|Z}$ can be estimated consistently with linear regression with normal errors in each of the three cases. This estimate can be used to more accurately estimate

$$\phi(Z) = \int f_{W|Z}(w | Z) \left[E(K_{\beta_n^0} | Z, W)^2 + \right.$$

Table 2

Comparison of (the variances of) the initial estimator and two one-step estimators. The generic one-step estimator is efficient in each case. The special one-step estimator assumes $W = T$ and is therefore efficient only in case 1. The generic one-step estimator has not reached the (asymptotic) efficiency bound in this simulation ($N = 1000$) but the special one-step estimator has in the first two cases where W is a perfect or good predictor of T .

Estimator	Available Covariate	Asymptotic Relative Efficiency	Relative Efficiency (baseline=Initial)	Relative Efficiency (baseline=Generic)
Initial ($\beta_{n,1}^0$)	$W_1 = T$	0.40	1	
	W_2	0.44	1	
	W_3	0.78	1	
Generic ($\beta_{n,1}^1$)	$W_1 = T$	0.90	2.20	1
	W_2	0.93	2.12	1
	W_3	0.94	1.19	1
Special ($\beta_{n,1}^{1*}$)	$W_1 = T$	1.03	2.32	1.15
	W_2	1.08	2.48	1.08
	W_3	0.90	1.15	0.96

$$\int \frac{F(t | Z, w) \bar{F}(t | Z, w) K'_{\beta_n^0}(t, Z)}{g(t | Z, w)} dt \Big] dw$$

(see equation (A.8)).

The special one-step estimator based on the assumption $W = T$ is easier to compute because the assumption implies $F(t | Z, W) = I(t \leq W)$, $E(K'_{\beta_n^0} | Z, W) = K'_{\beta_n^0}(W, Z)$, and

$$\phi(Z) = \int f_{W|Z}(w | Z) K_{\beta_n^0}(w, Z)^2 dw.$$

Results in Table 2 are based on 1000 repetitions.

5 California Partners' Study

The methods described in this paper were applied to a dataset extracted from the California Partners' Study. Each case consists of a monogamous

heterosexual couple in which the male is HIV-positive due to a prior sexual contact. The “failure time variable” on which current status data is available is the time (in months) until infection of the female partner. Several time-independent covariates are available including an indicator of condom use (never=1, ever=0), an indicator of bleeding (ever=1, never=0), an indicator of a sexually transmitted disease (STD) history in the female (ever=1, never=0), an estimate of the rate of sexual contact (contacts per month), and the age of the female (years). There are 87 subjects with complete information on these five covariates. More detailed descriptions of the data are available in Padian, et al. [24], Shiboski and Jewell [25], Jewell and Shiboski [7], and Padian, et al. [26].

Our ultimate goal is to estimate the regression parameters in the model $T = Z^\top \beta + \epsilon$, where T is the log of the transmission time. Define the following notation: $Z_0 \equiv 1$ is the intercept, $Z_1 = I(\text{No condom use})$, $Z_2 = I(\text{STD History})$, $Z_3 = Z_1 Z_2$. We expect the coefficients of Z_1 and Z_2 to be negative, indicating these risk factors lower the expected time until transmission of the disease. We include the interaction term because the effect of STD history may not be observed if condoms are used.

Before estimating β , we must model the censoring mechanism. The distribution of C may be dependent on the covariates in the model and possibly other external to the regression model. Several classes of models for the conditional distribution of C given covariates are feasible including simple linear regression and Cox proportional hazards. In each of these classes the only significant dependence is between the monitoring time and Z_1 . As noted in the introduction, it may be safer to include more rather than fewer covariates and to specify a semi-parametric rather than parametric model to protect against dependence between T and C as much as possible. With that in mind we chose to use the Cox proportional hazards model and to include all five covariates mentioned in the paragraph describing the dataset.

With a model for the censoring mechanism in hand, we proceed to computing an initial estimate of β based on equation (12). The length of the support window of K' can be varied (as can the functional form of K) to obtain results for a range of estimators from smoothed median regression to trimmed mean regression. Table 3 displays how the initial and one-step estimates depend on the selection of the window length. For the analysis of log transmission time the estimates do not change substantially with τ . In a similar analysis of the untransformed transmission time, the estimates changed due to the right skewness of the distribution. For example, the intercept, which represents the time until infection in pairs with neither risk factor, was largest for large τ and smallest for small τ . A wide window indicates the tail of the distribution will have an effect while a small window indicates only the center of the data is measured.

