Introduction to Empirical Processes and Semiparametric Inference
Lecture 27: More Semiparametrics

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

The Profile Sampler

Lee, Kosorok and Fine (2005) proposed inference based on sampling from a posterior distribution based on the profile likelihood.

The quadratic expansion of the previous section can generate confidence sets for θ by inverting the log-likelihood ratio.

Translating this elegant theory into practice can be computationally challenging.

In principle, having an estimator of θ and its variance simplifies this issue considerably.

However, the computation of these quantities using the semiparametric likelihood poses stiff challenges relative to those encountered with parametric models, as has been illustrated in several places in this book.

Finding the maximizer of the profile likelihood is done implicitly and typically involves numerical approximations.

When the nuisance parameter is not $\sqrt{n}$ estimable, nonparametric functional estimation of η for fixed θ may be required, which depends heavily on the proper choice of smoothing parameters.

Even when η is estimable at the parametric rate, and without smoothing, $\tilde{I}_0$ does not ordinarily have a closed form.

When it does have a closed form, it may include linear operators which are difficult to estimate well, and inverting the estimated linear operators may not be straightforward.

The validity of such variance estimators must be established on a case-by-case basis.

The bootstrap is a possible solution to some of these problems.

Theoretical justification for the bootstrap is possible but quite challenging for semiparametric models where the nuisance parameter is not $\sqrt{n}$ consistent.

Even when the bootstrap can be shown to be valid, the computational burden is quite substantial, since maximization over both θ and η is needed for each bootstrap sample.

A different approach to variance estimation is possible via Corollary 19.4 (presented previously) which verifies that the curvature of the profile likelihood near $\hat{\theta}_n$ is asymptotically equal to $\tilde{I}_0$.

In practice, one can perform second order numerical differentiation by

- evaluating the profile likelihood on a hyperrectangular grid of 3^k equidistant points centered at $\hat{\theta}_n$,
- taking the appropriate differences,
- and then dividing by $4h^2$,
- where p is the dimension of θ
- and h is the spacing between grid points.

While the properties of h for the asymptotic validity of this approach are well known, there are no clear cut rules on choosing the grid spacing in a given data set.

Thus, it would seem difficult to automate this technique for practical usage.

As an alternative, Lee, Kosorok and Fine propose an application of Markov chain Monte Carlo to the semiparametric profile likelihood.

The method involves generating a Markov chain $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ with stationary density proportional to

$$p_{\theta,n}(\theta) \equiv \exp\left(pL_n(\theta)\right) q(\theta),$$

where $q(\theta) = Q(d\theta)/(d\theta)$ for some prior measure Q.

This can be accomplished by using, for example, the Metropolis-Hastings algorithm (Metropolis, et al., 1953; and Hastings, 1970).

Here are the steps:

- Begin with an initial value $\theta^{(1)}$ for the chain.
- For each $k = 2, 3, \ldots,$ obtain a proposal $\tilde{\theta}^{k+1}$ by random walk from $\theta^{(k)}$.
- Compute $p_{\tilde{\theta}^{k+1},n}(\tilde{\theta}^{k+1})$, and decide whether to accept $\tilde{\theta}^{k+1}$ by evaluating the ratio

$$\frac{p_{\tilde{\theta}^{k+1},n}(\tilde{\theta}^{k+1})}{p_{\theta^{(k)},n}(\theta^{(k)})}$$

  and applying an acceptance rule.

After generating a sufficiently long chain, one may compute the mean of the chain to estimate the maximizer of $pL_n(\theta)$ and the variance of the chain to estimate $\tilde{I}_0^{-1}$.

The output from the Markov chain can also be directly used to construct various confidence sets, including minimum volume confidence rectangles.

Whether or not a Markov chain is used to sample from the "posterior" proportional to

$$\exp\left(pL_n(\theta)\right) q(\theta),$$

the procedure based on sampling from this posterior is referred to as the *profile sampler*.

Part of the computational simplicity of this procedure is that $pL_n(\theta)$ does not need to be maximized, it only needs to be evaluated.

The profile likelihood is generally fairly easy to compute as a consequence of algorithms such as the stationary point algorithm for maximizing over the nuisance parameter.

On the other hand, sometimes the profile likelihood can be very hard to compute.

