

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, \dots subjects, plots, ...
Covariate $x(u_1), x(u_2), \dots$ (non-random, vector-valued)
Response $Y(u_1), Y(u_2), \dots$ (random, real-valued)

Regression model:

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

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

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

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



Binary regression model: computation

Computational problem:

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

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)

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

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 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, \dots, 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 $E(Y(u') \mid \text{data})$

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



Point process model for auto-generated units

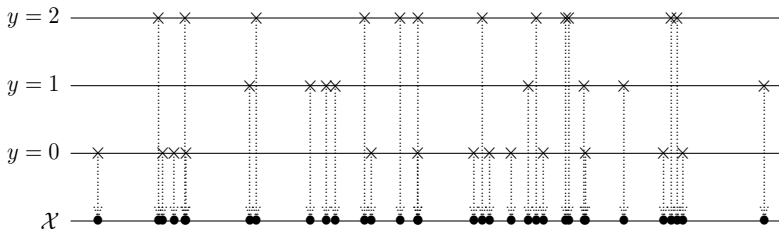


Figure 1: A point process on $\mathcal{C} \times \mathcal{X}$ for $k = 3$, and the superposition process on \mathcal{X} .

Intensity $\lambda_r(x)$ for class r

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

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):

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

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

Calculation is correct in a sense, but irrelevant. . .

. . . there might not be an event at x !



Correct calculation for auto-generated units

$$\text{pr}(\text{event of type } r \text{ in } dx \mid \lambda) = \lambda_r(x) dx + o(dx)$$

$$\text{pr}(\text{event of type } r \text{ in } dx) = E(\lambda_r(x)) dx + o(dx)$$

$$\text{pr}(\text{event in SPP in } dx \mid \lambda) = \lambda_{\cdot}(x) dx + o(dx)$$

$$\text{pr}(\text{event in SPP in } dx) = E(\lambda_{\cdot}(x)) dx + o(dx)$$

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

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

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

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

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 $E(\lambda_{\cdot}(x)) = m_{\cdot}(x)$

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



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:

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



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 \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 autogenerated \mathbf{x}

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



Attenuation

Quota sampling:

Conventional calculation for fixed subject u

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

implies marginally after integration

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

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

Calculation is correct for quota samples (\mathbf{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

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

implies marginally after integration

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

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

Calculation is correct for quota samples (\mathbf{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

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

implies marginally after integration

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

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

Calculation is correct for quota samples (\mathbf{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

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

implies marginally after integration

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

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

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

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

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

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

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

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

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

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



Consequences: inconsistency

Conventional Bayesian likelihood for predetermined \mathbf{x} :

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

'Correct' likelihood for auto-generated \mathbf{x}

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

If conventional likelihood is used with autogenerated \mathbf{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 : $m_r(x) = E(\lambda_r(x))$
 $\pi(x) = m_1(x)/m_{\cdot}(x)$; $\rho(x) = E(\lambda_1(x)/\lambda_{\cdot}(x))$

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

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

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

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

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

has zero mean for auto-generated \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 $\pi(x) = E(Y | x \in \text{SPP})$
because conventional notation $E(Y | x)$ 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;

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

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

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

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

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

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

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

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

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?

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

