A New Look at Some Efficiency Results for Semiparametric Models with Missing Data

Bin Nan

Department of Biostatistics
University of Michigan

August 6, 2004
Workshop on Missing Data Problems, Toronto

Joint work with Mary Emond and Jon Wellner

Major References

- Nan, Emond, and Wellner (2000). Information bounds for regression models with missing data. *Technical Report 378*, Department of Statistics, University of Washington.

  It is downloadable at

  - http://www.stat.washington.edu/jaw/RESEARCH/PAPERS/available.html
  - http://www.sph.umich.edu/∼bnan/research/

- Robins, Rotnitzky, and Zhao (1994). Estimation of regression coefficients when some regressors are not always observable. *Journal of the American Statistical Association* 89, 846-866.

A Missing Data Problem

Underlying complete data: i.i.d. copies of

$U^0 \equiv (U_1^0, U_2^0) \sim Q_{\theta,\eta}$ with density $q_{\theta,\eta}$.

- θ: parameter of interest (finite-dimensional)
- η: nuisance parameter(s)

Observed data: i.i.d. copies of

$$\begin{aligned} U &\equiv (U_1^0, RU_2^0, R) \equiv \begin{cases} (U_1^0, U_2^0), & R = 1 \\ U_1^0, & R = 0 \end{cases} \\ &\sim P_{\theta,\eta} \text{ with density } p_{\theta,\eta}. \end{aligned}$$

where $P(R = 1|U_1^0) = \pi(U_1^0) \geq \sigma > 0$.

A Missing Data Problem (Continued)

Observed vs complete:

$$p_{\theta,\eta}(u_1^0, ru_2^0, r) = \left(\pi(u_1^0) q_{\theta,\eta}(u_1^0, u_2^0)\right)^r \cdot \left((1-\pi(u_1^0)) \int q_{\theta,\eta}(u_1^0, u_2^0)\, d\mu(u_2^0)\right)^{1-r},$$

here μ is a dominating measure.

Notation:

- $\dot{\mathcal{Q}}_\eta$: Tangent space for η in model Q, which is the closed linear span of scores for η.
- $\dot{\mathcal{Q}}_\eta^\perp$: The orthogonal complement of $\dot{\mathcal{Q}}_\eta$ in $L_2^0(Q)$, a space of all zero mean and square integrable functions w.r.t. Q.
- l_θ^{*0}: Efficient score function for θ in model Q.
- l_θ^*: Efficient score function for θ in model P.

Question: What is l_θ^*?

Major Result

Theorem 1: (Can be obtained from Robins, Rotnitzky, and Zhao, JASA'94.)

The efficient score for θ in model P has the following form

$$l_\theta^*(U) = \frac{R}{\pi} h^*(U^0) - \frac{R-\pi}{\pi} E\{h^*(U^0)|U_1^0\} \ ,$$

where h^* is the unique function in $\dot{\mathcal{Q}}_\eta^\perp$ satisfying the following operator equation

$$\Pi\left(\frac{1}{\pi} h^* - \frac{1-\pi}{\pi} E(h^*|U_1^0)\Big|\dot{\mathcal{Q}}_\eta^\perp\right) = l_\theta^{*0} \ .$$

Application requires knowing:

- l_θ^{*0}
- $\dot{\mathcal{Q}}_\eta^\perp$
- $\Pi(a|\dot{\mathcal{Q}}_\eta^\perp)$ for all $a \in L_2^0(Q)$

Conditional Mean Restricted Regression Model

Underlying complete data: (S, Y, X)

- Y: response variable
- X: covariate (vector)
- S: auxiliary variable

The model:

$$Y = g(X; \theta) + \epsilon, \qquad E(\epsilon|X) = 0$$

where g is known up to θ.

The model is semiparametric because

$$q_{\theta, f_1, f_2, f_3}(s, y, x) = f_1(s|y, x) f_2(y - g(x; \theta)|x) f_3(x) \ ,$$

where the (conditional) density functions f_1, f_2, and f_3 are unknown infinite dimensional nuisance parameters, i.e.,

$$\eta = (f_1, f_2, f_3) \ .$$

Mean Model: Complete Data

Let $[\,\cdot\,]$ denote the closed linear span. We have

$$\begin{aligned}
\dot{\mathcal{Q}}_{f_1} &= [a_1(S,Y,X) : E(a_1|Y,X) = 0, Ea_1^2 < \infty] \\
\dot{\mathcal{Q}}_{f_2} &= [a_2(Y,X) : E(a_2|X) = 0, E(\epsilon a_2|X) = 0, Ea_2^2 < \infty] \\
\dot{\mathcal{Q}}_{f_3} &= [a_3(X) : E(a_3) = 0, Ea_3^2 < \infty] \ ,
\end{aligned}$$

and they are mutually orthogonal. Hence,

$$\dot{\mathcal{Q}}_\eta = \dot{\mathcal{Q}}_{f_1} + \dot{\mathcal{Q}}_{f_2} + \dot{\mathcal{Q}}_{f_3} \ .$$