When this is the case, numerical differentiation via Corollary 19.4 may be advantageous since it requires fewer evaluations of the profile likelihood.

However, numerical evidence in Section 4.2 of Lee, Kosorok and Fine (2005) seems to indicate that, at least for moderately small samples, numerical differentiation does not perform as well in general as the profile sampler.

This observation is supported by theoretical work on the profile sampler by Cheng and Kosorok (2007a, 2007b) who show that the profile sampler yields frequentist inference that is second-order accurate.

Thus the profile sampler may be beneficial even when the profile likelihood is hard to compute.

The procedure's validity is established in Theorem 1 below which extends Theorem 19.5 to allow the quadratic expansion of the log-likelihood around $\hat{\theta}_n$ to be valid in a fixed, bounded set, rather than only in a shrinking neighborhood.

The conclusion of these arguments is that the "posterior" distribution of the profile likelihood with respect to a prior on θ is asymptotically equivalent to the distribution of $\hat{\theta}_n$.

In order to do this, the new theorem will require an additional assumption on the profile likelihood.

Define

$$\Delta_n(\theta) \equiv n^{-1}(pL_n(\theta) - pL_n(\hat{\theta}_n)).$$

Here is the theorem:

THEOREM 1. *Assume Θ is compact, $\tilde{I}_0$ is positive definite, $Q(\Theta) < \infty$, $q(\theta_0) > 0$, and q is continuous at θ_0.*

Assume also that $\hat{\theta}_n$ is efficient and that (19.10) holds for $\check{\theta}_n = \hat{\theta}_n$ and for any possibly random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$.

Assume moreover that for every random sequence $\tilde{\theta}_n \in \Theta$

$$\Delta_n(\tilde{\theta}_n) = o_{P_0}(1) \quad \textit{implies that} \quad \tilde{\theta}_n = \theta_0 + o_{P_0}(1). \tag{1}$$

Then, for every measurable function $g : \mathbb{R}^k \mapsto \mathbb{R}$ *satisfying*

$$\limsup_{k\to\infty} k^{-2} \log \left(\sup_{u\in\mathbb{R}^k : \|u\| \le k} |g(u)| \right) \quad \le \quad 0, \tag{2}$$

we have

$$\begin{aligned} &\frac{\int_\Theta g\left(\sqrt{n}(\theta - \hat{\theta}_n)\right) p_{\theta,n}(\theta) d\theta}{\int_\Theta p_{\theta,n} d\theta} \\ = \quad &\int_{\mathbb{R}^k} g(u)(2\pi)^{-k/2} |\tilde{I}_0|^{1/2} \exp\left[-\frac{u'\tilde{I}_0 u}{2}\right] du + o_{P_0}(1). \end{aligned} \tag{3}$$

Note that when $g(u) = O(1 + \|u\|)^d$, for any $d < \infty$, Condition (2) is readily satisfied.

This means that

- indicators of measurable sets
- and the first two moments of $\sqrt{n}(T - \hat{\theta}_n)$, where T has the posterior density proportional to $t \mapsto p_{t,n}(t)$,

are consistent for the corresponding probabilities and moments of the limiting Gaussian distribution.

Specifically,

$$\mathsf{E}(T) = \hat{\theta}_n + o_{P_0}(n^{-1/2})$$

and

$$n\mathsf{var}(T) = \tilde{I}_0^{-1} + o_{P_0}(1).$$

Thus we can calculate all the quantities needed for inference on θ without having to actually maximize the profile likelihood directly or compute derivatives.

Note that the interesting Condition (1) is not implied by the other conditions and is not implied by the identifiability of the Kulback-Leibler information from the full likelihood.

Nevertheless, if it can be shown that $\Delta_n(\theta)$ converges uniformly over Θ to the profiled Kulback-Leibler information $\Delta_0(\theta)$, then identifiability of the Kulback-Leibler information for $L_n(\theta, \eta)$ is sufficient.

This approach works for the Cox model for right-censored data, as we will see below.

However, this strategy based on $\Delta_0(\theta)$ is usually not fruitful, and it seems to be easier to establish (1) directly.

The Condition (1) is needed because the integration in (3) is over all of Θ, and thus it is important to guarantee that there are no other distinct modes besides $\hat{\theta}_n$ in the limiting posterior.

