Semi-parametric efficiency bounds for the Maximum Partial Likelihood Estimator in Nested-Case Control Sampling

L. Goldstein and H. Zhang

Example: Colorado Uranimum Miners Cohort

3,347 miners who worked underground for at least one month.

258 lung cancer deaths between 1950 and 1982.

Determine the relation between radon exposure, smoking, and lung cancer.

Cohort is large, not feasible to obtain accurate exposure and smoking information on all members.

Survival Model

Failure rate at time t for individual $i \in \{1, \ldots, \eta\}$ with covariate $Z_i(t)$ is given by

$$\lambda_i(t) = Y_i(t)\lambda_0(t)e^{\theta_0 Z_i(t)}$$

where θ_0 is the unknown parameter expressing the strength of the relationship between the covariate and the chance of failure, and $\lambda_0(t)$ is an (unknown) baseline hazard rate. The exponential relative risk form is common, but not necessary.

The function $Y_i(t)$ is a censoring indicator, equal to 1 if i is observable at time t, and 0 otherwise.

Semi-parametric model

Primary interest is in θ_0, often in testing $\theta_0 = 0$. The function $\lambda_0(t)$ is a 'nuisance' parameter. Generality is obtained by leaving it unspecified, leading to the semi-parametric model with unknowns (θ_0, λ_0).

Would like to estimate θ_0 efficiently.

Partial Likelihood Estimator: Cox 1972

Suppose we observe individuals $i_1, \ldots, i_n$ fail at times $t_1 < t_2 < \cdots < t_n$. Let

$$\mathcal{R}(t) = \{i : Y_i(t) = 1\} \quad \text{and} \quad \mathcal{R}_j = \mathcal{R}(t_j).$$

The MPLE is a value $\widehat{\theta}$ maximizing

$$L(\theta) = \prod_{j=1}^{n} \frac{e^{\theta Z_{i_j}(t_j)}}{\sum_{k \in \mathcal{R}_j} e^{\theta Z_k(t_j)}},$$

a product of conditional probabilities. The unknown, and unspecified, baseline hazard conveniently drops out.

Requires covariate data on all cohort members.

Cohort Sampling Schemes

May consider sampling schemes (BGL 1995) specified by

$$\pi_t(r|i), \quad r \subset \{1, 2, \ldots, \eta\}$$

the probability of using the set r as the 'sampled risk set' should i fail at time t. May absorb $Y_i(t)$ into $\pi_t(r|i)$, no sampling for i failing when not at risk.

Simple example: 'nested case-control sampling' (Thomas 1977) of $m-1$ controls, with $\eta(t) = |\mathcal{R}(t)|$,

$$\pi_t(r|i) = \binom{\eta(t)-1}{m-1}^{-1} \quad i \in r \subset \mathcal{R}(t) \text{ with } |r| = m.$$

One to One Counter Matching

Suppose we sample one control per case when forming our risk sets.

A control with the same radon exposure as the case is uninformative for inference on the relation between lung cancer and radon exposure. Such a control may be selected in a simple random sample.

In some instances, we may have an easily available 'proxy' for the covariate which may help us avoid uninformative pairs.

One to One Counter Matching

Study of the relation between EMF exposure and leukemia, easily obtainable house 'wire code' information is correlated with (expensive to measure) house radiation level. Based on wire code information, one can sample a control 'counter' to the case, who is more likely to be informative than one chosen by simple random sampling.

Can counter match more generally, m groups with targets m_l in group l. (BL 1995)

Estimation under sampling

Partial likelihood may be formed using same principle as before, take product of conditional probabilities that i failed and r was sampled at time t, given a failure at t.

Nested case control sampling under null $\theta_0 = 0$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d \mathcal{N}\left(0, \frac{m}{m-1}\sigma^2_{\text{Full}}\right),$$

so Asymptotic Relative Efficiency of sampling is

$$\text{ARE} = \frac{m-1}{m}.$$

Efficiency Concerns

There is a price to pay for the generality gained by leaving the baseline unspecified: we are not maximizing a true likelihood. We are taking a product when events are not independent. In addition, all information between failure times is ignored.

Is the resulting estimator, nevertheless, still efficient?

Semiparametric Efficiency: Stein (1956), Hajek (1970,1972), LeCam (1972).

Efficiency: Full Cohort Case

Begun, Hall, Huang, Wellner (1983): MPLE is efficient, covariates constant over time.

Greenwood Wefelmeyer (1990): MPLE is efficient, covariates are random processes.

Efficiency and the MPLE under Sampling

The MPLE is not efficient in the constant covariate case under nested case-control sampling (Robins et al. (1994)). A covariate sampled for a control at time t may contain information available but not used by the MPLE at some later time.

