Introduction to Empirical Processes and Semiparametric Inference
Lecture 25: Semiparametric Models

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

Score Functions and Estimating Equations

A parameter $\psi(P)$ of particular interest is the parametric component θ of a semiparametric model

$$\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\},$$

where Θ is an open subset of $\mathbb{R}^k$ and H is an arbitrary set that may be infinite dimensional.

Tangent sets can be used to develop an efficient estimator for $\psi(P_{\theta,\eta}) = \theta$ through the formation of an *efficient score function*.

In this setting, we consider submodels of the form $\{P_{\theta+ta,\eta_t}, t \in N_\epsilon\}$ that are differentiable in quadratic mean with score function

$$\left.\frac{\partial \log dP_{\theta+ta,\eta_t}}{\partial t}\right|_{t=0} = a'\dot{\ell}_{\theta,\eta} + g,$$

where

- $a \in \mathbb{R}^k$,
- $\dot{\ell}_{\theta,\eta} : \mathcal{X} \mapsto \mathbb{R}^k$ is the ordinary score for θ when η is fixed,
- and where $g : \mathcal{X} \mapsto \mathbb{R}$ is an element of a tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ for the submodel

  $$\mathcal{P}_\theta = \{P_{\theta,\eta} : \eta \in H\}$$

  (holding θ fixed).

This tangent set is the *tangent set for* η and should be rich enough to reflect all parametric submodels of $\mathcal{P}_\theta$.

The tangent set for the full model is

$$\dot{\mathcal{P}}_{P_{\theta,\eta}} = \left\{ a' \dot{\ell}_{\theta,\eta} + g : a \in \mathbb{R}^k, g \in \dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}} \right\}.$$

While $\psi(P_{\theta+ta,\eta_t}) = \theta + ta$ is clearly differentiable with respect to t, we also require that there there exists a function $\tilde{\psi}_{\theta,\eta} : \mathcal{X} \mapsto \mathbb{R}^k$ such that

$$\left.\frac{\partial \psi(P_{\theta+ta,\eta_t})}{\partial t}\right|_{t=0} = a = P\left[\tilde{\psi}_{\theta,\eta}\left(\dot{\ell}'_{\theta,\eta} a + g\right)\right], \qquad (1)$$

for all $a \in \mathbb{R}^k$ and all $g \in \dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$.

After setting $a = 0$, we see that such a function must be uncorrelated with all of the elements of $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$.

Define $\Pi_{\theta,\eta}$ to be the orthogonal projection onto the closed linear span of $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$ in $L_2^0(P_{\theta,\eta})$.

The *efficient score function* for θ is

$$\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta},$$

while the *efficient information matrix* for θ is

$$\tilde{I}_{\theta,\eta} = P\left[\tilde{\ell}_{\theta,\eta}\tilde{\ell}'_{\theta,\eta}\right].$$

Provided that $\tilde{I}_{\theta,\eta}$ is nonsingular, the function

$$\tilde{\psi}_{\theta,\eta} = \tilde{I}_{\theta,\eta}^{-1} \tilde{\ell}_{\theta,\eta}$$

satisfies (1) for all $a \in \mathbb{R}^k$ and all $g \in \dot{\mathcal{P}}_{P_{\theta,\eta}}^{(\eta)}$.

Thus the functional (parameter) $\psi(P_{\theta,\eta}) = \theta$ is differentiable at $P_{\theta,\eta}$ relative to the tangent set $\dot{\mathcal{P}}_{P_{\theta,\eta}}$, with efficient influence function $\tilde{\psi}_{\theta,\eta}$.

Recall that the search for an efficient estimator of θ is over if one can find an estimator T_n satisfying

$$\sqrt{n}(T_n - \theta) = \sqrt{n}\mathbb{P}_n\tilde{\psi}_{\theta,\eta} + o_P(1).$$

Note that

$$\tilde{I}_{\theta,\eta} = I_{\theta,\eta} - P\left[\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}\left(\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}\right)'\right],$$

where

$$I_{\theta,\eta} = P\left[\dot{\ell}_{\theta,\eta}\dot{\ell}'_{\theta,\eta}\right].$$

An intuitive justification for the form of the efficient score is that some information for estimating θ is lost due to a lack of knowledge about η.

The amount subtracted off of the efficient score, $\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$, is the minimum possible amount for regular estimators when η is unknown.