Condition (19.10) is not sufficient for this since it only applies to shrinking neighborhoods of θ_0 and not to all of Θ as required.

The examples and simulation studies in Lee, Kosorok and Fine demonstrate that the profile sampler works very well and is in general computationally efficient.

The Metropolis algorithm applied to $p_{\theta,n}(\theta)$ with a Lebesgue prior measure is usually quite easy to tune and seems to achieve equilibrium quickly.

By the ergodic theorem, there exists a sequence of finite chain lengths $\{M_n\} \to \infty$ so that

- the chain mean

$$\bar{\theta}_n \equiv M_n^{-1} \sum_{j=1}^{M_n} \theta^{(j)}$$

satisfies

$$\bar{\theta}_n = \hat{\theta}_n + o_{P_0}(n^{-1/2});$$

- the standardized sample variance

$$V_n \equiv M_n^{-1} \sum_{j=1}^{M_n} n(\theta^{(i)} - \bar{\theta}_n)(\theta^{(i)} - \bar{\theta}_n)'$$

is consistent for $\tilde{I}_0^{-1}$; and

- the empirical measure

$$G_n(A) \equiv M_n^{-1} \sum_{j=1}^{M_n} \mathbf{1}\left\{\sqrt{n}(\theta^{(j)} - \bar{\theta}_n) \in A\right\},$$

for a bounded convex $A \subset \mathbb{R}^k$, is consistent for the probability that a mean zero Gaussian deviate with variance $\tilde{I}_0^{-1}$ lies in A.

Hence the output of the chain can be used for inference about θ_0, provided M_n is large enough so that the sampling error from using a finite chain is negligible.

We now verify the additional Assumption (1) for the Cox model for right censored data and for the Cox model for current status data.

Example 1: The Cox Model for Right Censored Data

For this example, we can use the identifiability of the profile Kulback-Leibler information since the profile likelihood does not involve the nuisance parameter.

Let B be the compact parameter space for β, where β_0 is known to be in the interior of B, and assume that $\|Z\|$ is bounded by a constant.

We know from our previous discussions of this model that $n^{-1}pL_n(\beta)$ equals, up to a constant that does not depend on β,

$$H_n(\beta) \equiv \mathbb{P}_n \left[\int_0^\tau \left(\beta' Z - \log \left[\mathbb{P}_n Y(s) e^{\beta' Z} \right] \right) dN(s) \right].$$

By arguments which are by now familiar to the reader, it is easy to verify that $\|H_n - H_0\|_B \xrightarrow{\mathsf{P}} 0$, where

$$H_0(\beta) \equiv P_0 \left[\int_0^\tau \left(\beta' Z - \log P_0 \left[Y(s) e^{\beta' Z} \right] \right) dN(s) \right].$$

It is also easy to verify that H_0 has

- first derivative

$$U_0(\beta) \equiv P_0 \left[\int_0^\tau (Z - E(s, \beta))\, dN(s) \right],$$

  where $E(s, \beta)$ is as defined in Section 4.2.1,

- and second derivative $-V(\beta)$, where $V(\beta)$ is defined in (4.5).

By the boundedness of $\|Z\|$ and B combined with the other assumptions of the model, it can be shown (see Exercise 19.5.8 below) that there exists a constant $c_0 > 0$ not depending on β such that

$$V(\beta) \geq c_0 \text{var} Z,$$

where for $k \times k$ matrices A and B, $A \geq B$ means that $c'Ac \geq c'Bc$ for every $c \in \mathbb{R}^k$.

Thus H_0 is strictly concave and thus has a unique maximum on B.

It is also easy to verify that $U_0(\beta_0) = 0$ (see Part (b) of Exercise 19.5.8), and thus the unique maximum is located at $\beta = \beta_0$.

Hence

$$\|\Delta_n(\beta) - \Delta_0(\beta)\|_B \xrightarrow{\mathsf{P}} 0,$$

where

$$\Delta_0(\beta) = H_0(\beta) - H_0(\beta_0) \leq 0$$

is continuous, with the last inequality being strict whenever $\beta \neq \beta_0$.

This immediately yields Condition (1) for β replacing θ.

The Cox Model for Current Status Data

For this example, we verify (1) directly.

Let $\tilde{\beta}_n$ be some possibly random sequence satisfying

$$\Delta_n(\tilde{\beta}_n) = o_{P_0}(1),$$

where β is replacing θ.

