Sampling bias in logistic models

Peter McCullagh

Department of Statistics
University of Chicago

Taipei, June 2007

Outline

Conventional regression model

Units: $u_1, u_2, \ldots$ subjects, patients, plots,
Covariate $x(u_1), x(u_2), \ldots$ (non-random)
Response $Y(u_1), Y(u_2), \ldots$ (random)

Regression model
For each finite subset $u_1, \ldots, u_n$ with $\mathbf{x} = (x(u_1), \ldots, x(u_n))$
Distribution $p_{\mathbf{x}}(\mathbf{y})$ on $\mathcal{R}^n$ depends on $\mathbf{x}$

Example

$$p_{\mathbf{x}}(A; \theta) = N_n(X\beta, \sigma_0^2 I_n + \sigma_1^2 K)(A)$$

$A \subset \mathcal{R}^n$, $K_{ij} = K(x_i, x_j)$
block-factor models, spatial models, generalized spline models,...

Binary regression model

Units: $u_1, u_2, \ldots$ subjects, patients, plots (labelled)
Covariate $x(u_1), x(u_2), \ldots$ (non-random, $\mathcal{X}$ valued)
Process η on $\mathcal{X}$ (Gaussian for example)
Responses $Y(u_1), \ldots$ conditionally independent given η

$$\operatorname{logit} \operatorname{pr}(Y(u) = 1 \mid \eta) = \alpha + \beta x(u) + \eta(x(u))$$

Joint distribution

$$p_{\mathbf{x}}(\mathbf{y}) = E_\eta \prod_{i=1}^{n} \frac{e^{\alpha+\beta x_i+\eta(x_i)}}{1 + e^{\alpha+\beta x_i+\eta(x_i)}}$$

parameters α, β, K. $K(x, x') = \operatorname{cov}(\eta(x), \eta(x'))$.

Binary regression model: computation

Computational problem: The marginal distribution

$$p_{\mathbf{x}}(\mathbf{y}) = \int_{\mathcal{R}^n} \prod_{i=1}^{n} \frac{e^{\alpha+\beta x_i+\eta(x_i)}}{1 + e^{\alpha+\beta x_i+\eta(x_i)}}\, \phi(\eta; K)\, d\eta$$

is not easy to compute.

Options:
Taylor approximation: Laird and Ware; Schall; Breslow and Clayton
Laplace approximation: Wolfinger 1993
Numerical approximation: Egret
Monte Carlo:

But $p_{\mathbf{x}}(\mathbf{y})$ is not the correct likelihood!

Binary regression model (contd)

$$\operatorname{logit} \operatorname{pr}(Y(u) = 1 \mid \eta) = \alpha + \beta x(u) + \eta(x(u))$$

Approximate one-dimensional marginal distribution

$$\operatorname{logit} \operatorname{pr}(Y(u) = 1) = \alpha^* + \beta^* x(u)$$

$|\beta^*| < |\beta|$ (parameter attenuation)
Subject-specific approach versus population-average approach

$$E(Y(u)) = \frac{e^{\alpha^* + \beta^* x(u)}}{1 + e^{\alpha^* + \beta^* x(u)}}$$

$$\operatorname{cov}(Y(u), Y(u')) = V(x(u), x(u'))$$

PA more acceptable than SS?

Properties of conventional regression model

(i) Population $\mathcal{U}$ is a fixed set of labelled units

(ii) Two sets of units having same $\mathbf{x}$ also have same response distribution. (exchangeability, no unmeasured confounders,...)

