Sampling bias in logistic models

Peter McCullagh

Department of Statistics
University of Chicago

Royal Statistical Society
Feb 6, 2008
Outline

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

2. Auto-generated units
   - Point process model

3. Consequences of auto-generation
   - Sampling bias
   - Randomization
   - Notation
   - Estimating functions
   - Interference

4. Inference and prediction
Conventional regression models

Auto-generated units

Consequences of auto-generation

Inference and prediction

Gaussian models

Binary regression model

Properties of regression models

Problems with conventional models

Conventional regression model

Fixed set $\mathcal{U}$ (always infinite): $u_1, u_2, \ldots$ called 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:

Sample = finite ordered subset $u_1, \ldots, u_n$ (distinct in $\mathcal{U}$)

For each sample with configuration $\mathbf{x} = (x(u_1), \ldots, x(u_n))$

Distribution $p_{\mathbf{x}}(\mathbf{y})$ on $\mathbb{R}^n$ depends on $\mathbf{x}$

Example:

$$p_{\mathbf{x}}(\mathbf{y} \in A; \theta) = N_n(X\beta, \sigma^2_0 I + \sigma^2_1 K[\mathbf{x}])(A)$$

$A \subset \mathbb{R}^n$, $K[\mathbf{x}] = \{K(x_i, x_j)\}$

block-factor models, spatial models, spline models,...
Binary regression model (GLMM)

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

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

Joint distribution for sample having configuration $x$

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

parameters $\alpha, \beta, K$, $K(x, x') = \text{cov}(\eta(x), \eta(x'))$. 
Binary regression model: computation

GLMM computational problem:

\[ p_x(y) = \int \prod_{i=1}^{n} \frac{e^{(\alpha + \beta x_i + \eta(x_i))y_i}}{1 + e^{\alpha + \beta x_i + \eta(x_i)}} \phi(\eta; K) \, d\eta \]

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,...
But, . . ., wait a minute...

But . . . \( p_x(y) \) is not usually the relevant distribution!

Why not?

Sampling: fixed \( x \) versus random \( x \)
stratum distribution versus conditional distribution
Mathematical structure of regression model

(i) Population $\mathcal{U}$ with covariate structure: $x : \mathcal{U} \rightarrow \mathcal{X}$

(ii) Sample $\mapsto$ configuration $\mathbf{x} = (x(u_1), \ldots, x(u_n))$

(iii) Configuration $\mapsto$ distribution $p_x(\cdot)$ on $\mathcal{R}^n$

(iv) Consistency of $\{p_x(\cdot)\}$ implies process: $Y : \mathcal{U} \rightarrow \mathcal{R}$

(v) Identify $Y$ with response

$\Rightarrow$ exchangeability, no confounders, no interference,...

Does not imply independence of components

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

objectively random versus happenstance samples
Problems in the application of conventional models

Clinical trials / market research / traffic studies / crime...

(i) Sample units generated by a random process
    sequential recruitment, purchase events, traffic studies...

(ii) Population also generated by a random process in time
    animal populations, purchase events, crime events,...

(iii) Sample as a fixed subset: what does this mean?
    — either mathematically or practically

(iv) Samples: random, sequential, quota

(v) Conditional distribution given observed random $x$
    versus stratum distribution for fixed $x$
AND MAKE SURE THAT THE SAMPLE OF VOLUNTEERS IS RANDOM OR MY STUDY IS RUINED!

Questionnaire:
1. Are you random?
   □ YES □ NO

by William C. Volt
A point process on $\mathcal{C} \times \mathcal{X}$ for $\mathcal{C} = \{0, 1, 2\}$, and the superposition process on $\mathcal{X}$.

Intensity $\lambda_r(x)$ for class $r$: $r = 0, 1, 2$.

$x$-values auto-generated by the superposition process with intensity $\lambda_r(x)$.

To each auto-generated unit there corresponds an $x$-value and a $y$-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 $\lambda$ on $\{0, 1\} \times \mathcal{X}$.
Conventional GLMM calculation (Bayesian and frequentist):

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

$$\Pr(Y = 1 \mid x) = E\left(\frac{\lambda_1(x)}{\lambda_0(x)}\right) = E\left(\frac{e^{\eta(x)}}{1 + e^{\eta(x)}}\right)$$