Table 3

Dependence of Estimates on Window Length. K' is zero outside $Z^\top \beta \pm \tau$. If τ is larger than 0.3, this window extends beyond the support of g_n . If τ is smaller than 0.15, the initial estimator has numerous solutions.

τ	Parameter	Z_0	Z_1	Z_2	Z_3
0.17	β_n^0	4.43	-0.52	-0.27	0.15
	β_n^1	4.42	-0.49	-0.26	0.26
0.21	β_n^0	4.43	-0.53	-0.29	0.17
	β_n^1	4.43	-0.50	-0.26	0.30
0.25	β_n^0	4.44	-0.54	-0.31	0.20
	β_n^1	4.43	-0.50	-0.27	0.42
0.29	β_n^0	4.45	-0.56	-0.33	0.24
	β_n^1	4.44	-0.51	-0.28	0.45

For $\tau = 0.25$ the initial estimator is $\beta_n^0 = (4.44, -0.54, -0.31, 0.24)$; that is, the conditional log time until infection is centered at $4.44 - 0.54Z_1 - 0.31Z_2 + 0.24Z_3$.

The remaining item is to compute the one-step estimator. The covariates in this data set are time-independent so equation (6) applies. The cumulative distribution function $F(t | Z, W)$ was estimated using the generalized additive model as in equation (14) with $Z = (Z_0, Z_1, Z_2, Z_3)$ and logit link function. The indicator of bleeding and the age of the female were used as covariates outside the regression model (that is, W). Adjusting for these covariates is not possible with any other technique in the literature. The one-step estimator is $\beta_n^1 = (4.43, -0.50, -0.27, 0.42)$; that is, the conditional log time until infection is centered at $4.43 - 0.50Z_1 - 0.27Z_2 + 0.42Z_3$.

The individual standard errors of the coefficients of the two main effect indicator variables are 0.19 and 0.11, respectively. Thus the indicators of no condom use and of STD history are significant ($0.01 < p < 0.02$) factors in predicting the log time until transmission. The coefficient of the interaction is not statistically significant.

6 Discussion

We provide locally efficient estimators of regression coefficients based on current status data with time-dependent covariates with a general linear regression failure-time model, $T = Z^\top \beta + \epsilon$, where the distribution of the error term has conditional location parameter equal to zero. Although the curse of

dimensionality prevents a globally efficient estimator, the proposed estimator attains the efficiency bound at a user-supplied submodel of interest and is consistent and asymptotically normal over the whole model.

Another advantage of this locally efficient estimation approach is that the censoring process need not be independent of the failure time; only coarsening at random is required. Unlike other regression estimation approaches, this estimator allows the effects of other unmodelled covariates to be incorporated in a very general way. Thus if a surrogate covariate for T is available, it may be used to improve the estimation of the regression parameters even though the surrogate is not included in the model. Furthermore, the unmodelled covariates may even be time-dependent processes.

The estimator exists in closed form and has been implemented with generally available software. It was shown in simulations to perform according to its asymptotic theoretical properties in finite samples and was applied to data from the California Partners' Study.

A Appendix

Identifiability of β . The crucial and structural condition for the consistency of the solution of an inverse probability of censoring weighted estimating equation is that the estimating function is unbiased. The following theorem provides the necessary and minimal conditions guaranteeing that the inverse weighting of the full data estimating function works.

Theorem 2 *Assume that (i) K is constant outside $[-\tau, \tau]$ and strictly increasing with two continuous derivatives on $[-\tau, \tau]$, (ii) $\Pr(-\tau < \epsilon < \tau \mid Z) > \delta_1 > 0$ with probability one for some δ_1 , (iii) the support of $g(\cdot \mid X)$ is an open interval (α_W, α^Y) with $W = (Z, L(0))$ being the baseline covariates, and (iv) $\Pr(Z \in \mathcal{Z}(\beta, G)) > \delta_2 > 0$ for some δ_2 , where*

$$\mathcal{Z}(\beta, G) \equiv \{z : \alpha^Y - \beta z > \min(T - \beta z, \tau), \alpha_W - \beta z < \max(T - \beta z, -\tau)\}. \quad (\text{A.1})$$

where the inequalities need to hold $F_{X|Z=z}$ a.e. Assume that $\mathcal{Z}(\beta, G)$ is non-empty. Define

$$\mathcal{H}^F(\beta, G) = \{h(Z)I(Z \in \mathcal{Z}(\beta, G)) : \sup_z |h(z)| < \infty\}.$$

If $h \in \mathcal{H}^F(\beta, G)$, then $E(IC_0(Y \mid G, D_h(\cdot \mid \beta)) \mid X) = D_h(X \mid \beta)$.