Fix some $\alpha \in (0,1)$ and note that since $\Delta_n(\tilde{\beta}_n) = o_{P_0}(1)$ and $\Delta_n(\beta_0) \leq 0$ almost surely, we have

$$n^{-1} \sum_{i=1}^{n} \log \left\{ \frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; X_i)}{f(\beta_0, F_0; X_i)} \right\} \geq o_{P_0}(1),$$

where

$$f(\beta, F; X) \equiv \delta \left\{ 1 - \overline{F}(Y)^{\exp(\beta' Z)} \right\} + (1-\delta) \overline{F}(Y)^{\exp(\beta' Z)},$$

$\overline{F} \equiv 1 - F = \exp(-\Lambda)$, and

$$\hat{F}_\beta \equiv 1 - \exp(-\hat{\Lambda}_\beta)$$

is the maximizer of the likelihood over the nuisance parameter for fixed β.

This now implies

$$n^{-1}\sum_{i=1}^{n}\log\left[1+\alpha\left\{\frac{f(\tilde{\beta}_n,\hat{F}_{\tilde{\beta}_n};X_i)}{f(\beta_0,F_0;X_i)}-1\right\}\right]\geq o_{P_0}(1),$$

because

$$\alpha\log(x)\leq\log(1+\alpha\{x-1\})$$

for any $x>0$.

This implies that

$$P_0\log\left[1+\alpha\left\{\frac{f(\tilde{\beta}_n,\hat{F}_{\tilde{\beta}_n};X_i)}{f(\beta_0,F_0;X_i)}-1\right\}\right] \quad\geq\quad o_{P_0}(1) \qquad (4)$$

by Lemma 1 below, since $x\mapsto\log(1+\alpha x)$ is Lipschitz continuous for $x\geq 0$ and $f(\theta_0,F_0;X)\geq c$ almost surely, for some $c>0$.

Because $x \mapsto \log x$ is concave, we now have by Jensen's inequality that

$$P_0 \log \left[1 + \alpha \left\{ \frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; X_i)}{f(\beta_0, F_0; X_i)} - 1 \right\} \right] \leq 0.$$

This combined with (4) implies that

$$P_0 \log \left[1 + \alpha \left\{ \frac{f(\tilde{\beta}_n, \hat{F}_{\tilde{\beta}_n}; X_i)}{f(\beta_0, F_0; X_i)} - 1 \right\} \right] = o_{P_0}(1).$$

This forces the result

$$P_0 \left| \overline{F}_{\tilde{\beta}_n}(Y)^{\exp(\tilde{\beta}_n' Z)} - \overline{F}_0(Y)^{\exp(\beta_0' Z)} \right| = o_{P_0}(1)$$

by the strict concavity of $x \mapsto \log x$.

This, in turn, implies that

$$P_0 \left[\left\{ (\tilde{\beta}_n - \beta_0)'(Z - E[Z|Y]) - c_n(Y) \right\}^2 \middle| Y \right] = o_{P_0}(1),$$

for almost surely all Y, where $c_n(Y)$ is uncorrelated with $Z - \mathsf{E}[Z|Y]$.

Hence

$$\tilde{\beta}_n = \theta_0 + o_{P_0}(1),$$

and Condition (1) now follows.

LEMMA 1. *The class*

$$\mathcal{F} \equiv \{f(\beta, F; X) : \beta \in B, F \in \mathcal{M}\},$$

where $\mathcal{M}$ is the class of distribution functions on $[0, \tau]$, is P_0-Donsker.

Other Methods

- The penalized profile sampler (Cheng and Kosorok, 2007c).
- The exchangible bootstrap (Cheng and Huang, In press, *Annals of Statistics*).
- m within n subsampling (Bickel, Götze and van Zwet, 1997).
- Subsampling (Politis and Ramono, 1994).
- Block jacknife (Ma and Kosorok, 2005a).
- Bayesian methods (Shen, 2002).
- Others.

Efficient Inference for Infinite Dimensional Parameters

For most semiparametric models where the joint parameter is regular, we can assume a little more structure than in the previous paragraphs.

For many jointly regular models, we have that $\eta = A$, where $t \mapsto A(t)$ is restricted to a subset $H \in D[0, \tau]$ of functions bounded in total variation, where $\tau < \infty$.