GLMM calculation is correct in a sense, but irrelevant. . .
. . . there might not be an event at $x$!
Correct calculation for auto-generated units

\[
\begin{align*}
\text{pr}(\text{event of type } r \text{ in } dx \mid \lambda) &= \lambda_r(x) \, dx + o(dx) \\
\text{pr}(\text{event of type } r \text{ in } dx) &= E(\lambda_r(x)) \, dx + o(dx) \\
\text{pr}(\text{event in SPP in } dx \mid \lambda) &= \lambda.(x) \, dx + o(dx) \\
\text{pr}(\text{event in SPP in } dx) &= E(\lambda.(x)) \, dx + o(dx)
\end{align*}
\]

\[
\begin{align*}
\text{pr}(Y(x) = r \mid \text{SPP event at } x) &= \frac{E \lambda_r(x)}{E \lambda.(x)} \neq E \left( \frac{\lambda_r(x)}{\lambda.(x)} \right)
\end{align*}
\]

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 \( x \)

\[
\text{pr}(Y = 1 \mid \lambda, x) = \frac{\lambda_1(x)}{\lambda.(x)}
\]

\[
\text{pr}(Y = 1; x) = E\left(\frac{\lambda_1(x)}{\lambda.(x)}\right) = E(Y_i \mid i: X_i = x)
\]

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 \( E(\lambda.(x)) = m.(x) \)

\[
\text{pr}(Y = 1 \mid x \in \text{SPP}) = \frac{E\lambda_1(x)}{E\lambda.(x)} \neq E\left(\frac{\lambda_1(x)}{\lambda.(x)}\right) = E(Y_i \mid i: X_i = x)
\]
Log Gaussian illustration of sampling bias

\[ \eta_0(x) \sim GP(0, K), \quad \lambda_0(x) = \exp(\eta_0(x)) \]
\[ \eta_1(x) \sim GP(\alpha + \beta x, K), \quad \lambda_1(x) = \exp(\eta_1(x)) \]
\[ \eta(x) = \eta_1(x) - \eta_0(x) \sim GP(\alpha + \beta x, 2K), \quad K(x, x) = \sigma^2 \]

One-dimensional sampling distributions:

\[ \rho(x) = p_x(Y = 1) = E \left( \frac{e^{\eta(x)}}{1 + e^{\eta(x)}} \right) \]
\[ \text{logit}(\rho(x)) \simeq \alpha^* + \beta^* x \quad (|\beta^*| < |\beta|) \]
\[ \pi(x) = \text{pr}(Y = 1 \mid x \in \text{SPP}) = \frac{E\lambda_1(x)}{E\lambda_0(x)} = \frac{e^{\alpha + \beta x + \sigma^2/2}}{e^{\sigma^2/2} + e^{\alpha + \beta x + \sigma^2/2}} \]
\[ \text{logit pr}(Y = 1 \mid x \in \text{SPP}) = \alpha + \beta x \]

No random samples; no approximation; no attenuation
Randomization: Simple treatment effect

Event $x \rightarrow$ pair $(t(x), y(x))$, treatment status, outcome

<table>
<thead>
<tr>
<th></th>
<th>$H_0$</th>
<th>$H_a$</th>
<th>Expected intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(t, y)$</td>
<td>$\lambda_y(x)\gamma_t$</td>
<td>$\lambda_y(x)\gamma_{t,y}$</td>
<td>$m_y(x)\gamma_{t,y}$</td>
</tr>
<tr>
<td>$(0, 0)$</td>
<td>$\lambda_0(x)\gamma_0$</td>
<td>$\lambda_0(x)\gamma_{00}$</td>
<td>$m_0(x)\gamma_{00}$</td>
</tr>
<tr>
<td>$(0, 1)$</td>
<td>$\lambda_1(x)\gamma_0$</td>
<td>$\lambda_1(x)\gamma_{01}$</td>
<td>$m_1(x)\gamma_{01}$</td>
</tr>
<tr>
<td>$(1, 0)$</td>
<td>$\lambda_0(x)\gamma_1$</td>
<td>$\lambda_0(x)\gamma_{10}$</td>
<td>$m_0(x)\gamma_{10}$</td>
</tr>
<tr>
<td>$(1, 1)$</td>
<td>$\lambda_1(x)\gamma_1$</td>
<td>$\lambda_1(x)\gamma_{11}$</td>
<td>$m_1(x)\gamma_{11}$</td>
</tr>
</tbody>
</table>