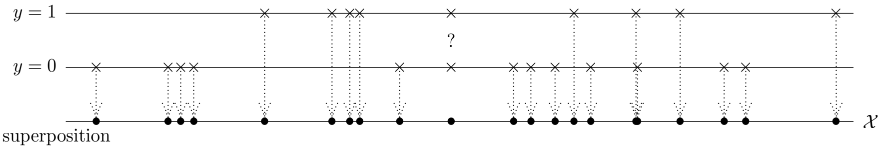
(iii) Distribution of $Y(u)$ depends only on $x(u)$, not on $x(u')$
(no interference, Kolmogorov consistency)

(iii) $u_1, \ldots, u_n$ is a fixed set of units $\Rightarrow$ $\mathbf{x}$ fixed
No concept of random sampling of units

(iv) Does not imply independence of components:
fitted value $E(Y(u')) \neq$ predicted $E(Y(u') \mid \text{data})$

What if ... $u_1, \ldots, u_n$ were generated at random?

Point process model



Intensity $\lambda_1(x)$ for $y = 1$: $m_1(x) = E(\lambda_1(x))$

Intensity $\lambda_0(x)$ for $y = 0$: $m_0(x) = E(\lambda_0(x))$

Intensity $\lambda_{\bullet}(x)$ for superposition: $m_{\bullet}(x) = E(\lambda_{\bullet}(x))$

$\mathrm{pr}(Y(x) = 1 \mid \lambda, x) = \lambda_1(x)/\lambda_{\bullet}(x)$

$\mathrm{pr}(Y(x) = 1 \mid x) = ? \qquad E\left(\frac{\lambda_1(x)}{\lambda_{\bullet}(x)}\right) \quad \text{or} \quad \frac{m_1(x)}{m_{\bullet}(x)} = \frac{E(\lambda_1(x)}{E(\lambda_{\bullet}(x))}$

Point process model

Intensity process $\lambda_0(x)$ for class 0, $\lambda_1(x)$ for class 1
Log ratio: $\eta(x) = \log \lambda_1(x) - \log \lambda_0(x)$
Events form a PP with intensity λ on $\{0, 1\} \times \mathcal{X}$.

$$\mathrm{pr}(Y = 1 \mid x, \lambda) = \frac{\lambda_1(x)}{\lambda_\bullet(x)} = \frac{e^{\eta(x)}}{1+e^{\eta(x)}}$$

$$\mathrm{pr}(Y = 1 \mid x) = E\left(\frac{e^{\eta(x)}}{1+e^{\eta(x)}}\right)$$

Conventional Bayesian calculation, but wrong!

$$\mathrm{pr}(Y(x) = 1 \mid \text{superposition event at } x) = \frac{E\lambda_1(x)}{E\lambda_\bullet(x)}$$

(Correct calculation)
Sampling bias: fixed x versus x in superposition set.

Two ways of thinking

First way: (Conventional Bayesian calculation)

Fix $x \in \mathcal{X}$ and wait for an event to occur at x

$\mathrm{pr}(Y = 1 \mid \lambda, x) = \frac{\lambda_1(x)}{\lambda_\bullet(x)}$

$\mathrm{pr}(Y = 1; x) = E\left(\frac{\lambda_1(x)}{\lambda_\bullet(x)}\right)$

Mathematically correct but seldom relevant

Second way:

First SPP event occurs at x, a random point in $\mathcal{X}$

joint density at (y, x) proportional to $E(\lambda_y(x)) = m_y(x)$

x has marginal density proportional to $E(\lambda_.(x)) = m_.(x)$

$$\mathrm{pr}(Y = 1 \mid x) = \left(\frac{E\lambda_1(x)}{E\lambda_.(x)}\right) \neq E\left(\frac{\lambda_1(x)}{\lambda_.(x)}\right)$$

Explanation of sampling bias

Fix x, x' non-random points in $\mathcal{X}$
No reason to think that $\lambda_{.}(x) > \lambda_{.}(x')$ versus $\lambda_{.}(x') > \lambda_{.}(x)$

Now let x^* be the point where first superposition event occurs
Good reason to think that $\lambda_{.}(x^*) > \lambda_{.}(x)$
because x-values have density $\lambda_{.}(x)$

Correct calculation for predetermined non-random $\mathbf{x}$:

$$p_{\mathbf{x}}(\mathbf{y}) = E \prod_{j=1}^{n} \frac{\lambda_{y_j}(x_j)}{\lambda_{.}(x_j)}$$

