The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1863-13

Advanced School and Conference on Statistics and Applied Probability in Life Sciences

24 September - 12 October, 2007

Sampling bias in logistic models

Peter McCullagh

Department of Statistics
University of Chicago
Chicago IL 60637, USA

Sampling bias in logistic models

Peter McCullagh

Department of Statistics
University of Chicago

Trieste, October 2007

www.stat.uchicago.edu/~pmcc/reports/bias.pdf

Outline

1. Conventional regression models
   - Gaussian models
   - Binary regression model
   - Properties of conventional models
2. Auto-generated units
   - Point process model
3. Consequences of auto-generation
   - Sampling bias
   - Non-attenuation
   - Inconsistency
   - Estimating functions
   - Robustness
   - Interference
4. Arguments pro and con

Conventional regression model

Fixed set $\mathcal{U}$ (usually infinite): $u_1, u_2, \ldots$ subjects, plots,...
Covariate $x(u_1), x(u_2), \ldots$ (non-random, vector-valued)
Response $Y(u_1), Y(u_2), \ldots$ (random, real-valued)

Regression model:
For each sample $u_1, \ldots, u_n$ 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:

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

$A \subset \mathcal{R}^n$, $K_{ij} = K(x_i, x_j)$
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

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

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

Binary regression model: computation

Computational problem:

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

Options:

Taylor approx: Laird and Ware; Schall; Breslow and Clayton, McC and Nelder, Drum and McC,...
Laplace approximation: Wolfinger 1993; Shun and McC 1994
Numerical approximation: Egret
E.M. algorithm: McCulloch 1994 for probit models
Monte Carlo: Z&L,...

Just a minute...

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

Why not?

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 samples 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)

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

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

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

Point process model for auto-generated units

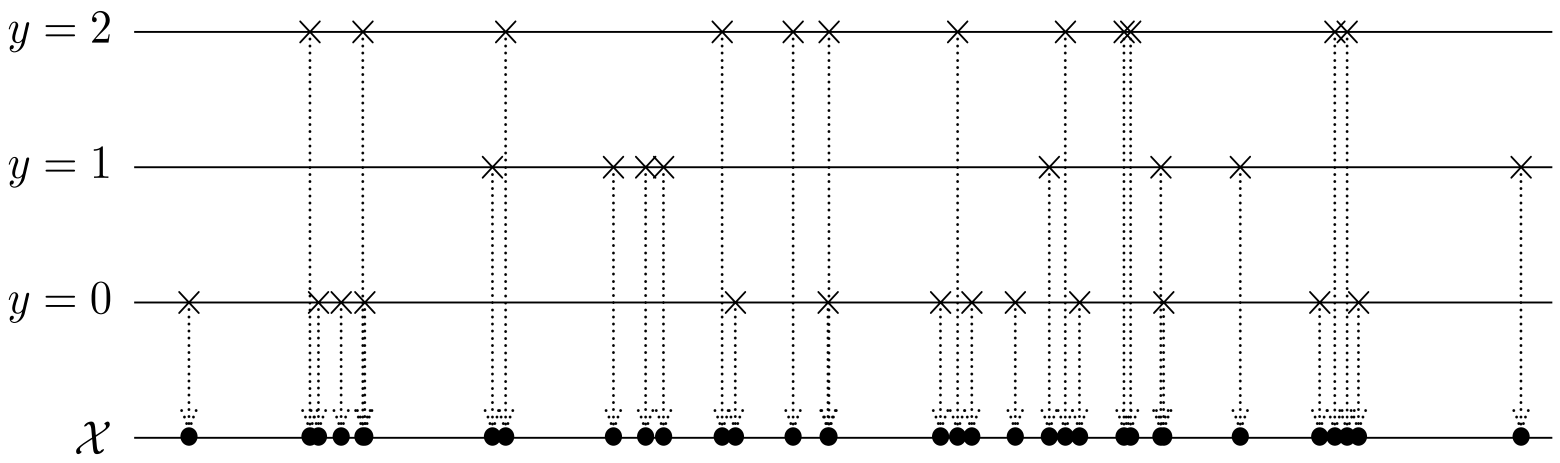

Figure 1: A point process on $\mathcal{C} \times \mathcal{X}$ for $k = 3$, and the superposition process on $\mathcal{X}$.

Intensity $\lambda_r(x)$ for class r

x-values auto-generated by the superposition process with intensity $\lambda_{\bullet}(x)$.