Conditional on $\lambda$:

$$\text{odds}(Y(x) = 1 \mid t, \lambda) = \frac{\lambda_1(x)\gamma_{t1}}{\lambda_0(x)\gamma_{t0}}$$

$$\text{odds ratio} = \tau = \gamma_{11}\gamma_{00}/(\gamma_{01}\gamma_{10})$$

Treatment effect constant in $x$ and independent of $\lambda$
GLMM analysis

Sample \((X, Y, t)_1, (X, Y, t)_2, \ldots\) observed sequentially

GLMM analysis for a single event at \(x\)

\[
\begin{align*}
\text{pr}(Y(x) = 1 \mid t, \lambda) &= \frac{\lambda_1(x) \gamma_{t1}}{\lambda_1(x) \gamma_{t1} + \lambda_0(x) \gamma_{t0}} \\
\text{pr}(Y(x) = 1 \mid t) &= E\left(\frac{\lambda_1(x) \gamma_{t1}}{\lambda_1(x) \gamma_{t1} + \lambda_0(x) \gamma_{t0}}\right) \\
\text{odds ratio} &= \frac{\text{odds}(Y(x) = 1 \mid t = 1)}{\text{odds}(Y(x) = 1 \mid t = 0)} = \tau^* \\
\end{align*}
\]

Conclusion: treatment effect is attenuated \(|\log \tau^*| < |\log \tau|\)
Point process analysis

Sample \((X, Y, t)_1, (X, Y, t)_2, \ldots\) observed sequentially
Correct analysis for a single event at \(x\)

\[
\begin{align*}
\text{pr}(Y(x) = 1 \mid t, \lambda) &= \frac{\lambda_1(x)\gamma_{t1}}{\lambda_1(x)\gamma_{t1} + \lambda_0(x)\gamma_{t0}} \\
\text{pr}(Y(x) = 1 \mid t, x \in \text{SPP}) &= \frac{E(\lambda_1(x))\gamma_{t1}}{E(\lambda_1(x))\gamma_{t1} + E(\lambda_0(x))\gamma_{t0}} \\
\text{odds}(Y(x) = 1 \mid t, x \in \text{SPP}) &= \frac{m_1(x)\gamma_{t1}}{m_0(x)\gamma_{t0}} \\
\text{odds ratio} &= \frac{\gamma_{11}\gamma_{00}}{\gamma_{01}\gamma_{10}} = \tau
\end{align*}
\]

Conclusion: No attenuation of treatment effect.
Notation: meaning of $E(Y_i \mid X_i = x)$

Exchangeable sequence $(Y_1, X_1), (Y_2, X_2), \ldots$ with binary $Y$ conditionally iid given $\lambda$

Stratum $x$: $\mathcal{U}_x = \{i : X_i = x\}$ an infinite random subsequence

Stratum average: $\text{ave}\{Y_i : i \in \mathcal{U}_x\} = \lambda_1(x)/\lambda.(x)$

Stratum mean = expected value of stratum average:

$$\rho(x) = E\left(\frac{\lambda_1(x)}{\lambda.(x)}\right) \approx \frac{e^{\alpha^* + \beta^* x}}{1 + e^{\alpha^* + \beta^* x}}$$

is declared target in much biostatistical work (PA)

In a PP observed sequentially for fixed time $t$

$$\pi(x) = \text{pr}(Y(x) = 1 \mid x \in \text{SPP}) = \frac{E(\lambda_1(x))}{E(\lambda.(x))} = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$
Consequences of ambiguous notation

Sample \((Y_1, X_1), (Y_2, X_2), \ldots\) observed sequentially

\[
\pi(x) = \frac{E(\lambda_1(x))}{E(\lambda(\cdot)(x))} = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} = E(Y_i | X_i = x)
\]

\[
\rho(x) = E\left(\frac{\lambda_1(x)}{\lambda(\cdot)(x)}\right) \approx \frac{e^{\alpha^* + \beta^* x}}{1 + e^{\alpha^* + \beta^* x}} = E(Y_i | i: X_i = x)
\]

(\(\rho(x)\) computed by logistic-normal integral)

Conventional PA estimating function \(Y_i - \rho(x_i)\) is such that

\[
E(Y_1 - \rho(x) | X_1 = x) = \pi(x) - \rho(x) \neq 0
\]

and same for \(Y_2, \ldots\).
Mean intensity for class $r$: $m_r(x) = E(\lambda_r(x))$

$\pi_r(x) = m_r(x)/m.(x); \quad \rho_r(x) = E(\lambda_r(x)/\lambda.(x))$

$E(Y_i) = \rho(x)$ for each $i$ in $U_x = \{u : X_u = x\}$

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

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

has zero mean for auto-generated configurations $x$.
Note: $E(T|x) \neq 0$; average is also over configurations
**Explanation of unbiasedness**

\[ \mathbf{z} = \{ (x_1, y_1), \ldots \} \] configuration generated in \((0, t)\).
\[ \mathbf{x} = \{ x_1, \ldots, \} \] marginal configuration (SPP)

