Sampling bias in logistic models

Peter McCullagh

Department of Statistics
University of Chicago

University of Wisconsin
Oct 24, 2007

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

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

2. Auto-generated units
   - Point process model

3. Consequences of auto-generation
   - Sampling bias
   - Non-attenuation
   - Inconsistency of estimates
   - Estimating functions
   - Robustness
   - Interference

Peter McCullagh
Conventional regression model

Fixed set $\mathcal{U}$ of units (master-list, usually infinite)
   elements $u_1, u_2, \ldots$ subjects, plots,\ldots
Covariate $x(u_1), x(u_2), \ldots$ (non-random, vector-valued)
Response $Y(u_1), Y(u_2), \ldots$ (random, real-valued)
Sample: $U \subset \mathcal{U}$ finite and non-random
Observation: $(Y, x) = (Y(u), x(u))$ for $u \in U$

Regression model: $p_x(A; \theta) = \Pr(Y \in A; x, \theta)$
   distribution depends on the covariate configuration $x$
   distribution defined for every sample $U \subset \mathcal{U}$
   distributions mutually compatible
Examples

Example I: GLM

independent components, exponential family,...

\[ E(Y(u)) = \mu(x(u)); \quad \eta(u) = g(\mu(x(u))); \quad \eta = X\beta \]

Example II: Linear Gaussian model

\[ \rho_x(Y \in A; \theta) = N_n(X\beta, \sigma_0^2 I_n + \sigma_1^2 K)(A) \]

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

block-factor models, spatial models, generalized spline models,...
Conventional regression models
Auto-generated units
Consequences of auto-generation
Arguments pro and con

Gaussian models
Binary regression model
Stratification versus conditioning
Properties of conventional 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)
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

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

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

Computational problem:

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

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,...
But ... $p_x(y)$ is not the correct distribution!

Why not?

What is the correct distribution?

Good news and bad news ...
Exchangeable sequences and conditional distributions

Let \((Y_1, X_1), (Y_2, X_2), \ldots\) be an exchangeable sequence
\(n\)-dimensional joint distributions \(p_n(y_1, \ldots, y_n, x_1, \ldots, x_n)\)

Conditional distribution: \(p_n(y \mid x) = \frac{p_n(y, x)}{p_n(x)}\)
\(p_1(y \mid x) = \Pr(Y_u = y \mid X_u = x)\) (fixed \(u\) and random \(x\))

Strata distributions:
fix \(x \in \mathcal{X}\) and let \(u\) be the first unit such that \(X_u = x\).
Let \(f_x(\cdot)\) be the distribution of \(Y_u\)
— for fixed stratum \(x\) and random unit \(u\)

Under what conditions is \(f_x(y)\) equal to \(p_1(y \mid x)\)?
iid yes, otherwise not usually.
Consequences of GLMM: attenuation

\[
\logit \Pr(Y(u) = 1 \mid \eta) = \alpha + \beta x(u) + \eta(x(u))
\]

Approximate one-dimensional marginal distribution

\[
\logit \Pr(Y(u) = 1) = \alpha^* + \beta^* x(u)
\]

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

Subject-specific approach versus population-average approach

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

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

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 $E(Y(u') \mid data)$

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

Auto-generated units

Consequences of auto-generation

Arguments pro and con

Point process model

Point process model for auto-generated units

![Diagram of point process model](image)

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 \( \lambda(.) \).

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

Peter McCullagh
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 calculation (Bayesian and frequentist):

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

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

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

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

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

Sampling bias:
Distn for fixed \(x\) versus distn for autogenerated \(x\).
Stratification vs. conditioning

Stratification: waiting for Godot!

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

$$\text{pr}(Y = r \mid \lambda; x) = \frac{\lambda_r(x)}{\lambda_.(x)}$$

$$\rho_r(x) = \text{pr}(Y = r; x) = \mathbb{E}\left(\frac{\lambda_r(x)}{\lambda_.(x)}\right)$$

Conventional GLMM calculation

$\rho_r(x)$ is the response distribution in stratum $x$
not a conditional distribution ($x$ not a random variable)
mathematically correct, but seldom relevant...
Stratification vs. conditioning

Conditional distribution: 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 $E(\lambda(.)|x)) = m(.)$

Conditional distribution

$$
\pi_r(x) = pr(Y = r | x) = \frac{E\lambda_r(x)}{E\lambda(.)(x)} \neq E\left(\frac{\lambda_r(x)}{\lambda(.)(x)}\right) = \rho_r(x)
$$
Log Gaussian illustration of sampling bias

\[
\eta_0(x) \sim \mathcal{GP}(\nu_0(x), K), \quad \lambda_0(x) = \exp(\eta_0(x)) \\
\eta_1(x) \sim \mathcal{GP}(\nu_1(x), K), \quad \lambda_1(x) = \exp(\eta_1(x)) \\
\eta(x) = \eta_1(x) - \eta_0(x) \sim \mathcal{GP}((\nu_0 - \nu_1)(x), 2K), \quad K(x, x) = \sigma^2
\]