Proof of theorem. The conditional expectation is given by:

$$I(Z \in \mathcal{Z})h(Z) \int_{\alpha_W}^{\max(\min(T, \alpha^Y), \alpha_W)} K'_\beta(c, Z)dc + K_\beta(\alpha_W, Z).$$

This can be rewritten as:

$$\begin{aligned} & I(Z \in \mathcal{Z})h(Z)K_\beta(\max(\min(T, \alpha^Y), \alpha_W), Z) \\ &= I(Z \in \mathcal{Z})h(Z)K(\max(\min(\epsilon, \alpha^Y - \beta Z), \alpha_W - \beta Z)) \\ &= I(Z \in \mathcal{Z})h(Z)K(\max(\min(\epsilon, \alpha^Y - \beta Z), -\tau)) \text{ by (iv) and (A.1)} \\ &= I(Z \in \mathcal{Z})h(Z)K(\min(\epsilon, \alpha^Y - \beta Z)) \text{ by (i)} \\ &= I(Z \in \mathcal{Z})h(Z)K(\min(\epsilon, \tau)) \text{ by (iv) and (A.1)} \\ &= I(Z \in \mathcal{Z})h(Z)K(\epsilon) \text{ by (i).} \square \end{aligned}$$

Asymptotic Efficiency of β_n^1 . Theorem 1 (restated here with the regularity conditions) shows that the one-step estimator β_n^1 is indeed asymptotically linear and consistent regardless of the model for F , and if F is correctly specified, estimator has influence curve ℓ_{eff}^* .

Before stating the regularity conditions, we note that condition (ii) in the theorem is a general empirical process condition. For empirical process theory we refer to van der Vaart and Wellner [27]. We decided not to derive more primitive conditions that imply condition (ii) because it is technical and model dependent. In our technical report it is shown that condition (i) assures the initial estimator exists and is \sqrt{n} -consistent and that the structural condition (A.4) as needed in the proof holds. Condition (iii) requires that g_n converges uniformly to g over a set A and that $F_n(t \mid Z, \bar{L}(u))$ converges uniformly to something (not necessarily the truth) over a set B , where A and B are intersections of the support of g and K : in other words, one only needs convergence over sets at which F and g are identifiable (under condition (i)). In addition, condition (iii) requires that the product of the rates is $o_P(n^{-1/2})$. Condition (iv) requires that one uses an efficient procedure for estimation of the monitoring mechanism $g(c \mid X)$ such as a maximum likelihood estimator.

Theorem 1 *Assume*

- (i) *Conditions of lemma 1 hold.*
- (ii) *$\ell_{\text{eff}}^*(\cdot \mid F_n, G_n, h_n, c_n, \beta_n^0)$ is contained in a $P_{F_X, G}$ -Donsker class with probability tending to one.*
- (iii) *For some F^\dagger we have that $\sqrt{nr_1nr_2n} \rightarrow 0$ where*

$$r_{1n} \equiv \sup_A |g_n(u | X) - g(u | X)| \rightarrow 0$$

$$r_{2n} \equiv \sup_B \left| F_n(t | Z, \bar{L}(u)) - F^\dagger(t | Z, \bar{L}(u)) \right| \rightarrow 0$$

and the uniform convergence statements need to hold in probability over the sets

$$A \equiv \{(u, Z, \bar{L}(u)) : K'(u - Z^\top \beta) > 0, \max(\bar{F}^\dagger, \bar{F}_n, \bar{F})(u | Z, \bar{L}(u)) > 0\}$$

$$B \equiv \{(t, Z, \bar{L}(u)) : t \geq u, K'(t - Z^\top \beta) > 0, \max(g_n, g)(u | Z, \bar{L}(u)) > 0\}.$$

(iv) $\Phi(G_n) \equiv E_Y(\ell_{\text{eff}}^*(Y | F^\dagger, G_n, h, c, \beta))$ is an efficient estimator of $\Phi(G)$.