The composite parameter is thus $\psi = (\theta, A)$.

We endow the parameter space with the uniform norm since this is usually the most useful in applications.

Examples include

- many right-censored univariate regression models,
- including the proportional odds model of Section 15.3,
- certain multivariate survival models, and
- certain biased sampling models.

The index set $\mathcal{H}$ we assume consists of all finite variation functions in $D[0, \tau]$, and we assign to

$$\mathcal{C} = \mathbb{R}^k \times \mathcal{H}$$

the norm

$$\|c\|_{\mathcal{C}} \equiv \|a\| + \|h\|_v,$$

where

- $c = (a, h)$,
- $\|\cdot\|$ is the Euclidean norm, and
- $\|\cdot\|_v$ is the total variation norm on $[0, \tau]$.

We let

$$\mathcal{C}_p \equiv \{c \in \mathcal{C} : \|c\|_{\mathcal{C}} \leq p\},$$

where the inequality is strict when $p = \infty$.

This is the same structure utilized in Section 15.3.4 for the proportional odds model aside from some minor changes in the notation.

The full composite parameter $\psi = (\theta, A)$ can be viewed as an element of $\ell^\infty(\mathcal{C}_p)$ if we define

$$\psi(c) \equiv a'\theta + \int_0^\tau h(s)dA(s), \quad c \in \mathcal{C}_p, \quad \psi \in \Omega \equiv \Theta \times H.$$

As described in Section 15.3.4, Ω thus becomes a subset of $\ell^\infty(\mathcal{C}_p)$, with norm

$$\|\psi\|_{(p)} \equiv \sup_{c \in \mathcal{C}_p} |\psi(c)|.$$

Moreover, if $\|\cdot\|_\infty$ is the uniform norm on Ω, then, for any $1 \leq p < \infty$,

$$\|\psi\|_\infty \leq \|\psi\|_{(p)} \leq 4p\|\psi\|_\infty.$$

Thus the uniform and $\|\cdot\|_{(p)}$ norms are equivalent.

For a direction $h \in \mathcal{H}$, we will perturb A via the one-dimensional submodel

$$t \mapsto A_t(\cdot) = \int_0^{(\cdot)} (1 + th(s))dA(s).$$

We now modify the score notation slightly.

For any $c \in \mathcal{C}$, let

$$\begin{aligned} U[\psi](c) &= \frac{\partial}{\partial t}\ell\left(\theta + ta, A(\cdot) + t\int_0^{(\cdot)} h(s)dA(s)\right)\Big|_{t=0} \\ &= \frac{\partial}{\partial t}\ell\left(\theta + ta, A(\cdot)\right)\Big|_{t=0} + \frac{\partial}{\partial t}\ell\left(\theta, A(\cdot) + t\int_0^{(\cdot)} h(s)dA(s)\right)\Big|_{t=0} \\ &\equiv U_1[\psi](a) + U_2[\psi](h). \end{aligned}$$

Note that

$$\Psi_n(\psi)(c) = \mathbb{P}_n U[\psi](c),$$

and

$$\Psi(\psi)(c) = P_0 U[\psi](c),$$

where $P_0 = P_{\psi_0}$.

In this context, $P_\psi U_2[\psi](h) = 0$ for all $h \in \mathcal{H}$ by definition of the maximum and under identifiability of the model.

It is important to note that the map $\psi \mapsto U[\psi](\cdot)$ actually has domain $\text{lin}\,\Omega$ and range contained in $\ell^\infty(\mathcal{C})$.

We now consider properties of the second derivative of the log-likelihood.

Let $\bar{a} \in \mathbb{R}^k$ and $\bar{h} \in \mathcal{H}$.

For ease of exposition, we will use the somewhat redundant notation $c = (a, h) \equiv (c_1, c_2)$.