Such reuse results in biases, as controls that are able to survive to be used for repeated failures are no longer representative of the population to whom the failure should be compared.

Efficiency and the MPLE under Sampling

Modified estimators have additional complexity and their efficiency gains may be small.

They may also require additional modeling assumptions, such as assumed parametric forms to model the dependence of informative censoring (Chen (2004)).

Efficiency: Sampling

Having some idea of when the MPLE is efficient for sampling

a. Informs the search for estimators which might improve the MPLE

b. Mandates the use of the MPLE in situations which approximate ones where it is efficient. Thus avoids the use of modified estimators which may only have efficiency gains under conditions not satisfied in general.

Guessing when the MPLE might be efficient

Under conditions and modeling assumptions, the modified estimators take advantage of the use of predicted values of unsampled covariates. When covariates fluctuate rapidly (highly stratified situations) one cannot predict well the values of the covariates at some future time given their values at the present time.

Leads us to consider situations where sets sampled at each failure time are independent.

Is the MPLE efficient again upon ruling out such situations?

Efficiency: Sampling

For simplicity, consider the case where there is no censoring.

Application

Study of occupational exposure to EMF and leukemia (Floderus et al. 1993) in adult male population in mid-Sweden over 1983-1987. 250 cases of Leukemia.

Nested case-control sampling with two controls sampled based on the age of the case, no censoring.

Efficiency: Parametric Model

Information inequality for $p(x;\theta); \theta \in \mathbb{R}^p$,

$$I(\theta) = \mathsf{Var}\left(\frac{\partial \log p(x;\theta)}{\partial \theta}\right),$$

variance of the (mean zero) 'score function' $U(\theta)$. In particular

$$I_{jk}(\theta) = \mathsf{Cov}(U_j(\theta), U_k(\theta)).$$

Under some regularity, for unbiased $\widehat{\theta}$,

$$\mathsf{Var}(\widehat{\theta}) \geq I^{-1}(\theta).$$

Example for $\mathbb{R}^2$

Variance bound,

$$I^{-1}(\theta) = \frac{1}{I_{11}I_{22} - I_{12}^2} \begin{bmatrix} I_{22} & -I_{12} \\ -I_{12} & I_{11} \end{bmatrix}.$$

Information bound for θ_1 in the presence of unknown θ_2 is

$$\frac{I_{11}I_{22} - I_{12}^2}{I_{22}} = I_{11} - \frac{I_{12}^2}{I_{22}} = I_{11}(1 - \frac{\mathsf{Cov}^2(U_1, U_2)}{I_{11}I_{22}}).$$

Effective information for θ_1, with $r = \mathsf{Corr}(U_1, U_2)$ is,

$$I_{11}^* = I_{11}(1 - r^2).$$

Hellinger Derivative

For $p = p(x;\theta)$ a density function, letting p' denote partial with respect to θ,

$$\frac{\partial}{\partial\theta}p^{1/2} = \frac{1}{2}p^{-1/2}p'$$

so in particular

$$\begin{aligned} 4||\frac{\partial}{\partial\theta}p^{1/2}||^2 &= \int\left(\frac{(p')^2}{p}\right) = \mathsf{Var}\left(\frac{p'}{p}\right) \\ &= \mathsf{Var}\left(\frac{\partial}{\partial\theta}\log p(x;\theta)\right) = I(\theta). \end{aligned}$$

Hellinger Score Function

Four times the square of the L^2 norm of the 'Hellinger Score' function $\rho = \partial p^{1/2}/\partial\theta$ equals the information:

$$I(\theta) = 4||\frac{\partial}{\partial\theta}p^{1/2}||^2 = 4||\rho||^2.$$

Information for θ_1 in the presence of θ_2:

$$I_{11}^* = I_{11}(1 - r^2) = 4||\rho_1 - A\rho_2||^2.$$

In general, when there are more parameters, subtract the projection on space spanned by score functions of the other parameters.

Convolution Theorem for Semiparametric Models

The 'effective information' is given by

$$I^* = 4||\rho_0 - A\alpha^*||^2,$$

where ρ_0 is parametric score, and $A\alpha^*$ is the projection of the non-parametric score onto the space spanned by ρ_0. Then under regularity,

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to Z \star S \quad \text{where} \quad Z \sim \mathcal{N}(0, (I^*)^{-1}).$$

Also asymptotically minimax.

