Sampling bias in logistic models

Peter McCullagh

Department of Statistics
University of Chicago

Taipei, June 2007
Outline

1. Conventional regression models
   - Gaussian models
   - Binary regression model
   - Properties of conventional models

2. Unlabelled units
   - Point process model

3. Consequences
   - Sampling bias
   - Non-attenuation
   - Inconsistency
   - Estimating functions
   - Robustness
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

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

$A \subset \mathcal{R}^n$, $K_{ij} = K(x_i, x_j)$
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, $X$ valued)
Process $\eta$ on $X$ (Gaussian for example)
Responses $Y(u_1), \ldots$ conditionally independent given $\eta$

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

Joint distribution

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

parameters $\alpha, \beta, K$. $K(x, x') = \text{cov}(\eta(x), \eta(x'))$. 

Peter McCullagh
Random effects
Conventional regression models
Unlabelled units
Consequences

Gaussian models
Binary regression model
Properties of conventional models

Binary regression model: computation

Computational problem: The marginal distribution

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

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!
Conventional regression models
Unlabelled units
Consequences

Gaussian models
Binary regression model
Properties of conventional models

Binary regression model (contd)

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

Approximate one-dimensional marginal distribution

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

$$|\beta^*| < |\beta|$$ (parameter attenuation)

Subject-specific approach versus population-average approach

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

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

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 \( E(Y(u')) \neq \text{predicted } E(Y(u') \mid \text{data}) \)

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 $\lambda.(x)$ for superposition: $m.(x) = E(\lambda.(x))$

$\Pr(Y(x) = 1 \mid \lambda, x) = \lambda_1(x)/\lambda.(x)$

$\Pr(Y(x) = 1 \mid x) = ? E\left(\frac{\lambda_1(x)}{\lambda.(x)}\right) \text{ or } \frac{m_1(x)}{m.(x)} = \frac{E(\lambda_1(x))}{E(\lambda.(x))}$
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 $\lambda$ on $\{0, 1\} \times \mathcal{X}$.

$$\Pr(Y = 1 \mid x, \lambda) = \frac{\lambda_1(x)}{\lambda_0(x)} = \frac{e^{\eta(x)}}{1 + e^{\eta(x)}}$$

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

Conventional Bayesian calculation, but wrong!

$$\Pr(Y(x) = 1 \mid \text{superposition event at } x) = \frac{E\lambda_1(x)}{E\lambda_0(x)}$$
(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 $x$

$$\text{pr}(Y = 1 \mid \lambda, x) = \frac{\lambda_1(x)}{\lambda^*(x)}$$
$$\text{pr}(Y = 1; x) = E\left(\frac{\lambda_1(x)}{\lambda^*(x)}\right)$$

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 $E(\lambda^*(x)) = m^*(x)$

$$\text{pr}(Y = 1 \mid x) = \left(\frac{E\lambda_1(x)}{E\lambda^*(x)}\right) \neq E\left(\frac{\lambda_1(x)}{\lambda^*(x)}\right)$$
Explanation of sampling bias

Fix $x, x'$ non-random points in $\lambda'$
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 $x$:

$$p_x(y) = E \prod_{j=1}^{n} \frac{\lambda_y(x_j)}{\lambda(x_j)}$$

Correct calculation for random $x$

$$p(y \mid x) = \frac{E \prod \lambda_y(x_j)}{E \prod \lambda(x_j)}$$
Consequences of a miscalculation: attenuation

In conventional Bayesian calculation

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

implies marginally after integration

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

with $\tau = |\beta^*|/|\beta| < 1$, sometimes as small as $1/3$.

$\beta$ called subject-specific effect; $\beta^*$ population-average effect;

Correct calculation for random $x$

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

No labelled units, no attenuation, same coefficient $\beta$

Distinction between SS effect and PA effect is spurious
Consequences of a miscalculation: inconsistency

Conventional Bayesian likelihood for predetermined $\mathbf{x}$:

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

Correct likelihood for random $\mathbf{x}$

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

If $\mathbf{x}$ is randomly generated

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

\((y, x)\) generated at random by PP

Mean intensity for class \(r\): \(m_r(x) = E(\lambda_r(x))\)

\(\pi(x) = \frac{m_1(x)}{m.(x)}; \quad \rho(x) = E(\lambda_1(x)/\lambda.(x))\)

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

\[T = \sum_x h(x)(Y(x) - \rho(x))\]

has zero mean for predetermined \(x\). (PA estimating function)

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

\[T = \sum_{x \in \text{SPP}} h(x)(Y(x) - \pi(x))\]

has zero mean for random \(x\).
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 $\pi(x) = E(Y \mid x \in \text{SPP})$ because conventional notation $E(Y \mid x)$ is ambiguous.

PA is remarkably robust but does not consistently estimate the variance.
(y, x) generated by point process;

\[ T(x, y) = \sum_{x \in \text{SPP}} h(x)(Y(x) - \pi(x)) \]

\[ E(T(x, y)) = 0 \quad E(T | x) \neq 0 \]

\[ \text{var}(T) = \int_{x} h(x)\pi(x)(1 - \pi(x)) m.(x) \, dx \]

\[ + \int_{x^2} h(x)h(x')[\pi_{11}(x, x') - \pi_{1.}^2(x, x')] m..(x, x') \, dx \, dx' \]

\[ + \int_{x^2} h(x)h(x')[\pi_{1.}(x, x') - \pi(x)]^2 m..(x, x') \, dx \, dx' \]