ASYMPTOTIC EFFICIENCY BOUNDS IN SEMI-PARAMETRIC REGRESSION MODELS FOR CASE-CONTROL DATA

Alan Lee

Department of Statistics

University of Auckland

The Problem:

- When modelling the effect of covariates on the incidence of disease, prospective sampling often doesn't generate enough cases, particularly if the disease is rare.
- This leads to poor estimates of regression coefficients

The solution

- Sample separately from the "case" and "control" populations
- But…
  - Inference now depends on the distribution of the covariates – we can ignore this when sampling prospectively
  - Could model the covariate distribution but this is usually too hard
  - An alternative is to treat it non-parametrically

The likelihood

- Let $f_0(x,\beta)$, (resp $f_1(x, \beta)$) be the probability of being a control (resp a case), given covariates x
- Let g(x) be the density of x
- $\pi_1=\int f_1(x,\beta)g(x)dx$ is the probability of being a case
- Conditional density of x given a case is $f_1(x,b)g(x)/\pi_1$, similarly for a control

- This is what we are sampling from

Likelihood (cont)

- Likelihood is $l(\beta, g)$
- Scott-Wild technique is to profile out g and maximize profile likelihood over β
- Is this efficient?
- What does efficiency mean in this context anyway?

Semi-parametric efficiency

- If $g \in G$ where G is an infinite dimensional index set, consider a ***finite-dimensional submodel*** g_t, $t \in T$, where g_t is in G for all t in T
- The true g, g_0 say, is g_{t0} for some t_0 in T
- Consider the space spanned by S_t
- Take closure of unions of all such spaces, this is the *nuisance tangent space (NTS)*

Efficient score

- Projection of S_β onto the orthogonal complement of the NTS is the *efficient score* S_{eff}
- $S_{eff} = S_\beta - \eta_{MIN}$ where η_{MIN} is the element in the NTS that minimizes

$$E|| S_\beta - \eta ||^2$$

- This is the *projection theorem*

Information bound

- For a "reasonable" estimate of β, the asymptotic var of the estimate satisfies

  Avar est $\geq$ B

  where $B = E(S_{eff}S_{eff}^T)^{-1}$ is the *information bound*

Case-control

- If we have J disease states, we need a multi-sample version of the previous theory, corresponding to the densities

  $$p_j = f_j(x,b)g(x)/\pi_j$$

  for the jth disease state.
- The efficient score now has J elements, one for each disease state

Case-control (cont)

- Can still use the projection theorem
- The analogue of $E||S_\beta-\eta||^2$ is

$$w_1 E_1||S_{\beta 1}-\eta_1||^2 + \ldots + w_J E_J||S_{\beta J}-\eta_J||^2$$

- We can get an explicit expression for this in terms of inner products in a certain L_2 space

Minimising the squared norm

Last expression can be written

$$(h, Ah)_2 - 2(\phi, h)_2 + const$$

$$= ((h^* - h), A(h - h^*))_2 + (h^*, Ah^*)_2 + const$$

where h is in $L_2(G_0)$, G_0 is df of true density g_0,

A is a positive-definite self-adjoint operator,

and h^* solves the "operator equation" $Ah^* = \phi$.

The squared norm is minimized at $h = h^*$

Solving the operator equation

- The operator equation can be solved explicitly: we get

$$h^* = \phi / f^* + c_1 P_1^* + \ldots + c_J P_J^*$$

which gives a formula for the efficient score and hence the information bound

- IB is inverse of $I_{\beta\beta} - I_{\beta\rho} I_{\rho\rho}^{-1} I_{\rho\beta}$

- See next slide for definitions

Math stuff (J=2)

$$f^* = \frac{w_0}{\pi_0} f_0 + \frac{w_1}{\pi_1} f_1, \quad P_j^* = \frac{\frac{w_j}{\pi_j} f_j}{f^*}, \quad \rho = \frac{w_0 \pi_1}{w_1 \pi_0}$$

$$I_{\beta\beta} = \int (S_0 - S_1)(S_0 - S_1)^T P_0^* P_1^* f^* dG_0$$

$$I_{\beta\rho} = \int (S_0 - S_1) P_0^* P_1^* f^* dG_0$$

$$I_{\rho\rho} = \int P_0^* P_1^* f^* dG_0$$

Scott-Wild estimating equations (1)

These are derived by (partly) profiling out g. The g maximizing the likelihood takes the form

$$\hat{g}(x,\beta) = \frac{f^*(x) g_0(x)}{\sum_{j=1}^{J} \mu_j f_j(x,\beta)}$$

where the μ's satisfy

$$\sum_{i=1}^{n_j} P_j^*(x_{ij},\beta,\mu) = n_j,\ j = 1,2,\ldots,J,\ \ P_j^*(x,\beta,\mu) = \frac{\mu_j f_j(x,\beta)}{\sum_{j=1}^{J} \mu_j f_j(x,\beta)}$$

Scott-Wild estimating equations (2)