Hellinger Type Differentiability

Begun, Hall, Huang, Wellner (1983). Semiparametric model (θ, g). Density of i.i.d observations with distribution $\mathbf{X}$ is given by $f(\mathbf{x}; \theta, g)$. Space of perturbations of (θ, g), are $\{\theta_n, g_n\}_{n \geq 0}$ such that there exists $\tau \in \mathbb{R}$ such that

$$|\sqrt{n}(\theta_n - \theta) - \tau| \to 0,$$

and an α such that

$$||\sqrt{n}(g_n^{1/2} - g^{1/2}) - \alpha|| \to 0 \quad \text{as } n \to \infty.$$

Hellinger Type Differentiability

There exists ρ_0 and a linear operator A such that the sequence of densities given by $f_n = f(\cdot; \theta_n, g_n)$ for $n = 0, 1, \ldots$ satisfies

$$||\sqrt{n}(f_n^{1/2} - f_0^{1/2}) - \zeta|| \to 0 \quad \text{as } n \to \infty$$

for $\zeta = \tau\rho_0 + A\alpha$.

Parametric score ρ_0, nonparametric score operator A, calculate α^* using the normal equations

$$A^*A\alpha^* = A^*\rho_0.$$

Apply Theory to the Cox Model

Relation between hazard and survival functions, at the null,

$$\lambda_0(t) = \frac{g(t)}{1 - G(t)} = \frac{g(t)}{\overline{G}(t)}.$$

Cox model away from the null,

$$\lambda(t) = e^{\theta z}\lambda_0(t) = \frac{e^{\theta z} g(t)}{\overline{G}(t)} \Longleftrightarrow \overline{G}_\theta(t; z) = \overline{G}^{e^{\theta z}}(t).$$

Mix over covariate density $h(z)$ when z unobserved,

$$\overline{G}_\theta(t) = \int \overline{G}_\theta(t; z) h(z) dz.$$

Sampling

Density $f(X;\theta,g,h)$ of observation of failure, failure time, and covariates of failure and $m-1$ controls sampled from the remaining $\eta-1$ is

$$\binom{\eta-1}{m-1}^{-1} e^{\theta z_i} g(t)\overline{G}(t)^{\sum_{j\in r} e^{\theta z_j}-1}\overline{G}_\theta(t)^{\eta-m}h(z_r).$$

For $m=\eta$ we obtain the case of BHHW without censoring.

Covariate Density

Consider covariate density as unknown; makes no difference in the full cohort case, as the covariates are observed. In the sampling case it continues to be true that no covariate distribution needs to be assumed in order to apply the MPLE.

Semiparametric model with parameter (θ, g, h). We consider the special case of the model of greatest interest, the null $\theta_0 = 0$, and denote the null density of the observations as f_0.

Score Operators

The nonparametric baseline and covariate density g and h give rise to operators A and B, on perturbations α and β in $L^2(\nu^+)$ and $L^2(\nu)$, respectively,

$$A\alpha = \left(g^{-1/2}(t)\alpha(t) + \frac{(\eta-1)\int_t^\infty g^{1/2}\alpha d\nu}{\overline{G}(t)} \right) f_0^{1/2},$$

and

$$B\beta = \left(\sum_{j\in r} h^{-1/2}(z_j)\beta(z_j) \right) f_0^{1/2}.$$

Score and Solutions of Normal Equation

$$\rho_0 = \frac{1}{2}\left[z_i + \log \overline{G}(t) \sum_{j \in r}(z_j - EZ) + \eta EZ \log \overline{G}(t)\right] f_0^{1/2},$$

$$\alpha^* = \frac{EZ}{2}\left[1 + \log \overline{G}(t)\right] g^{1/2}(t)$$

and

$$\beta^* = \frac{1}{2} h^{1/2}(z) \frac{\eta - m}{m\eta}(z - EZ)$$

Solving Normal Equations: Result

Let $\eta \geq m \geq 5$. Then for the nested case control model the effective information is

$$I_*^{\eta}(\theta_0) \quad = \quad \mathsf{Var}(Z) \left(\frac{m-1}{m} + \frac{m}{\eta^2} \right).$$

Special case $m = \eta$ recovers full cohort information BHHW result. Taking the limit in η,

$$I_*(\theta_0) = \lim_{\eta \to \infty} I_*^{\eta}(\theta_0) = I_{\mathsf{Full}}(\theta_0) \left(\frac{m-1}{m} \right).$$

Hence the Cox MPLE is again efficient.

Baseline Hazard Estimator

Lower bound holds in both the distributional and minimax senses for subconvex loss functions, $\ell : \mathbb{R} \to \mathbb{R}^+$ such that

$$\{x : \ell(x) \leq y\}$$

is closed, convex and symmetric for all $y \geq 0$.

Similar remarks apply to the Breslow estimator of the baseline hazard, e.g. $\ell(x) = \sup_t |x(t)|$, $\int_0^1 x^2(t)dt$.

Further Directions

1. Compute bound away from the null
2. Include censoring: the operator C for censoring makes for three.
3. Explore other sampling models, such as counter matching