Therefor, it can be directly verified that

- $\dot{\mathcal{Q}}_\eta^\perp = \Big\{ h(X)\epsilon : E\{h^2(X)\epsilon^2\} < \infty \Big\}$
- $\Pi(a|\dot{\mathcal{Q}}_\eta^\perp) = E\{a(S,Y,X)\epsilon|X\}\epsilon \Big/ E(\epsilon^2|X)$
- $l_\theta^{*0} = \dot{g}_\theta(X;\theta)\epsilon \Big/ E(\epsilon^2|X), \qquad \dot{g}_\theta = \partial g(\cdot;\theta)/\partial\theta.$

Mean Model: Surrogate Outcome Problem

Observed data:

- Y is missing at random
- X and S are always observable

Efficient score: Applying Theorem 1, we obtain

$$l_\theta^* = \frac{\dot{g}_\theta(X;\theta)}{E(\epsilon^{*2}|X)}\epsilon^* \ ,$$

where

$$\epsilon^* = \frac{R}{\pi}Y - \frac{R-\pi}{\pi}E(Y|S,X) - g(X;\theta) \ .$$

Note: This is the only example I have seen that the efficient score l_θ^* in a semiparametric model with missing data has an explicit closed form.

Mean Model: Surrogate Outcome Problem (Continued)

Applications:

- Chen and Breslow (2004) developed an efficient estimator when both X and S are discrete.
- Nan (2004) developed estimating methods based on l_θ^* by modelling/estimating $E(Y|S, X)$ and $E(\epsilon^{*2}|X)$.

References: Details can be found in

- Chen and Breslow (2004). Semiparametric efficient estimation for the auxiliary outcome problem with the conditional mean model. *Canadian Journal of Statistics* 32, to appear.
- Nan (2004). New estimating methods for surrogate outcome data. Manuscript. (http://www.sph.umich.edu/∼bnan/research/)

Classical Regression Model

Underlying complete data: (Y, X)

- Y: response variable
- X: covariate (vector)

The model: The density function of (Y, X) is

$$q_{\theta,g}(y, x) = f(y|x; \theta)g(x) \ ,$$

where

- $f(y|x : \theta)$ is a known regular parametric conditional density function up to an Euclidean parameter θ,
- $g(x) = \eta$ is an unknown density function of X (an infinite dimensional nuisance parameter).

Note:

- The model is an adaptive semiparametric model due to the factorization of f and g in $q_{\theta,g}$. Thus g is ignorable for the estimation of θ if there is no missing data.
- The adaptation is destroyed when X is sometimes missing.

Classical Regression Model: Complete Data

It is easily seen that

- $\dot{\mathcal{Q}}_\eta = L_2^0(G)$, where G is the c.d.f. of X with density g.

Thus,

- $l_\theta^{*0} = \dot{l}_\theta^0 \equiv \partial \log f(y|x;\theta)/\partial\theta$, the ordinary score function for θ;
- $\dot{\mathcal{Q}}_\eta^\perp = \{h(Y,X) : h \perp L_2^0(G)\}$;
- $\Pi(b|\dot{\mathcal{Q}}_\eta^\perp) = b - E(b|X)$ for all $b \in L_2^0(Q)$.

Classical Regression: Observed Data

Observed data: Let $S(Y, X)$ be a fixed measurable function. Then the observed incomplete data are

$$\begin{cases} (Y, X) & \text{if } R = 1 \\ S(Y, X) & \text{if } R = 0. \end{cases}$$

Efficient score: Applying Theorem 1, we obtain

$$l_\theta^* = \frac{R}{\pi} h^*(Y, X) - \frac{R - \pi}{\pi} E\{h^*(Y, X)|S\},$$

where h^* is the unique solution of the following integral equation:

$$\begin{aligned} h^* &= \pi\{\dot{l}_\theta^0 - E(\dot{l}_\theta^0|X, R = 1)\} + (1 - \pi)E(h^*|S) \\ &\quad - \pi E\Big\{\frac{1 - \pi}{\pi} E(h^*|S)\Big| X, R = 1\Big\} . \end{aligned}$$

Classical Regression: Observed Data (Continued)

Applications:

- Robins, Hsieh, and Newey (1995, *JRSSB*) developed an efficient estimator based on l_θ^* when all components in $S(Y, X) = (Y, X_2)$ are discrete where $X = (X_1, X_2)$.

  Note: Their l_θ^* formula needs some minor correction.

- Breslow, McNeney, and Wellner (2003, *Annals*) studied the missing data problem of Lawless, Kalbfleisch, and Wild (1999, *JRSSB*) where

$$S(Y, X) = \sum_{j=1}^{J} j \, 1\{(Y, X) \in \mathcal{S}_j\}$$

  for a partition $\{\mathcal{S}_j\}_{j=1}^{J}$ of the sample space of (Y, X). They showed that the NPMLE achieves the information bound determined by l_θ^*.

Cox Model

Underlying complete data: (Y, Δ, X)

- $Y = \min(T, C)$: observed time, where

  T: failure time

  C: censoring time

- $\Delta = 1(T \leq C)$: failure indicator

- X: covariate (vector)

Assume T and C are independent given X.

