# ML estimation: Random-intercepts logistic model

$$\log\left[\frac{p_{ij}}{1-p_{ij}}\right] = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \upsilon_i \qquad \text{with } \upsilon_i \sim N(0, \sigma_\upsilon^2)$$

Standardizing the random effect, $\theta_i = \upsilon_i/\sigma_\upsilon$, yields

$$\log\left[\frac{p_{ij}}{1-p_{ij}}\right] = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \sigma_\upsilon\theta_i \qquad \text{with } \theta_i \sim N(0, 1)$$

Conditional probability of a positive response is:

$$p_{ij} \;=\; P(Y_{ij} \;=\; 1 \mid \theta_i) = \Psi(\mathrm{z}_{ij})$$

where the standard logistic cdf is given as

$$\Psi(\mathrm{z}_{ij}) = \frac{1}{1+\exp(-\mathrm{z}_{ij})} \qquad \text{and } \mathrm{z}_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \sigma_\upsilon\theta_i$$

- observations within a subject are assumed independent given the random subject effect (conditional independence)
- Thus, we can multiply the conditional probabilities across the $n_i$ timepoints within a subject together to yield the conditional probability for the $n_i \times 1$ response vector $\boldsymbol{Y}_i$

$$\ell(\boldsymbol{Y}_i \mid \theta) = \prod_{j=1}^{n_i} \Psi(\mathrm{z}_{ij})^{Y_{ij}} [1 - \Psi(\mathrm{z}_{ij})]^{1-Y_{ij}}$$

the marginal probability for $\boldsymbol{Y}_i$ in the population of subjects is

$$h(\boldsymbol{Y}_i) = \int_\theta \ell(\boldsymbol{Y}_i \mid \theta)\, g(\theta)\, d\theta$$

where $g(\theta)$ represents the population distribution of the (standardized) random effects, namely, $N(0, 1)$

Comments on the marginal probability $h(\boldsymbol{Y}_i)$

- obtained by considering the conditional likelihood, which depends on the random effect, for all possible values of the random effect, thereby yielding an aggregated or marginal likelihood

- akin to a weighted average probability, the values of $\theta$ modify the response function $\mathrm{z}_{ij}$ and thereby modify the conditional likelihood $\ell(\boldsymbol{Y}_i \mid \theta)$, which is weighted by the probability at that point in the distribution $g(\theta)$ as one goes over all values of $\theta$

Marginal likelihood of the response patterns $\boldsymbol{Y}_i$ from all subjects (since subjecs are independent of each other):

$$L = \prod_{i=1}^{N} h(\boldsymbol{Y}_i)$$

or taking logs,

$$\log L = \sum_{i=1}^{N} \log h(\boldsymbol{Y}_i)$$

Let $\boldsymbol{\eta}$ represent either $\boldsymbol{\beta}$ or $\sigma_{\upsilon}$, then taking derivatives

$$\frac{\partial \log L}{\partial \boldsymbol{\eta}} = \sum_{i=1}^{N} h^{-1}(\boldsymbol{Y}_i) \frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\eta}}$$

express the marginal likelihood $h(\boldsymbol{Y}_i)$ in the following way:

$$= \int_\theta \ell(\boldsymbol{Y}_i \mid \theta)\, g(\theta)\, d\theta$$

$$= \int_\theta \left( \prod_{j=1}^{n_i} \Psi(\mathrm{z}_{ij})^{Y_{ij}} [1 - \Psi(\mathrm{z}_{ij})]^{1-Y_{ij}} \right) g(\theta)\, d\theta$$

$$= \int_\theta \left[ \exp \left( \log \left\{ \prod_{j=1}^{n_i} \Psi(\mathrm{z}_{ij})^{Y_{ij}} [1 - \Psi(\mathrm{z}_{ij})]^{1-Y_{ij}} \right\} \right) \right] g(\theta)\, d\theta$$

$$= \int_\theta \left[ \exp \left( \sum_{j=1}^{n_i} Y_{ij} \log[\Psi(\mathrm{z}_{ij})] + (1 - Y_{ij}) \log[1 - \Psi(\mathrm{z}_{ij})] \right) \right] g(\theta)\, d\theta$$

notice that the expression in the large ( ) is precisely the form of the likelihood in ordinary (fixed-effects) logistic regression

remember also that $\partial \exp f(x)/\partial x \;=\; f(x) \partial f(x)/\partial x$