To each auto-generated unit there corresponds an x-value and a y-value. y-value

Binary 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 $\{0, 1\} \times \mathcal{X}$.
Conventional calculation (Bayesian and frequentist):

$$\begin{aligned} \mathrm{pr}(Y = 1 \mid x, \lambda) &= \frac{\lambda_1(x)}{\lambda_\cdot(x)} = \frac{e^{\eta(x)}}{1 + e^{\eta(x)}} \\ \mathrm{pr}(Y = 1 \mid x) &= E\left(\frac{\lambda_1(x)}{\lambda_\cdot(x)}\right) = E\left(\frac{e^{\eta(x)}}{1 + e^{\eta(x)}}\right) \end{aligned}$$

Calculation is correct in a sense, but irrelevant. . .
. . . there might not be an event at x!

Correct calculation for auto-generated units

$$\begin{aligned}
\mathrm{pr}(\text{event of type } r \text{ in } dx \mid \lambda) &= \lambda_r(x)\,dx + o(dx) \\
\mathrm{pr}(\text{event of type } r \text{ in } dx) &= E(\lambda_r(x))\,dx + o(dx) \\
\mathrm{pr}(\text{event in SPP in } dx \mid \lambda) &= \lambda_.(x)\,dx + o(dx) \\
\mathrm{pr}(\text{event in SPP in } dx) &= E(\lambda_.(x))\,dx + o(dx)
\end{aligned}$$

$$\mathrm{pr}(Y(x) = r \mid \text{SPP event at } x) = \frac{E\lambda_r(x)}{E\lambda_.(x)} \neq E\left(\frac{\lambda_r(x)}{\lambda_.(x)}\right)$$

Sampling bias:
Distn for fixed x versus distn for autogenerated x.

Two ways of thinking

First way: waiting for Godot!

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

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

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

Conventional, mathematically correct, but seldom relevant

Second way: come what may!

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 $E(\lambda_\bullet(x)) = m_\bullet(x)$

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

Log Gaussian illustration of sampling bias

$$
\begin{aligned}
\eta_0(x) &\sim GP(0, K), & \lambda_0(x) &= \exp(\eta_0(x)) \\
\eta_1(x) &\sim GP(\alpha + \beta x, K), & \lambda_1(x) &= \exp(\eta_1(x)) \\
\eta(x) = \eta_1(x) - \eta_0(x) &\sim GP(\alpha + \beta x, 2K), & K(x, x) &= \sigma^2
\end{aligned}
$$

One-dimensional sampling distributions:

$$
\begin{aligned}
\rho(x(u)) = \mathrm{pr}(Y(u) = 1) &= E\left(\frac{e^{\alpha+\beta x(u)+\eta(x)}}{1 + e^{\alpha+\beta x(u)+\eta(x)}}\right) \\
\mathrm{logit}(\rho(x)) &\simeq \alpha^* + \beta^* x \qquad (|\beta^*| < |\beta|) \\
\pi(x) = \mathrm{pr}(Y = 1 \mid x \in \mathrm{SPP}) &= \frac{E\lambda_1(x)}{E\lambda_.(x)} = \frac{e^{\alpha+\beta x+\sigma^2/2}}{e^{\sigma^2/2} + e^{\alpha+\beta x+\sigma^2/2}} \\
\mathrm{logit}\,\mathrm{pr}(Y = 1 \mid x \in \mathrm{SPP}) &= \alpha + \beta x
\end{aligned}
$$

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 **x**:

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

Correct calculation for random autogenerated **x**

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

Attenuation

Quota sampling:
Conventional calculation for fixed subject u

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

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.

- Calculation is correct for quota samples (x fixed)
- Both probabilities specific to unit u
- No averaging over units $u \in \mathcal{U}$
- Nevertheless β is called the subject-specific effect
  - β^* is called population averaged effect

Attenuation

Quota sampling:
Conventional calculation for fixed subject u

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

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.

- Calculation is correct for quota samples (x fixed)
- Both probabilities specific to unit u
- No averaging over units $u \in \mathcal{U}$
- Nevertheless β is called the subject-specific effect
  - β^* is called population averaged effect

Attenuation

Quota sampling:
Conventional calculation for fixed subject u

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

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.

Calculation is correct for quota samples (x fixed)
Both probabilities specific to unit u
No averaging over units $u \in \mathcal{U}$
Nevertheless β is called the subject-specific effect
β^* is called population averaged effect

Attenuation

Quota sampling:
Conventional calculation for fixed subject u

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

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.