Consider again the semiparametric regression model:

$$Y = \beta' Z + e,$$

where

- $\mathsf{E}[e|Z] = 0$ and $\mathsf{E}[e^2|Z] \leq K < \infty$ almost surely,
- and where we observe (Y, Z),
- with the joint density η of (e, Z) satisfying $\int_{\mathbb{R}} e\eta(e, Z)de = 0$ almost surely.

Assume η has partial derivative with respect to the first argument, $\dot{\eta}_1$, satisfying

$$\frac{\dot{\eta}_1}{\eta} \in L_2(P_{\beta,\eta}),$$

and hence

$$\frac{\dot{\eta}_1}{\eta} \in L_2^0(P_{\beta,\eta}),$$

where $P_{\beta,\eta}$ is the joint distribution of (Y, Z).

The Euclidean parameter of interest in this semiparametric model is $\theta = \beta$.

The score for β, assuming η is known, is

$$\dot{\ell}_{\beta,\eta} = -Z(\dot{\eta}_1/\eta)(Y - \beta' Z, Z),$$

where we use the shorthand $(f/g)(u,v) = f(u,v)/g(u,v)$ for ratios of functions.

One can show that the tangent set $\dot{\mathcal{P}}^{(\eta)}_{P_{\beta,\eta}}$ for η is the subset of $L_2^0(P_{\beta,\eta})$ which consists of all functions $g(e,Z) \in L_2^0(P_{\beta,\eta})$ which satisfy

$$\mathsf{E}[eg(e,Z)|Z] = \frac{\int_{\mathbb{R}} eg(e,Z)\eta(e,Z)de}{\int_{\mathbb{R}} \eta(e,Z)de} = 0,$$

almost surely.

One can also show that this set is the orthocomplement in $L_2^0(P_{\beta,\eta})$ of all functions of the form $ef(Z)$, where f satisfies $P_{\beta,\eta}f^2(Z) < \infty$.

This means that

$$\tilde{\ell}_{\beta,\eta} = (I - \Pi_{\beta,\eta})\dot{\ell}_{\beta,\eta}$$

is the projection in $L_2^0(P_{\beta,\eta})$ of $-Z(\dot{\eta}_1/\eta)(e, Z)$ onto $\{ef(Z) : P_{\beta,\eta}f^2(Z) < \infty\}$, where I is the identity.

Thus

$$\begin{aligned}\tilde{\ell}_{\beta,\eta}(Y,Z) &= \frac{-Ze\int_{\mathbb{R}}\dot{\eta}_1(e,Z)ede}{P_{\beta,\eta}[e^2|Z]} \\ &= -\frac{Ze(-1)}{P_{\beta,\eta}[e^2|Z]} = \frac{Z(Y-\beta'Z)}{P_{\beta,\eta}[e^2|Z]},\end{aligned}$$

where

- the second-to-last step follows from the identity

$$\int_{\mathbb{R}}\dot{\eta}_1(e,Z)ede = \left.\frac{\partial\int_{\mathbb{R}}\eta(te,Z)de}{\partial t}\right|_{t=1},$$

- and the last step follows since $e = Y - \beta'Z$.

When the function $z \mapsto P_{\beta,\eta}[e^2|Z = z]$ is non-constant in z,

- $\tilde{\ell}_{\beta,\eta}(Y, Z)$ is not proportional to $Z(Y - \beta' Z)$,
- and the usual least-squares estimator $\hat{\beta}$ will not be efficient.

This is discussed in greater detail in Chapter 4.

Two very useful tools for computing efficient scores are score and information operators.

Returning to the generic semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, sometimes it is easier to represent an element g in the tangent set for η, $\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}}$, as $B_{\theta,\eta}b$, where

- b is an element of another set $\mathbb{H}_\eta$ and
- $B_{\theta,\eta}$ is an operator satisfying

$$\dot{\mathcal{P}}^{(\eta)}_{P_{\theta,\eta}} = \{B_{\theta,\eta}b : b \in \mathbb{H}_\eta\}.$$

Such an operator is a score operator.

The information operator is:

$$B^*_{\theta,\eta} B_{\theta,\eta} : H_\eta \mapsto \overline{\text{lin}}\, H_\eta.$$

If $B^*_{\theta,\eta} B_{\theta,\eta}$ has an inverse, then it can be shown that the efficient score for θ has the form

$$\tilde{\ell}_{\theta,\eta} = \left(I - B_{\theta,\eta} \left[B^*_{\theta,\eta} B_{\theta,\eta} \right]^{-1} B^*_{\theta,\eta} \right) \dot{\ell}_{\theta,\eta}.$$

To illustrate these methods, consider again the Cox model for right-censored data.