\( \mathbf{z} \) is a random measure with mean \( t \, m_r(x) \, \nu(dx) \) at \((r, x)\)

\( \mathbf{x} \) is marginal random measure with mean \( t \, m_r(x) \, \nu(dx) \) at \(x\)

\[ \pi_r(x) = m_r(x)/m_r(x) \]

Hence \( E(\mathbf{z}(r, dx)) = \pi_r(x) E(\mathbf{x}(dx)) \) for all \((r, x)\) implies

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

\[
= \int_x h(x) \left( \mathbf{z}(r, dx) - \pi_r(x)\mathbf{x}(dx) \right)
\]

has zero expectation. But \( E(T \mid \mathbf{x}) \neq 0 \).
Conventional regression models
Auto-generated units
Consequences of auto-generation
Inference and prediction

Sampling bias
Randomization
Notation
Estimating functions
Interference

**variance calculation: binary case**

\((y, x)\) generated by point process;

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

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

\[
\text{var}(T) = \int_X h^2(x) \pi(x)(1 - \pi(x)) m.(x) \, dx
\]

\[
+ \int_{X^2} h(x)h(x') V(x, x') m..(x, x') \, dx \, dx'
\]

\[
+ \int_{X^2} h(x)h(x') \Delta^2(x, x') m..(x, x') \, dx \, dx'
\]

\(V\): spatial or within-cluster correlation;
\(\Delta\): interference
What is interference?

Physical interference:
- distribution of $Y(u)$ depends on $x(u')$

Sampling interference for autogenerared units

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

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

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

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

Hence $\pi_{r..}(x, x') = \text{pr}(Y(x) = r | x, x' \in SPP)$

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

No second-order sampling interference if $\Delta_r(x, x') = 0$
Inference: Conventional Gaussian model

Model \( \rho_x(A) = N_n(X\beta, \Sigma_x = \sigma^2_0 I_n + K[x])(A) \)

Rationalization \( Y(i) = x'_i \beta + \epsilon_i + \eta(x(i)) \)

Stratum average: \( \bar{Y}(U_x) = \text{ave}\{ Y_i | i: x_i = x \} = x' \beta + \eta(x) \)

Conditional distribution of \( Y_u \) for \( u \in U_x \) given observation \( y, X \)

\[
Y_u | \text{data} \sim N(x' \beta + k' \Sigma_x^{-1}(y - X\beta), \Sigma_{uu} - k' \Sigma_x^{-1} k)
\]

\[
\bar{Y}(U_x) | \text{data} \sim N(x' \beta + k' \Sigma_x^{-1}(y - X\beta), K(x, x) - k' \Sigma_x^{-1} k)
\]

\( k_i = K(x, x_i), \) (such as \( e^{-|x - x_i|} \) or \(|x - x'|^3\))

Stratum average is a random variable, not a parameter
Estimate is a distribution (not a function of sufficient statistic)
Likewise for GLMMs
Inference and prediction for the PP model

For a sequential sample
Observation \((x, y) \equiv (x^{(0)}, \ldots, x^{(k-1)})\)
Product density \(m_r(x^{(r)}) = E(\prod_{x \in x^{(r)}} \lambda_r(x))\) for class \(r\)
Conditional distribution as a random labelled partition of \(x\):

\[ p(y \mid x) \propto m_0(x^{(0)}) \cdot \cdots \cdot m_{k-1}(x^{(k-1)}) \]

For a subsequent autogenerated event

\[ p(Y(x') = r \mid \text{data, } x' \in \text{SPP}) \propto m_r(x^{(r)}, x')/m_r(x^{(r)}) \]

Use likelihood or estimating function to estimate parameters
Use conditional distribution for inference/prediction
Brief summary of conclusions

(i) Reasonable case for fixed-population model in certain areas
labatory work; field trials; veterinary trials;...
(ii) Good case for autogenerated units in other areas
clinical trials; marketing; crime; animal behaviour
(iii) The choice matters in random-effects models
(iv) $\pi(x) = E(Y_i \mid X_i = x)$ versus $\rho(x) = E(Y_i \mid i: X_i = x)$
attenuation or non-attenuation
(v) What is modelled and estimated by PA?
claims to estimate $\rho(x)$ but actually estimates $\pi(x)$
(vi) What does the GLMM likelihood estimate?
Difficult to say; probably neither