- Calculation is correct for quota samples (x fixed)
- Both probabilities specific to unit u
- No averaging over units $u \in \mathcal{U}$
- Nevertheless β is called the subject-specific effect
  - β^* is called population averaged effect

Non-attenuation

Sequential sampling for auto-generated units

$$\operatorname{logit}\operatorname{pr}(\,Y(x) = 1 \mid \lambda, \text{event at } x) = \alpha + \beta x + \eta(x)$$

implies marginally after integration

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

- Calculation is correct for autogenerated units
- Both probabilities specific to unit at x
- No averaging over units
- No parameter attenuation for autogenerated units

Non-attenuation

Sequential sampling for auto-generated units

$$\operatorname{logit} \operatorname{pr}(Y(x) = 1 \mid \lambda, \text{event at } x) = \alpha + \beta x + \eta(x)$$

implies marginally after integration

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

Calculation is correct for autogenerated units

Both probabilities specific to unit at x

No averaging over units

No parameter attenuation for autogenerated units

Non-attenuation

Sequential sampling for auto-generated units

$$\operatorname{logit} \operatorname{pr}(Y(x) = 1 \mid \lambda, \text{event at } x) = \alpha + \beta x + \eta(x)$$

implies marginally after integration

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

Calculation is correct for autogenerated units

Both probabilities specific to unit at x

No averaging over units

No parameter attenuation for autogenerated units

Non-attenuation

Sequential sampling for auto-generated units

$$\operatorname{logit} \operatorname{pr}(Y(x) = 1 \mid \lambda, \text{event at } x) = \alpha + \beta x + \eta(x)$$

implies marginally after integration

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

Calculation is correct for autogenerated units

Both probabilities specific to unit at x

No averaging over units

No parameter attenuation for autogenerated units

Consequences: 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 auto-generated **x**

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

If conventional likelihood is used with autogenerated **x**

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

Consequences: estimating functions

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

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

(PA estimating function for $\rho(x)$)

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

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

has zero mean for auto-generated **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: binary case

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

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

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

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

V: spatial or within-cluster correlation;
Δ: interference

What is interference?

Physical interference:

distribution of $Y(u)$ depends on $x(u')$

Sampling interference for autogenerated units

$m_r(x) = E(\lambda_r(x)); \quad m_{rs}(x, x') = E(\lambda_r(x)\lambda_s(x'))$

Univariate distributions: $\pi_r(x) = m_r(x)/m_{\bullet}(x)$

Bivariate: $\pi_{rs}(x, x') = m_{rs}(x, x')/m_{\bullet\bullet}(x, x')$

$\pi_{rs}(x, x') = \mathrm{pr}(Y(x) = r,\ Y(x') = s \mid x, x' \in \mathrm{SPP})$

Hence $\pi_{r\bullet}(x, x') = \mathrm{pr}(Y(x) = r \mid x, x' \in \mathrm{SPP})$

$\Delta_r(x, x') = \pi_{r\bullet}(x, x') - \pi_r(x)$

No second-order sampling interference if $\Delta_r(x, x') = 0$

Autogeneration of units in observational studies

Q: Subject was observed to engage in behaviour X.
What form Y did the behaviour take?

Application	X	Y
Marketing	car purchase	brand
Ecology	sex	activity class
Ecology	play	relatives or non-relatives
Traffic study	highway use	speed
Traffic study	highway speeding	colour of car/driver
Law enforcement	burglary	firearm used?
Epidemiology	birth defect	type of defect
Epidemiology	cancer death	cancer type

Units/events auto-generated by the process

Auto-generation as a model for self-selection

Economics:
Event: single; in labour force; seeks job training
Attributes (Y): (age, job training (Y/N), income)

Epidemiology:
Event: birth defect
Attributes: (age of M, type of defect, state)

Clinical trial:
Event: seeks medical help; diagnosed C.C.; informed consent;
Attributes: (age, sex, treatment status, survival)

What is the population of statistical units?

Mathematical considerations

Restriction: if $p_k()$ is the distribution for k classes,
what is the distribution for $k-1$ classes?
Does restricted model have same form?
Answer:

Weighted sampling

Closure under weighted or case-control sampling

Closure under aggregation of homogeneous classes

Mathematical considerations

Restriction: if $p_k()$ is the distribution for k classes,
what is the distribution for $k-1$ classes?
Does restricted model have same form?
Answer:

Weighted sampling
Closure under weighted or case-control sampling

Closure under aggregation of homogeneous classes

Mathematical considerations

Restriction: if $p_k()$ is the distribution for k classes,
what is the distribution for $k-1$ classes?
Does restricted model have same form?
Answer:

Weighted sampling
Closure under weighted or case-control sampling

Closure under aggregation of homogeneous classes