Recall in this setting that we observe a sample of n realizations of $X = (V, d, Z)$, where

- $V = T \wedge C$,
- $d = \mathbf{1}\{V = T\}$,
- $Z \in \mathbb{R}^k$ is a covariate vector,
- T is a failure time, and
- C is a censoring time.

We also assume

- that T and C are independent given Z,
- that T given Z has integrated hazard function $e^{\beta' Z}\Lambda(t)$,
- for β in an open subset $B \subset \mathbb{R}^k$
- and Λ is continuous and monotone increasing with $\Lambda(0) = 0$,
- and that the censoring distribution does not depend on β or Λ (i.e., censoring is uninformative).

Recall also the counting and at-risk processes $N(t) = \mathbf{1}\{V \leq t\}d$ and $Y(t) = \mathbf{1}\{V \geq t\}$, and let

$$M(t) = N(t) - \int_0^t Y(s)e^{\beta' Z} d\Lambda(s).$$

For some $0 < \tau < \infty$ with $P\{C \geq \tau\} > 0$, let H be the set of all Λ's satisfying our criteria with $\Lambda(\tau) < \infty$.

Now the set of models $\mathcal{P}$ is indexed by $\beta \in B$ and $\Lambda \in H$.

We let $P_{\beta,\Lambda}$ be the distribution of (V, d, Z) corresponding to the given parameters.

The likelihood for a single observation is thus proportional to

$$p_{\beta,\Lambda}(X) = \left[e^{\beta' Z}\lambda(V)\right]^d \exp\left[-e^{\beta' Z}\Lambda(V)\right],$$

where λ is the derivative of Λ.

Now let $L_2(\Lambda)$ be the set of measurable functions $b : [0, \tau] \mapsto \mathbb{R}$ with

$$\int_0^\tau b^2(s) d\Lambda(s) < \infty.$$

Note that if $b \in L_2(\Lambda)$ is bounded, then

$$\Lambda_t(s) = \int_0^s e^{tb(u)} d\Lambda(u) \in H$$

for all t.

The score function

$$\left.\frac{\partial \log p_{\beta+ta,\Lambda_t}(X)}{\partial t}\right|_{t=0}$$

is thus

$$\int_0^\tau \left[a'Z + b(s)\right] dM(s),$$

for any $a \in \mathbb{R}^k$.

The score function for β is therefore

$$\dot{\ell}_{\beta,\Lambda}(X) = ZM(\tau),$$

while the score function for Λ is

$$\int_0^\tau b(s)dM(s).$$

In fact, one can show that there exists one-dimensional submodels Λ_t such that $\log p_{\beta+ta,\Lambda_t}$ is differentiable with score

$$a'\dot{\ell}_{\beta,\Lambda}(X) + \int_0^\tau b(s)dM(s),$$

for any $b \in L_2(\Lambda)$ and $a \in \mathbb{R}^k$.

The operator

$$B_{\beta,\Lambda} : L_2(\Lambda) \mapsto L_2^0(P_{\beta,\Lambda}),$$

given by

$$B_{\beta,\Lambda}(b) = \int_0^\tau b(s) dM(s),$$

is the score operator which generates the tangent set for Λ,

$$\dot{\mathcal{P}}_{P_{\beta,\Lambda}}^{(\Lambda)} \equiv \{B_{\beta,\Lambda} b : b \in L_2(\Lambda)\}.$$

It can be shown that this tangent space spans all square-integrable score functions for Λ generated by parametric submodels.

The adjoint operator can be shown to be

$$B^*_{\beta,\Lambda} : L_2(P_{\beta,\Lambda}) \mapsto L_2(\Lambda),$$

where

$$B^*_{\beta,\Lambda}(g)(t) = \frac{P_{\beta,\Lambda}[g(X)dM(t)]}{d\Lambda(t)}.$$

The information operator

$$B^*_{\beta,\Lambda} B_{\beta,\Lambda} : L_2(\Lambda) \mapsto L_2(\Lambda)$$

is thus

$$\begin{aligned} B^*_{\beta,\Lambda} B_{\beta,\Lambda}(b)(t) &= \frac{P_{\beta,\Lambda}\left[\int_0^\tau b(s)dM(s)dM(u)\right]}{d\Lambda(u)} \\ &= P_{\beta,\Lambda}\left[Y(t)e^{\beta' Z}\right] b(t), \end{aligned}$$

using martingale methods.