Then $\beta_n^1 \equiv \beta_n^0 + n^{-1} \sum_{i=1}^n \ell_{\text{eff}}^*(Y_i | F_n, G_n, h_n, c_n, \beta_n^0)$ is asymptotically linear with influence curve

$$\Pi \left[\ell_{\text{eff}}^*(Y | F^\dagger, G, h, c, \beta) \middle| T_2^\perp(P_{F_X, G}) \right], \quad (\text{A.2})$$

where $T_2(P_{F_X, G})$ is the tangent space for the chosen CAR-model containing the true G and the matrix c is the limit of c_n . Furthermore, if $h = h_{\text{opt}}$ and $F^\dagger = F$ (so that $\ell_{\text{eff}}^*(\cdot | F^\dagger, G, h, c, \beta) = \ell_{\text{eff}}^*(\cdot | F, G, h_{\text{opt}}, c_{\text{opt}}, \beta)$), then β_n^1 is asymptotically efficient.

See our technical report for proof.

In the next lemma we assume that for F_X a.e. (Z, L)

$$\inf_{c \in A(Z, L)} g(c | X) > \gamma > 0 \text{ for some } \gamma > 0, \quad (\text{A.3})$$

where

$$A(Z, L) \equiv \{c : c \in (\alpha_X, \alpha^X), K'(c - Z^\top \beta) > 0, \bar{F}(c | Z, \bar{L}(c)) > 0\}.$$

This condition implies the identifiability result of Theorem 2 and could be weakened by choosing as choice for the initial estimating function an $h(Z)$ which equals zero for $Z \notin \mathcal{Z}(\beta, G)$, instead of $h(Z) = Z$.

Lemma 1 Let h_n be given, β_n^0 be the initial estimator defined in section 2, and $c_n \equiv c(h_n)$ be as in equation (19).

Suppose that the following conditions on the true data generating distribution hold: $(\partial/\partial\beta)E(ZU_G(K_\beta)(Y))$ is invertible at the true value of $\beta = \beta_0$, g' and K'' are bounded above, identifiability condition (A.3) holds for $\beta \in N_{\beta_0}$, where N_{β_0} is an arbitrarily small neighborhood of β_0 .

In addition, we make the following consistency assumptions: $\|g_n - g\|_{\infty, A} = O_P(n^{-1/4})$, $\|G_n - G\|_{\infty, A} = O_P(n^{-1/2})$, $h_n(Z)$ converges uniformly to an arbitrary $h(Z)$ with $c(h) \equiv -E(h(Z)Z^\top U_G(K'_\beta)(Y))$ invertible.

Then c_n converges to $c(h)$ and (under no conditions on F_n)

$$E_Y \left(\ell_{\text{eff}}^*(\cdot \mid F_n, G, h_n, c_n, \beta_n^0) \right) = \beta - \beta_n^0 + o_P(n^{-1/2}) \quad (\text{A.4})$$

See our technical report for proof.

Optimal Estimating Function, $D_{h_{\text{opt}}}$. The following theorem provides the closed-form representation of the optimal full data estimating function $D_{h_{\text{opt}}}(X)$, which is such that $IC(Y \mid F_X, G, D_{h_{\text{opt}}})$ equals the efficient influence curve for β in our observed model.

Theorem 3 *Suppose K is twice differentiable and assume condition (A.3). Then*

$$D_{h_{\text{opt}}}(X) \equiv h_{\text{opt}}(Z)K_\beta(T, Z) \equiv \frac{ZE(K'_\beta \mid Z)}{\phi(Z)}K_\beta(T, Z) \quad (\text{A.5})$$

where, for general \bar{L} ,

$$\phi(Z) = E(IC(Y \mid F, G, K_\beta)^2 \mid Z). \quad (\text{A.6})$$

Consider the case where $(Z, \bar{L}) = Z$ (i.e., no covariates other than those modelled). Then

$$\phi(Z) = \int \frac{F(t \mid Z)\bar{F}(t \mid Z)K'_\beta{}^2(t, Z)}{g(t \mid X)} dt. \quad (\text{A.7})$$

Consider the case where $(Z, \bar{L}) = (Z, W)$ with W time independent but not modelled. Then

$$\phi(Z) = E \left(E(K_\beta \mid Z, W)^2 + \int \frac{F(t \mid Z, W)\bar{F}(t \mid Z, W)K'_\beta{}^2(t, Z)}{g(t \mid X)} dt \mid Z \right). \quad (\text{A.8})$$

References

- [1] J.M. Robins, Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, Proceedings of