The model: The density function of (Y, Δ, X) is

$$q_{\theta,\lambda,\lambda_G,h}(y, \delta, x) = \Big\{\lambda(y)e^{\theta' x}\Big\}^{\delta} \exp\Big\{-\Lambda(y)e^{\theta' x}\Big\}$$
$$\cdot \Big\{\lambda_G(y|x)\Big\}^{1-\delta} \exp\Big\{-\Lambda_G(y|x)\Big\} h(x) \ ,$$

where

- λ, Λ: failure time baseline hazard and baseline cumulative hazard functions,

- λ_G, Λ_G: censoring time conditional hazard and conditional cumulative hazard functions given covariate,

- h: covariate density function. $\qquad \eta = (\lambda, \lambda_G, h)$.

Cox Model: Complete Data

Direct calculation yields:

$$\begin{aligned}
\dot{\mathcal{Q}}_\lambda &= \left[\int a(t)dM(t) : a \in L_2(Q_{YX,\,\Delta=1}) \right] \\
\dot{\mathcal{Q}}_{\lambda_G} &= \left[\int b(t,X)dM_G(t) : b \in L_2(Q_{YX,\,\Delta=0}) \right] \\
\dot{\mathcal{Q}}_h &= L_2^0(Q_X) \ ,
\end{aligned}$$

and they are mutually orthogonal. Here M and M_G are martingales for the failure and the censoring counting processes, respectively.

Proposition 1: *For any function $b \in L_2^0(Q)$, define*

$$\begin{aligned}
\mathbf{R}_1 b(y,x) &= b(y,1,x) - E\{b(Y,\Delta,X)|Y > y, X = x], \\
\mathbf{R}_2 b(y,x) &= b(y,0,x) - E\{b(Y,\Delta,X)|Y > y, X = x].
\end{aligned}$$

Then b can be decomposed as follows:

$$\begin{aligned}
b(Y,\Delta,X) &= \int \mathbf{R}_1 b(t,X)dM(t) \\
&\qquad + \int \mathbf{R}_2 b(t,X)dM_G(t) + E[b|X].
\end{aligned}$$

Cox Model: Complete Data (Continued)

Define

$$\begin{aligned}\mathbf{D}u &= \int u(t,X)dM(t)\ , \quad \text{and} \\ \mathbf{B}u &= \int \big[u(t,X) - E\{u(Y,X)|Y=t,\Delta=1\}\big]dM(t)\end{aligned}$$

for all $u \in L_2(Q_{YX,\,\Delta=1})$.

Applying Proposition 1, we have

- $\dot{\mathcal{Q}}_\eta^\perp = \big\{\mathbf{B}u : u \in L_2(Q_{YX,\,\Delta=1})\big\}$
- $\Pi(b|\dot{\mathcal{Q}}_\eta^\perp) = \mathbf{B} \circ \mathbf{R}_1 b$ for all $b \in L_2^0(Q)$.
- $l_\theta^{*0} = \mathbf{B}X = \int\{X - E(X|Y=t,\Delta=1)\}dM(t)$.

Cox Model: Case-Cohort Study

Observed data: Let $X = (X_1, X_2)$. The observed data are

$$\begin{cases} (Y, \Delta, X) & \text{if } R = 1 \\ (\Delta, X_2) & \text{if } R = 0 \end{cases}$$

with $\pi(\Delta, X_2) = 1$ if $\Delta = 1$.

Efficient score: Applying Theorem 1, we obtain

$$l_\theta^* = \frac{R}{\pi}\mathbf{D}u^*(Y, \Delta, X) - \frac{R-\pi}{\pi}E\{\mathbf{D}u^*(Y, \Delta, X)|\Delta, X_2\} \ ,$$

where u^* is the unique solution of the following integral equation

$$\begin{aligned} u^*(y, x) &- [\mathbf{K}u^*(y, x) - E\{\mathbf{K}u^*(Y, X)|Y = y, \Delta = 1\}] \\ &= x - E(X|Y = y, \Delta = 1), \end{aligned}$$

and the operator $\mathbf{K}$ is given by

$$\mathbf{K}u(y, x) = E\{b(Y, \Delta, X)|Y > y, X = x\}$$

with

$$b(Y, \Delta, X) = \frac{1-\pi}{\pi}[\mathbf{D}u(Y, \Delta, X) - E\{\mathbf{D}u(Y, \Delta, X)|\Delta, X_2\}].$$

Cox Model: Case-Cohort Study (Continued)

Applications:

- Nan, Emond, and Wellner (2004, *Annals*) calculated efficiency loss of the ad hoc estimator proposed by Self and Prentice (1988, *Annals*), and potential efficiency gain for the stratified case-cohort design.

- Nan (2004) developed a one-step estimator that approximately solves the efficient score equation with initial value obtained from an ad hoc estimating method. The procedure involves solving the integral equation for u^*. When covariates are discrete, the one-step estimator is semiparametric efficient.

  Reference: Nan (2004). Efficient estimation for case-cohort studies. *Canadian Journal of Statistics* 32, to appear.