Since

$$B^*_{\beta,\Lambda}\left(\dot{\ell}_{\beta,\Lambda}\right)(t) = P_{\beta,\Lambda}\left[ZY(t)e^{\beta' Z}\right],$$

we have that the efficient score for β is

$$\begin{aligned}\tilde{\ell}_{\beta,\Lambda} &= \left(I - B_{\beta,\Lambda}\left[B^*_{\beta,\Lambda}B_{\beta,\Lambda}\right]^{-1}B^*_{\beta,\Lambda}\right)\dot{\ell}_{\beta,\Lambda} \qquad (2)\\ &= \int_0^\tau \left\{ Z - \frac{P_{\beta,\Lambda}\left[ZY(t)e^{\beta' Z}\right]}{P_{\beta,\Lambda}\left[Y(t)e^{\beta' Z}\right]}\right\} dM(t).\end{aligned}$$

When

$$\tilde{I}_{\beta,\Lambda} \equiv P_{\beta,\Lambda}\left[\tilde{\ell}_{\beta,\Lambda}\tilde{\ell}'_{\beta,\Lambda}\right]$$

is positive definite, the resulting efficient influence function is

$$\tilde{\psi}_{\beta,\Lambda} \equiv \tilde{I}^{-1}_{\beta,\Lambda}\tilde{\ell}_{\beta,\Lambda}.$$

Since the estimator $\hat{\beta}_n$ obtained from maximizing the *partial likelihood*

$$\tilde{L}_n(\beta) \quad = \quad \prod_{i=1}^{n} \left(\frac{e^{\beta' Z_i}}{\sum_{j=1}^{n} \mathbf{1}\{V_j \geq V_i\} e^{\beta' Z_j}} \right)^{d_i} \tag{3}$$

can be shown to satisfy

$$\sqrt{n}(\hat{\beta}_n - \beta) = \sqrt{n}\mathbb{P}_n \tilde{\psi}_{\beta,\Lambda} + o_P(1),$$

this estimator is efficient.

Returning to our discussion of score and information operators, these operators are also useful for generating scores for the entire model, not just for the nuisance component.

With semiparametric models having score functions of the form

$$a'\dot{\ell}_{\theta,\eta} + B_{\theta,\eta}b,$$

for $a \in \mathbb{R}^k$ and $b \in \mathbb{H}_\eta$, we can define a new operator

$$A_{\beta,\eta} : \{(a,b) : a \in \mathbb{R}^k, b \in \text{lin}\,\mathbb{H}_\eta\} \mapsto L_2^0(P_{\theta,\eta})$$

where

$$A_{\beta,\eta}(a,b) = a'\dot{\ell}_{\theta,\eta} + B_{\theta,\eta}b.$$

More generally, we can define the score operator

$$A_\eta : \text{lin}\, \mathbb{H}_\eta \mapsto L_2(P_\eta)$$

for the model $\{P_\eta : \eta \in H\}$, where

- H indexes the entire model
- and may include both parametric and nonparametric components,
- and where lin $\mathbb{H}_\eta$ indexes directions in H.

Let the parameter of interest be $\psi(P_\eta) = \chi(\eta) \in \mathbb{R}^k$.

We assume there exists a linear operator

$$\dot{\chi} : \text{lin}\, \mathbb{H}_\eta \mapsto \mathbb{R}^k$$

such that, for every $b \in \text{lin}\, \mathbb{H}_\eta$, there exists a one-dimensional submodel

$$\{P_{\eta_t} : \eta_t \in H, t \in N_\epsilon\}$$

satisfying

$$\int \left[\frac{(dP_{\eta_t})^{1/2} - (dP_\eta)^{1/2}}{t} - \frac{1}{2} A_\eta b (dP_\eta)^{1/2} \right]^2 \to 0,$$

as $t \downarrow 0$, and

$$\left. \frac{\partial \chi(\eta_t)}{\partial t} \right|_{t=0} = \dot{\chi}(b).$$

We require $\mathbb{H}_\eta$ to be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_\eta$.