Correct calculation for random $\mathbf{x}$

$$p(\mathbf{y} \,|\, \mathbf{x}) = \frac{E \prod \lambda_{y_j}(x_j)}{E \prod \lambda_{.}(x_j)}$$

Consequences of a miscalculation: attenuation

In conventional Bayesian calculation

$$\operatorname{logit}\operatorname{pr}(Y(u) = 1 \mid \eta, x) = \alpha + \beta x(u) + \eta(x(u))$$

implies marginally after integration

$$\operatorname{logit}\operatorname{pr}(Y(u) = 1; x) \simeq \alpha^* + \beta^* x(u)$$

with $\tau = |\beta^*|/|\beta| < 1$, sometimes as small as 1/3.
β called subject-specific effect; β^* population-average effect;

Correct calculation for random x

$$\operatorname{logit}\operatorname{pr}(Y(x) = 1 \mid x \text{ in superposition}) = \alpha^* + \beta x$$

No labelled units, no attenuation, same coefficient β
Distinction between SS effect and PA effect is spurious

Consequences of a miscalculation: inconsistency

Conventional Bayesian likelihood for predetermined **x**:

$$p_{\mathbf{x}}(\mathbf{y}) = E \prod_{j=1}^{n} \frac{\lambda_{y_j}(x_j)}{\lambda_{\bullet}(x_j)}$$

Correct likelihood for random **x**

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{E \prod \lambda_{y_j}(x_j)}{E \prod \lambda_{\bullet}(x_j)}$$

If **x** is randomly generated

parameter estimates based on $p_{\mathbf{x}}(\mathbf{y})$ are inconsistent
bias is approximately $1/\tau > 1$

Consequences: estimating functions

$(\mathbf{y}, \mathbf{x})$ generated at random by PP
Mean intensity for class r: $m_r(x) = E(\lambda_r(x))$
$\pi(x) = m_1(x)/m_.(x); \quad \rho(x) = E(\lambda_1(x)/\lambda_.(x))$

For predetermined x, $E(Y) = \rho(x)$

$$T = \sum_x h(x)(Y(x) - \rho(x))$$

has zero mean for predetermined $\mathbf{x}$. (PA estimating function)

For random x, $E(Y|x \in \mathrm{SPP}) = \pi(x)$

$$T = \sum_{x \in \mathrm{SPP}} h(x)(Y(x) - \pi(x))$$

has zero mean for random $\mathbf{x}$.

Consequences: robustness of PA

Bayes/likelihood has the right target parameter initially
but ignores sampling bias in the likelihood
estimates the right parameter inconsistently.

Population-average estimating equation
establishes the wrong target parameter $\rho(x) = E(Y; x)$
misses the target because sampling bias is ignored
but consistently estimates $\pi(x) = E(Y \mid x \in \mathrm{SPP})$
because conventional notation $E(Y \mid x)$ is ambiguous

PA is remarkably robust but
does not consistently estimate the variance

variance calculation

$(\mathbf{y}, \mathbf{x})$ generated by point process;

$$T(\mathbf{x}, \mathbf{y}) = \sum_{x \in \mathrm{SPP}} h(x)(Y(x) - \pi(x))$$

$$E(T(\mathbf{x}, \mathbf{y})) = 0 \quad E(T \mid \mathbf{x}) \neq 0$$

$$\begin{aligned}
\mathrm{var}(T) &= \int_{\mathcal{X}} h(x)\pi(x)(1 - \pi(x))\, m_{.}(x)\, dx \\
&+ \int_{\mathcal{X}^2} h(x)h(x')[\pi_{11}(x, x') - \pi_{1.}^2(x, x')] m_{..}(x, x')\, dx\, dx' \\
&+ \int_{\mathcal{X}^2} h(x)h(x')[\pi_{1.}(x, x') - \pi(x)]^2 m_{..}(x, x')\, dx\, dx'
\end{aligned}$$