We assume the following derivative structure exists and is valid for $j = 1, 2$ and all $c \in \mathcal{C}$:

$$\begin{aligned}
&\frac{\partial}{\partial s} U_j[\theta + s\bar{a}, A + s\bar{h}](c_j)\Big|_{s=0} \\
&= \quad \frac{\partial}{\partial s} U_j[\theta + s\bar{a}, A](c_j)\Big|_{s=0} + \frac{\partial}{\partial s} U_j[\theta, A + s\bar{h}](c_j)\Big|_{s=0}, \\
&\equiv \quad \bar{a}' \hat{\sigma}_{1j}[\psi](c_j) + \int_0^\tau \hat{\sigma}_{2j}[\psi](c_j)(u) d\bar{h}(u),
\end{aligned}$$

where $\hat{\sigma}_{1j}[\psi](c_j)$ is a random k-vector and $u \mapsto \hat{\sigma}_{2j}[\psi](c_j)(u)$ is a random function contained in $\mathcal{H}$.

Denote $\sigma_{jk}[\psi] = P_0\hat{\sigma}_{jk}[\psi]$ and $\sigma_{jk} = \sigma_{jk}[\psi_0]$, for $j, k = 1, 2$, and where $P_0 = P_{\psi_0}$.

Let $\hat{\psi}_n = (\hat{\theta}_n, \hat{A}_n)$ be the maximizers of the log-likelihood.

Then $\Psi_n(\hat{\psi}_n)(c) = 0$ for all $c \in \mathcal{C}$.

Moreover, since lin Ω is contained in $\mathcal{C}$, we have that the map

$$\bar{c} \in \text{lin}\,\Omega \mapsto -\dot{\Psi}(\bar{c})(\cdot) \in \ell^\infty(\mathcal{C})$$

has the form $-\dot{\Psi}(\bar{c})(\cdot) = \bar{c}(\sigma(\cdot))$, where $\sigma \equiv \sigma[\psi_0]$ and

$$\sigma[\psi](c) \equiv \begin{pmatrix} \sigma_{11}[\psi] & \sigma_{12}[\psi] \\ \sigma_{21}[\psi] & \sigma_{22}[\psi] \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

and, for any $c, \bar{c} \in \mathcal{C}$,

$$\bar{c}(c) = \bar{c}_1' c_1 + \int_0^\tau c_2(u) d\bar{c}_2(u).$$

Provided $\sigma : \mathcal{C} \mapsto \mathcal{C}$ is continuously invertible and onto, we have that $\dot{\Psi} : \text{lin}\, \Omega \mapsto \mathcal{C}$ is also continuously invertible and onto with inverse satisfying $-\dot{\Psi}^{-1}(c)(\cdot) = c(\sigma^{-1}(\cdot))$.

In this set-up, we will need the following conditions for some $p > 0$:

$$\{U[\psi](c) : \|\psi - \psi_0\| \leq \epsilon, c \in \mathcal{C}_p\} \text{ is Donsker for some } \epsilon > 0, \quad (5)$$

$$\sup_{c \in \mathcal{C}_p} P_0 \left| U[\psi](c) - U[\psi_0](c) \right|^2 \to 0, \text{ as } \psi \to \psi_0, \text{ and} \quad (6)$$

$$\sup_{c \in \mathcal{C}_p} \left\| \sigma[\psi](c) - \sigma[\psi_0](c) \right\|_{(p)} \to 0, \text{ as } \|\psi - \psi_0\|_{(p)} \to 0. \quad (7)$$

Note by Exercise 20.3.1 that (7) implies Ψ is Fréchet-differentiable in $\ell^\infty(\mathcal{C}_p)$.

It is also not hard to verify that if Conditions (5)–(7) hold for some $p > 0$, then they hold for all $0 < p < \infty$ (see Exercise 20.3.2).

This yields the following corollary:

COROLLARY 1. *Assume Conditions (5)–(7) hold for some $p > 0$, that $\sigma : \mathcal{C} \mapsto \mathcal{C}$ is continuously invertible and onto, and that $\hat{\psi}_n$ is uniformly consistent for ψ_0 with*

$$\sup_{c \in \mathcal{C}_1} \left| \mathbb{P}_n \Psi_n(\hat{\psi}_n)(c) \right| = o_{P_0}(n^{-1/2}).$$

Then $\hat{\psi}_n$ is efficient with

$$\sqrt{n}(\hat{\psi}_n - \psi_0)(\cdot) \rightsquigarrow Z(\sigma^{-1}(\cdot))$$

in $\ell^\infty(\mathcal{C}_1)$, where Z is the tight limiting distribution of $\sqrt{n}\mathbb{P}_n U[\psi_0](\cdot)$.