The efficient influence function is the solution

$$\tilde{\psi}_{P_\eta} \in \overline{R}(A_\eta) \subset L_2^0(P_\eta)$$

of

$$A_\eta^* \tilde{\psi}_{P_\eta} \quad = \quad \tilde{\chi}_\eta, \tag{4}$$

where $\tilde{\chi}_\eta \in \mathbb{H}_\eta$ satisfies

$$\langle \tilde{\chi}_\eta, b \rangle_\eta = \dot{\chi}_\eta(b)$$

for all $b \in \mathbb{H}_\eta$.

When $A_\eta^* A_\eta$ is invertible, then the solution to (4) can be written

$$\tilde{\psi}_{P_\eta} = A_\eta \left(A_\eta^* A_\eta \right)^{-1} \tilde{\chi}_\eta.$$

In Chapter 4, we utilized this approach to derive efficient estimators for all parameters of the Cox model.

Returning to the semiparametric model setting, where

- $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$,
- Θ is an open subset of $\mathbb{R}^k$, and
- H is a set,

the efficient score can be used to derive *estimating equations* for computing efficient estimators of θ.

Recall that an estimating equation is a data dependent function $\Psi_n : \Theta \mapsto \mathbb{R}^k$ for which an approximate zero yields a Z-estimator for θ.

When $\Psi_n(\tilde{\theta})$ has the form $\mathbb{P}_n \hat{\ell}_{\tilde{\theta},n}$, where $\hat{\ell}_{\tilde{\theta},n}(X|X_1, \ldots, X_n)$ is a function for the generic observation X which depends on the value of $\tilde{\theta}$ and the sample data $X_1, \ldots, X_n$, we have the following estimating equation result:

THEOREM 1. *Suppose that the model* $\{P_{\theta,\eta} : \theta \in \Theta\}$*, where* $\Theta \subset \mathbb{R}^k$*, is differentiable in quadratic mean with respect to* θ *at* (θ, η) *and let the efficient information matrix* $\tilde{I}_{\theta,\eta}$ *be nonsingular.*

Let $\hat{\theta}_n$ *satisfy* $\sqrt{n}\mathbb{P}_n \hat{\ell}_{\hat{\theta}_n,n} = o_P(1)$ *and be consistent for* θ*.*

Also assume that $\hat{\ell}_{\hat{\theta}_n,n}$ *is contained in a* $P_{\theta,\eta}$*-Donsker class with probability tending to 1 and that the following conditions hold:*

$$P_{\hat{\theta}_n,\eta} \hat{\ell}_{\hat{\theta}_n,n} = o_P(n^{-1/2} + \|\hat{\theta}_n - \theta\|), \tag{5}$$

$$P_{\theta,\eta} \left\| \hat{\ell}_{\hat{\theta}_n,n} - \tilde{\ell}_{\theta,\eta} \right\|^2 \xrightarrow{\mathsf{P}} 0, \quad P_{\hat{\theta}_n,\eta} \left\| \hat{\ell}_{\hat{\theta}_n,n} \right\|^2 = O_P(1). \tag{6}$$

Then $\hat{\theta}_n$ is asymptotically efficient at (θ, η).

Returning to the Cox model example, the profile likelihood score is the partial likelihood score $\Psi_n(\tilde{\beta}) = \mathbb{P}_n \hat{\ell}_{\tilde{\beta},n}$, where

$$\begin{aligned}\hat{\ell}_{\tilde{\beta},n}(X = (V, d, Z) | X_1, \ldots, X_n) \\ = \int_0^\tau \left\{ Z - \frac{\mathbb{P}_n \left[Z Y(t) e^{\tilde{\beta}' Z} \right]}{\mathbb{P}_n \left[Y(t) e^{\tilde{\beta}' Z} \right]} \right\} dM_{\tilde{\beta}}(t),\end{aligned}$$

and

$$M_{\tilde{\beta}}(t) = N(t) - \int_0^t Y(u) e^{\tilde{\beta}' Z} d\Lambda(u).$$

We showed in Chapter 4 that all the conditions of Theorem 1 are satisfied for the root of $\Psi_n(\tilde{\beta}) = 0$, $\hat{\beta}_n$, and thus the partial likelihood yields efficient estimation of β.

Maximum Likelihood Estimation

The most common approach to efficient estimation is based on modifications of maximum likelihood estimation that lead to efficient estimates.

These modifications, which we will call "likelihoods," are generally not really likelihoods (products of densities) because of complications resulting from the presence of an infinite dimensional nuisance parameter.