Substituting back, we get

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} \frac{\partial \log P_j^*(x_{ij}, \beta, \mu)}{\partial \phi} = 0$$

$$\phi = (\beta, \mu)$$

Derivatives wrt μ equal zero iff last equation on previous slide is satisfied

Asymptotic distribution

Using a multi-sample version of standard results for M-estimators of finite-dimensional parameters, we get

$$\sqrt{n}(\hat{\phi}_n - \phi_0) = -V^{-1} n^{-1/2} \left\{ \sum_{i=1}^{n_1} \psi_{i1} + \ldots + \sum_{i=1}^{n_J} \psi_{iJ} \right\} + o_P(1)$$

$$\psi_{ij} = \frac{\partial \log P_j^*(x_{i0}, \beta_0, \mu_0)}{\partial \phi}$$

$$V = E(w_1 \frac{\partial \psi_1}{\partial \phi} + w_J \frac{\partial \psi_J}{\partial \phi})$$

Asymptotic distribution (2)

which implies

$$\sqrt{n}(\hat{\phi}_n - \phi_0) \approx N(0, V^{-1}(w_0 E_0 \psi_0 \psi_0^T + w_1 E_1 \psi_1 \psi_1^T) V^{-1})$$

In fact

$$V^{-1}(w_0 E_0 \psi_0 \psi_0^T + w_1 E_1 \psi_1 \psi_1^T) V^{-1} = \begin{bmatrix} (I_{\beta\beta} - I_{\beta\mu} I_{\mu\mu}^{-} I_{\mu\beta})^{-1} & * \\ * & * \end{bmatrix}$$

so that

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \approx N(0, (I_{\beta\beta} - I_{\beta\mu} I_{\mu\mu}^{-} I_{\mu\beta})^{-1})$$

Scott-Wild is efficient

- The asymptotic variance of the Scott-Wild estimator is the inverse of

$$I_{\beta\beta} - I_{\beta\rho} I_{\rho\rho}^{-1} I_{\rho\beta}$$

- This is the information bound!
- Thus, Scott-Wild is efficient.

Alternative approach

- Consider "population expected log-likelihood"

$$w_0 E_0 \log p_0(x, \beta, g) + w_1 E_1 \log p_1(x, \beta, g)$$

- For fixed β, let g(β) be the maximizer over g of the population expected log-likelihood
- A version of Newey's 1994 theorem shows that the efficient scores are

$$\frac{\partial \log p_j(x, \beta, g(\beta))}{\partial \beta}$$

Alternative (2)

- We can find the maximizing g explicitly, as

$$\hat{g}(x,\beta)=\frac{f^*(x)g_0(x)}{\sum_j \mu_j f_j(x,\beta)},\ \text{where}\ \int P_j^*(x,\beta,\mu)f^*(x)g_0(x)dx=w_j$$

and hence calculate the efficient score as

$$\frac{\partial \log p_j(x,\beta,\hat{g}(\beta))}{\partial\beta}=\frac{\partial \log P_j^*(x,\beta,\mu(\beta))}{\partial\beta}$$

- This gives another derivation of the information bound

Asymptotic distribution (2)

The Scott –Wild equation for β is

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{\partial \log P_j^*(x_{ij},\beta,\mu_n(\beta))}{\partial\beta}=0$$

so provided

$$\sqrt{n}(\hat{\beta}_n-\beta_0)=n^{-1/2}V^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{\partial \log P_j^*(x_{ij},\beta_0,\mu(\beta_0))}{\partial\beta}+o_p(1),$$

the estimator is efficient.

Asymptotic distribution (3)

This will follow under reasonable conditions since the estimate $\mu_n(\beta)$ is $\sqrt{n}$-consistent since it is an M-estimate, as it is part of the solution of the basic Scott-Wild estimating equation

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{\partial \log P_j^*(x_{ij},\beta,\mu)}{\partial \phi}=0$$

In fact…

- Murphy and van der Vaart (2000) prove that for any based on a density of the form

  $p(x,\beta,g)$, g infinite dimensional,

  the estimator of β obtained by profiling out g is efficient.
- Their theorem needs strong conditions for its validity, to cope with the infinite dimensional nature of g
- A "multi-sample" version of their theorem can be proved in the same way, and yields the efficiency directly.

But..

- The case-control problem is essentially finite-dimensional, so does not require such a high-powered approach
- A direct approach using the M-estimator result leads to a simpler proof, requiring only “classical assumptions”

Two-stage case-control

- Same approach works with 2-stage case control
- Sample N individuals prospectively, observe disease status
- Then sample n_j from those having status j
- Equivalent to multi-sample setup with an extra sample of N from Mult(N, $\pi_1,\ldots\pi_J$)

Scott-Wild estimating equations for 2-stage

These now take the form

$$\sum_{j=1}^{J}\sum\nolimits_{i=1}^{n_j}\frac{\partial \log P_j^*(x_{ij},\beta,\mu)}{\partial\phi}$$

$$-\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left\{\log\mu_j-(N_j/n_j-1)\log(N-\mu_j(N-n))\right\}=0$$

$$\phi=(\beta,\mu)$$

$$P_j^*(x,\beta,\mu)=\frac{\mu_j f_j(x,\beta)}{\sum_j \mu_j f_j(x,\beta)}$$

2-stage case-control (cont)

- Can still apply the multi-sample M-estimator theorem to get the asymptotic normality and asymptotic variance
- Using the same approach as before, we can calculate the IB and show that it equals the asymptotic variance of the SW estimator.