Example 1: The Cox Model for Right-Censored Data

This model has been explored extensively in previous sections, and both weak convergence and efficiency have already been established.

Nevertheless, it is useful to study this model again from the perspective of this section.

We make the usual assumptions for this model as done in Section 4.2.2, including requiring the baseline hazard to be continuous, except that we will use (θ, A) to denote the model parameters (β, Λ).

It is not hard to verify that

- $U_1[\psi](a) = \int_0^\tau Z' a dM_\psi(s)$
- and $U_2[\psi](h) = \int_0^\tau h(s) dM_\psi(s)$,
- where $M_\psi(t) \equiv N(t) - \int_0^t Y(s) e^{\theta' Z} dA(s)$
- and N and Y are the usual counting and at-risk processes.

It is also easy to show that the components of σ are defined by

$$\begin{aligned}
\sigma_{11}a &= \int_0^\tau P_0\left[ZZ'Y(s)e^{\theta_0'Z}\right] dA_0(s)a, \\
\sigma_{12}h &= \int_0^\tau P_0\left[ZY(s)e^{\theta_0'Z}\right] h(s)dA_0(s), \\
\sigma_{21}a &= P_0\left[Z'Y(\cdot)e^{\theta_0'Z}\right] a, \text{ and} \\
\sigma_{22}h &= P_0\left[Y(\cdot)e^{\theta_0'Z}\right] h(\cdot).
\end{aligned}$$

The maximum likelihood estimator is

$$\hat{\psi}_n = (\hat{\theta}_n, \hat{A}_n),$$

where $\hat{\theta}_n$ is the maximizer of the well-known partial likelihood and $\hat{A}_n$ is the Breslow estimator.

Conditions (5)–(7) are easy to verify by recycling arguments we have used previously, and most of the remaining conditions are easy to verify.

The only somewhat difficult condition to verify is that σ is continuously invertible and onto, but we refer to the book for details.

Example 2: A Biased Sampling Model

We will now consider a special case of a class of biased sampling models (generalizations of case-control designs to vaccine break-through infection studies) which were studied by Gilbert (2000).

The data consists of n i.i.d. realizations of $X = (\delta, Y)$.

Here, $\delta \in \{0, 1\}$ is a random stratum identifier, taking on the value j with selection probability $\lambda_j > 0$, $j = 0, 1$, with $\lambda_0 + \lambda_1 = 1$.

Given $\delta = j$, $Y \in [0, \tau]$ has distribution F_j defined on a sigma field of subsets $\mathcal{B}$ of $[0, \tau]$ by

$$F_j(B, \theta, A) \equiv W_j^{-1}(\theta, A) \int_B w_j(u, \theta) dA(u)$$

for $B \in \mathcal{B}$.

The w_j, $j = 0, 1$, are nonnegative (measurable) stratum weight functions assumed to be known up to the finite dimensional parameter $\theta \in \Theta \subset \mathbb{R}$.

We will assume hereafter that $w_0(t, \theta) = 1$ and that $w_1(t, \theta) = e^{\theta t}$.

$$W_j(\theta, A) \equiv \int_0^\tau w_j(u, \theta) dA(u)$$

is assumed to be finite for all $\theta \in \Theta$.

The probability measure A is the unknown infinite dimensional parameter of interest, and $\psi = (\theta, A)$ is the joint parameter.

We assume that A_0 is continuous with support on all of $[0, \tau]$.

The goal is to estimate ψ based on information from samples from the F_j distributions, $j = 0, 1$.

Thus the log-likelihood for a single observation is

$$\begin{aligned} \ell(\psi)(X) &= \log w_\delta(Y, \theta) + \log \Delta A(Y) - \log W_\delta(\theta, A), \\ &= \delta\theta Y + \log \Delta A(Y) - \log \int_0^\tau e^{\delta\theta s} dA(s), \end{aligned}$$

where $\Delta A(Y)$ is the probability mass of A at Y.