Recall the setting of estimation of an unknown real density $f(x)$ from an i.i.d. sample $X_1, \ldots, X_n$.

The likelihood is $\prod_{i=1}^{n} f(X_i)$, and the maximizer over all densities has arbitrarily high peaks at the observations, with zero at the other values, and is therefore not a density.

This problem can be fixed by using an empirical likelihood $\prod_{i=1}^{n} p_i$, where $p_1, \ldots, p_n$ are the masses assigned to the observations indexed by $i = 1, \ldots, n$ and are constrained to satisfy $\sum_{i=1}^{n} p_i = 1$.

This leads to the empirical distribution function estimator, which is known to be fully efficient.

Consider again the Cox model for right-censored data explored in the previous section.

The density for a single observation $X = (V, d, Z)$ is proportional to

$$\left[e^{\beta' Z}\lambda(V)\right]^d \exp\left[-e^{\beta' Z}\Lambda(V)\right].$$

Maximizing the likelihood based on this density will result in the same problem raised in the previous paragraph.

A likelihood that works is the following, which assigns mass only at observed failure times:

$$L_n(\beta, \Lambda) \quad = \quad \prod_{i=1}^{n} \left[e^{\beta' Z_i} \Delta\Lambda(V_i) \right]^{d_i} \exp\left[-e^{\beta' Z_i} \Lambda(V_i) \right], \quad (7)$$

where $\Delta\Lambda(t)$ is the jump size of Λ at t.

For each value of β, one can maximize or *profile* $L_n(\beta, \Lambda)$ over the "nuisance" parameter Λ to obtain the profile likelihood $pL_n(\beta)$, which for the Cox model is $\exp\left[-\sum_{i=1}^{n} d_i\right]$ times the partial likelihood (3).

Let $\hat{\beta}$ be the maximizer of $pL_n(\beta)$.

Then the maximizer $\hat{\Lambda}$ of $L_n(\hat{\beta}, \Lambda)$ is the "Breslow estimator"

$$\hat{\Lambda}(t) = \int_0^t \frac{\mathbb{P}_n dN(s)}{\mathbb{P}_n \left[Y(s) e^{\hat{\beta}' Z} \right]}.$$

We showed in Chapter 4 that $\hat{\beta}$ and $\hat{\Lambda}$ are both efficient.

Another useful class of likelihood variants are *penalized likelihoods*.

Penalized likelihoods add a penalty term (or terms) in order to maintain an appropriate level of smoothness for one or more of the nuisance parameters.

This method was used in the partly linear logistic regression model described in Chapter 1.

Other methods of generating likelihood variants that work are possible.

The basic idea is that using the likelihood principle to guide estimation of semiparametric models often leads to efficient estimators for the model components which are $\sqrt{n}$ consistent.

Because of the richness of this approach to estimation, one needs to verify for each new situation that a likelihood-inspired estimator is consistent, efficient and well-behaved for moderate sample sizes.

Verifying efficiency usually entails demonstrating that the estimator satisfies the efficient score equation described in the previous section.

Approximately Least-Favorable Submodels

One of the key challenges in this setting is to ensure that the efficient score is a derivative of the chosen log likelihood along some submodel.

Something helpful in this setting are *approximately least-favorable submodels*.

The basic idea is to find a function $\eta_t(\theta, \eta)$ such that

- $\eta_0(\theta, \eta) = \eta$, for all $\theta \in \Theta$ and $\eta \in H$, where $\eta_t(\theta, \eta) \in H$ for all t small enough, and such that
- $\tilde{\kappa}_{\theta_0,\eta_0} = \tilde{\ell}_{\theta_0,\eta_0}$, where

$$\tilde{\kappa}_{\theta,\eta}(x) = \left. \frac{\partial l_{\theta+t,\eta_t(\theta,\eta)}(x)}{\partial t} \right|_{t=0},$$

  $l_{\theta,\eta}(x)$ is the log-likelihood for the observed value x at the parameters (θ, η), and where (θ_0, η_0) are the true parameter values.

Note that we require $\tilde{\kappa}_{\theta,\eta} = \tilde{\ell}_{\theta,\eta}$ only when $(\theta, \eta) = (\theta_0, \eta_0)$.