One-dimensional distributions \((\nu_0(x) - \nu_1(x) = \alpha + \beta x)\):

\[
\rho(x) = \Pr(Y = 1; x) = E\left(\frac{e^{\eta(x)}}{1 + e^{\eta(x)}}\right) \quad \text{(stratum } x) \\
\logit(\rho(x)) \simeq \alpha^* + \beta^* x \quad (|\beta^*| < |\beta|)
\]

\[
\pi(x) = \Pr(Y = 1 \mid x \in \text{SPP}) = \frac{E\lambda_1(x)}{E\lambda_0(x)} = \frac{e^{\alpha + \beta x + \sigma^2/2}}{e^{\sigma^2/2} + e^{\alpha + \beta x + \sigma^2/2}}
\]

\[
\logit \Pr(Y = 1 \mid x \in \text{SPP}) = \alpha + \beta x
\]
Conventional regression models

Auto-generated units

Consequences of auto-generation

Arguments pro and con

Sampling bias

Non-attenuation

Inconsistency of estimates

Estimating functions

Robustness

Interference

Attenuation

Quota sampling:

Conventional calculation for fixed subject $u$

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

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 $1/3$.

Calculation is correct for quota samples ($x$ fixed)

Both probabilities specific to unit $u$

No averaging over units $u \in \mathcal{U}$

Nevertheless $\beta$ is called the subject-specific effect

$\beta^*$ is called population averaged effect

Peter McCullagh

Auto-generated units
Quota sampling:
Conventional calculation for fixed subject $u$

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

implies marginally after integration

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

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 $u \in U$

Nevertheless $\beta$ is called the subject-specific effect
$\beta^*$ is called population averaged effect
Attenuation

Quota sampling:
Conventional calculation for fixed subject $u$

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

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.

Calculation is correct for quota samples ($x$ fixed)
Both probabilities specific to unit $u$
No averaging over units $u \in U$

Nevertheless $\beta$ is called the subject-specific effect
$\beta^*$ is called population averaged effect
Quota sampling:
Conventional calculation for fixed subject $u$

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

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.

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

Sequential sampling for auto-generated units

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

implies marginally after integration

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

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

\[ \logit \Pr(Y(x) = 1 \mid \lambda, \text{event at } x) = \alpha + \beta x + \eta(x) \]

implies marginally after integration

\[ \logit \Pr(Y(x) = 1 \mid x \text{ in superposition}) = \alpha + \beta x \]

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

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

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

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

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

implies marginally after integration

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

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 \( \mathbf{x} \):

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

‘Correct’ likelihood for auto-generated \( \mathbf{x} \)

\[
\rho(\mathbf{y} | \mathbf{x}) = \frac{E \prod \lambda_{y_j}(x_j)}{E \prod \lambda.(x_j)}
\]

If conventional likelihood is used with autogenerated \( \mathbf{x} \)

parameter estimates based on \( \rho_{\mathbf{x}}(\mathbf{y}) \) are inconsistent

bias is approximately \( 1/\tau > 1 \)
Consequences: estimating functions

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

$\pi(x) = m_1(x)/m.(x)$;  $\rho(x) = E(\lambda_1(x)/\lambda.(x))$

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

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

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

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

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

has zero mean for auto-generated $x$. 
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 \( \pi(x) = E(Y | x \in SPP) \) because conventional notation \( E(Y | x) \) is ambiguous

PA is remarkably robust but does not consistently estimate the variance
variance calculation: binary case

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

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

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

\[ \text{var}(T) = \int_{\mathcal{X}} h^2(x)\pi(x)(1 - \pi(x)) m.(x) \, dx \]
\[ + \int_{\mathcal{X}^2} h(x)h(x') \, V(x, x') \, m..(x, x') \, dx \, dx' \]
\[ + \int_{\mathcal{X}^2} h(x)h(x') \Delta^2(x, x')m..(x, x') \, dx \, dx' \]

\(V\): spatial or within-cluster correlation;
\(\Delta\): interference
Conventional regression models
Auto-generated units
Consequences of auto-generation
Arguments pro and con

What is interference?

Physical interference:
\[ \text{distribution of } Y(u) \text{ depends on } x(u') \]

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: \[ \pi_r(x) = m_r(x)/m.(x) \]

Bivariate: \[ \pi_{rs}(x, x') = m_{rs}(x, x')/m.(x, x') \]
\[ \pi_{rs}(x, x') = \Pr(Y(x) = r, Y(x') = s \mid x, x' \in \text{SPP}) \]

Hence \[ \pi_{r.}(x, x') = \Pr(Y(x) = r \mid x, x' \in \text{SPP}) \]
\[ \Delta_r(x, x') = \pi_{r.}(x, x') - \pi_r(x) \]
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?

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

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?
Conventional regression models
Auto-generated units
Consequences of auto-generation
Arguments pro and con

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
Conventional regression models
Auto-generated units
Consequences of auto-generation
Arguments pro and con

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