- the Biopharmaceutical Section, American Statistical Association, (1993) 22–33.
- [2] D.F. Heitjan, D.B. Rubin, Ignorability and coarse data, *Annals of Statistics*, 19 (1991) 2244–2253.
 - [3] M. Jacobsen, N. Keiding, Coarsening at random in general sample spaces and random censoring in continuous time, *Annals of Statistics*, 23 (1995) 774–786.
 - [4] R.D. Gill, M.J. van der Laan, J.M. Robins, Coarsening at random: Characterizations, conjectures and counter-examples, in: D.Y. Lin, T.R. Fleming (Eds.), *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, Springer (1997) 255–295.
 - [5] M.J. van der Laan, J.M. Robins, Locally efficient estimation with current status data and time-dependent covariates, *Journal of the American Statistical Association*, 93 (1998) 693–701.
 - [6] I.D. Diamond, J.W. McDonald, I.H. Shah, Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan, *Demography*, 23 (1986) 607–620.
 - [7] N.P. Jewell, S.C. Shiboski, Statistical analysis of HIV infectivity based on partner studies, *Biometrics*, 46 (1990) 1133–1150.
 - [8] I.D. Diamond, J.W. McDonald, The analysis of current status data, in: J. Trussell, R. Hankinson, J. Tilton (Eds.), *Demographic Applications of Event History Analysis*, Oxford University Press (1991).
 - [9] N. Keiding, Age-specific incidence and prevalence (with discussion), *Journal of the Royal Statistical Society Ser. A*, 154 (1991) 371–412.
 - [10] J. Sun, D. Kalbfleisch, The analysis of current status data on point processes, *Journal of the American Statistical Association*, 88 (1993) 1449–1454.
 - [11] P. Groeneboom, J.A. Wellner, *Information bounds and nonparametric maximum likelihood estimation*, Birkhäuser Verlag, 1992.
 - [12] N.P. Jewell, H.M. Malani, E. Vittinghoff, Nonparametric estimation for a form of doubly censored data with application to two problems in AIDS, *Journal of the American Statistical Association*, 89, (1994) 7–18.
 - [13] S. van de Geer, Asymptotic normality in mixture models, preprint University of Leiden, the Netherlands, 1994.
 - [14] J. Huang, J.A. Wellner, Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I, *Statistica Neerlandica* 49 (1995) 153–163.
 - [15] D. Rabinowitz, A. Tsiatis, A. Aragon, Regression with interval-censored data, *Biometrika*, 82 (1995) 501–513.
 - [16] J. Huang, Efficient estimation for the proportional hazards model with interval censoring, *Annals of Statistics*, 24 (1996) 540–568.

- [17] A.J. Rossini, A.A. Tsiatis, A semiparametric proportional odds regression model for the analysis of current status data, *Journal of the American Statistical Association*, 91 (1996) 713–721.
- [18] X. Shen, Linear regression with current status data, *Journal of the American Statistical Association*, 95 (2000) 842–852.
- [19] P.J. Bickel, A.J. Klaassen, Y. Ritov, J.A. Wellner, *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore, 1993.
- [20] P.K. Andersen, O. Borgan, R.D. Gill, N. Keiding, *Statistical models based on counting processes*, Springer, New York, 1993.
- [21] J.M. Robins, A. Rotnitzky, *Recovery of information and adjustment for dependent censoring using surrogate markers*, *AIDS Epidemiology, Methodological Issues*, Birkhäuser, 1992.
- [22] M.J. van der Laan, J.M. Robins, *Unified methods for censored longitudinal data and causality*, Springer, New York, 2002.
- [23] D. Morrison, *Multivariate statistical methods*, 3rd ed., McGraw-Hill Publishing Co., San Francisco, 1990.
- [24] N. Padian, L. Marquis, D.P. Francis, R.E. Anderson, G.W. Rutherford, P.M. O’Malley, W. Winkelstein, Male-to-female transmission of Human Immunodeficiency Virus, *Journal of the American Medical Association*, 258 (1987) 788–790.
- [25] S.C. Shiboski, N.P. Jewell, Statistical analysis of the time dependence of HIV infectivity based on partner study data, *Journal of the American Statistical Association*, 87 (1992) 360–372.
- [26] N. Padian, S.C. Shiboski, S.O. Glass, E. Vittinghoff, Heterosexual transmission of HIV in northern California: results from a ten year study, *American Journal of Epidemiology*, 146 (1997) 350–357.
- [27] A.W. van der Vaart, J.A. Wellner, *Weak convergence and empirical processes*, Springer Verlag, 1996.