If $(\hat{\theta}_n, \hat{\eta}_n)$ is the maximum likelihood estimator, i.e., the maximizer of $\mathbb{P}_n l_{\theta,\eta}$, then the function

$$t \mapsto \mathbb{P}_n l_{\hat{\theta}_n + t, \eta_t(\hat{\theta}_n, \hat{\eta}_n)}$$

is maximal at $t = 0$, and thus $(\hat{\theta}_n, \hat{\eta}_n)$ is a zero of $\mathbb{P}_n \tilde{\kappa}_{\tilde{\theta},\tilde{\eta}}$.

Now if $\hat{\theta}_n$ and

$$\hat{\ell}_{\tilde{\theta},n} = \tilde{\kappa}_{\tilde{\theta},\hat{\eta}_n}$$

satisfy the conditions of Theorem 1 at $(\theta, \eta) = (\theta_0, \eta_0)$, then the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically efficient at (θ_0, η_0).

Consider now the maximum likelihood estimator $\hat{\theta}_n$ based on maximizing the joint empirical log-likelihood

$$L_n(\theta, \eta) \equiv n\mathbb{P}_n l(\theta, \eta),$$

where $l(\cdot, \cdot)$ is the log-likelihood for a single observation.

For now, η will be regarded as a nuisance parameter, and thus we can restrict our attention to the profile log-likelihood

$$\theta \mapsto pL_n(\theta) \equiv \sup_{\eta} L_n(\theta, \eta).$$

Note that L_n is a sum and not an average, since we multiplied the empirical measure by n.

While the solution of an efficient score equation need not be a maximum likelihood estimator, it is also possible that the maximum likelihood estimator in a semiparametric model may not be expressible as the zero of an efficient score equation.

This possibility occurs because the efficient score is a projection, and, as such, there is no assurance that this projection is the derivative of the log-likelihood along a submodel.

This is the main issue that motivates approximately least-favorable submodels.

An approximately least-favorable submodel approximates the true least-favorable submodel to a useful level of accuracy that facilitates analysis of semiparametric estimators.

We will now describe this process in generality: the specifics will depend on the situation.

As mentioned previously, we first need a general map from the neighborhood of θ into the parameter set for η, which map we will denote by $t \mapsto \eta_t(\theta, \eta)$, for $t \in \mathbb{R}^k$.

We require that

$$\begin{aligned} \eta_t(\theta, \eta) &\in \hat{H}, \quad \text{for all } \|t - \theta\| \text{ small enough, and} \qquad (8) \\ \eta_\theta(\theta, \eta) &= \eta \quad \text{for any } (\theta, \eta) \in \Theta \times \hat{H}, \end{aligned}$$

where $\hat{H}$ is a suitable enlargement of H that includes all estimators that satisfy the constraints of the estimation process.

Now define the map

$$\ell(t, \theta, \eta) \equiv l(t, \eta_t(\theta, \eta)).$$

We will require several things of $\ell(\cdot, \cdot, \cdot)$, at various point in our discussion, that will result in further restrictions on $\eta_t(\theta, \eta)$.

Define

$$\dot{\ell}(t, \theta, \eta) \equiv \frac{\partial}{\partial t)} \ell(t, \theta, \eta),$$

and let

$$\hat{\ell}_{\theta,n} \equiv \dot{\ell}(\theta, \theta, \hat{\eta}_n).$$

Clearly, $\mathbb{P}_n \hat{\ell}_{\hat{\theta}_n,n} = 0$, and thus $\hat{\theta}_n$ is efficient for θ_0, provided $\hat{\ell}_{\theta,n}$ satisfies the conditions of Theorem 1.

The reason it is necessary to check this even for maximum likelihood estimators is that, as mentioned previously, $\hat{\eta}_n$ is often on the boundary (or even a little bit outside of) the parameter space.

Recall again the Cox model setting for right censored data.

In this case, η is the baseline integrated hazard function which is usually assumed to be continuous.

However, $\hat{\eta}_n$ is the Breslow estimator, which is a right-continuous step function with jumps at observed failure times and is therefore not in the parameter space.

Thus direct differentiation of the log-likelihood at the maximum likelihood estimator will not yield an efficient score equation.

We will also require that

$$\dot{\ell}(\theta_0, \theta_0, \eta_0) \quad = \quad \tilde{\ell}_{\theta_0, \eta_0}. \tag{9}$$

Note that we are only insisting that this identity holds at the true parameter values.

This approximately least-favorable submodel structure is very useful for developing methods of inference for θ.