Thus the score functions are

$$U_1[\psi](a) = \delta \left[Y - \frac{\int_0^\tau s e^{\delta\theta s} dA(s)}{\int_0^\tau e^{\delta\theta s} dA(s)} \right] a, \quad \text{and}$$

$$U_2[\psi](h) = h(Y) - \frac{\int_0^\tau e^{\delta\theta s} h(s) dA(s)}{\int_0^\tau e^{\delta\theta s} dA(s)}.$$

The components of σ are obtained by taking the expectations under the true distribution P_0 of $\hat{\sigma}_{jk}$, $j, k = 1, 2$, where

$$\begin{aligned}
\hat{\sigma}_{11} a &= \left(E_\delta([\delta y]^2) - [E_\delta(\delta y)]^2 \right) a, \\
\hat{\sigma}_{21} a &= \frac{e^{\delta\theta_0(\cdot)}}{\int_0^\tau e^{\delta\theta_0 s} dA_0(s)} \left[\delta(\cdot) - E_\delta(\delta y) \right] a, \\
\hat{\sigma}_{12} h &= E_\delta(\delta y h(y)) - E_\delta(\delta y) E_\delta(h(y)), \quad \text{and} \\
\hat{\sigma}_{22} h &= \frac{e^{\delta\theta_0(\cdot)}}{\int_0^\tau e^{\delta\theta_0 s} dA_0(s)} \left[h(\cdot) - E_\delta(h(y)) \right],
\end{aligned}$$

where, for a $\mathcal{B}$-measurable function $y \mapsto f(y)$,

$$E_j(f(y)) \equiv \frac{\int_0^\tau f(y) e^{j\theta_0 y} dA_0(y)}{\int_0^\tau e^{j\theta_0 y} dA_0(y)},$$

for $j = 0, 1$.

Then $\sigma_{jk} = P_0 \hat{\sigma}_{jk}, j, k = 1, 2.$

We now show that $\sigma : \mathcal{C} \mapsto \mathcal{C}$ is continuously invertible and onto.

First, as with the previous example, it is easy to verify that $\sigma = \kappa_1 + \kappa_2$, where $\kappa_1 c \equiv (a, \rho_0(\cdot)h(\cdot))$, $\kappa_2 \equiv \sigma - \kappa_1$,

$$\rho_0(\cdot) \equiv P_0 \left[\frac{e^{\delta\theta_0(\cdot)}}{\int_0^\tau e^{\delta\theta_0 s} dA_0(s)} \right],$$

and where κ_2 is a compact operator and κ_1 is continuously invertible and onto.

Provided we can show that σ is one-to-one, we will be able to utilize again Lemma 6.17 to obtain that σ is continuously invertible and onto.

Our argument for showing that σ is one-to-one will be similar to that used for the Cox model for right censored data.

Accordingly, let $c = (a, h) \in \mathcal{C}$ satisfy $\sigma c = 0$, and let $\bar{c} = (a, \int_0^{(\cdot)} h(s) dA_0(s))$.

Thus $\bar{c}(\sigma c) = 0$.

After some algebra, it can be shown that this implies

$$\begin{aligned} 0 &= P_0 E_\delta \left(a\delta y + h(y) - E_\delta [a\delta y + h(y)] \right)^2 \\ &= \lambda_0 V_0(h(y)) + \lambda_1 V_1(ay + h(y)), \end{aligned} \tag{8}$$

where, for a measurable function $y \mapsto f(y)$, $V_j(f(y))$ is the variance of $f(Y)$ given $\delta = j$, $j = 0, 1$.

Recall that both λ_0 and λ_1 are positive.

Thus, since (8) implies $V_0(h(y)) = 0$, $y \mapsto h(y)$ must be a constant function.

Since (8) also implies $V_1(ay + h(y)) = 0$, we now have that $a = 0$.

Hence $h(Y) = E_\delta(h(y))$ almost surely.

Thus $P_0 h^2(Y) = 0$, which implies $h = 0$ almost surely.

Hence $c = 0$, and thus σ is one-to-one.

Conditions (5)–(7) can be established for $p = 1$ by recycling previous arguments (see Exercise 20.3.6).

Gilbert (2000) showed that the maximum likelihood estimator $\hat{\psi}_n = \arg\max_\psi \mathbb{P}_n l_\psi(X)$ is uniformly consistent for ψ_0.

Since $\Psi_n(\hat{\psi}_n)(c) = 0$ almost surely for all $c \in \mathcal{C}$ by definition of the maximum, all of the conditions of Corollary 1 hold for $\hat{\psi}_n$.

Thus $\hat{\psi}_n$ is efficient.