Denoting $\ell(\boldsymbol{Y}_i \mid \theta)$ by $\ell_i$, we get

$$\frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\eta}} = \int_\theta \sum_{j=1}^{n_i} \left[ \frac{Y_{ij}}{\Psi(\mathrm{z}_{ij})} \partial\Psi(\mathrm{z}_{ij}) + \frac{1 - Y_{ij}}{1 - \Psi(\mathrm{z}_{ij})} (-\partial\Psi(\mathrm{z}_{ij})) \right] \frac{\partial \mathrm{z}_{ij}}{\partial \boldsymbol{\eta}} \, \ell_i \, g(\theta) \, d\theta$$

$$= \int_\theta \sum_{j=1}^{n_i} \frac{Y_{ij} - \Psi(\mathrm{z}_{ij})}{\Psi(\mathrm{z}_{ij})(1 - \Psi(\mathrm{z}_{ij}))} \, \partial\Psi(\mathrm{z}_{ij}) \, \frac{\partial \mathrm{z}_{ij}}{\partial \boldsymbol{\eta}} \, \ell_i \, g(\theta) \, d\theta$$

and since $\partial\Psi(\mathrm{z}_{ij})$ equals the pdf, which for the logistic is $\Psi(\mathrm{z}_{ij})[1 - \Psi(\mathrm{z}_{ij})]$,

$$\frac{\partial \log L}{\partial \boldsymbol{\eta}} = \sum_{i=1}^{N} h^{-1}(\boldsymbol{Y}_i) \int_\theta \sum_{j=1}^{n_i} Y_{ij} - \Psi(\mathrm{z}_{ij}) \, \frac{\partial \mathrm{z}_{ij}}{\partial \boldsymbol{\eta}} \, \ell_i \, g(\theta) \, d\theta$$

with

$$\frac{\partial \mathrm{z}_{ij}}{\partial \boldsymbol{\beta}} = \boldsymbol{x}'_{ij} \qquad \frac{\partial \mathrm{z}_{ij}}{\partial \sigma_\upsilon} = \theta_i$$

## Fisher's method of scoring

provisional estimates for the vector of all parameters $\boldsymbol{\Theta}$, on iteration $\iota$ are improved by

$$\boldsymbol{\Theta}_{\iota+1} = \boldsymbol{\Theta}_{\iota} - \left\{ E \left[ \frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_{\iota} \, \partial \boldsymbol{\Theta}'_{\iota}} \right] \right\}^{-1} \frac{\partial \log L}{\partial \boldsymbol{\Theta}_{\iota}}$$

where, the information matrix, or minus the expectation of the matrix of second derivatives, is given by

$$-E \left[ \frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_{\iota} \, \partial \boldsymbol{\Theta}'_{\iota}} \right] = E \left[ \sum_{i=1}^{N} h^{-2}(\boldsymbol{Y}_i) \, \frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\Theta}_{\iota}} \left( \frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\Theta}_{\iota}} \right)' \right]$$

right-hand side sometimes called "outer product of the gradients"

in econometrics, often referred to as the BHHH method

## Numerical Quadrature for integration over $\theta$

- method to numerically perform an integration

$$\int_\theta f(\theta)g(\theta)d\theta \approx \sum_{q=1}^{Q} f(B_q)A(B_q)$$

where $B_q$ $(q = 1, \ldots, Q)$ are the quadrature nodes or points
$A(B_q)$ $(q = 1, \ldots, Q)$ are the weights (sum = 1)

- the more points you use, the more accurate the approximation, but the more time it takes

- For standard normal distribution, Gauss-Hermite quadrature

- does yield a likelihood value that can be used for LR tests

# Gauss-Hermite quadrature points and weights

Number of Quad Points = 3

| | | | |
|---|---|---|---|
| Quad Points = | -1.73205080756888 | 0.00000000000000 | 1.73205080756888 |
| Quad Weights= | 0.16666666872343 | 0.66666668665052 | 0.16666666872343 |

Number of Quad Points = 4

| | | | | |
|---|---|---|---|---|
| Quad Points = | -2.33441421833898 | -0.74196378430273 | 0.74196378430273 | 2.33441421833898 |
| Quad Weights= | 0.04587585477426 | 0.45412415611336 | 0.45412415611336 | 0.04587585477426 |

. . . . . . . . .

. . . . . . . . .

Number of Quad Points = 10

| | | | | |
|---|---|---|---|---|
| Quad Points = | -4.85946282833231 | -3.58182348355193 | -2.48432584163895 | -1.46598909439116 |
| | -0.48493570751550 | 0.48493570751550 | 1.46598909439116 | 2.48432584163895 |
| | 3.58182348355193 | 4.85946282833231 | | |
| Quad Weights= | 0.00000431065265 | 0.00075807095698 | 0.01911158107317 | 0.13548370704150 |
| | 0.34464234526294 | 0.34464234526294 | 0.13548370704150 | 0.01911158107317 |
| | 0.00075807095698 | 0.00000431065265 | | |

Using the quadrature points and weights, the response model is

$$\mathrm{z}_{ijq} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \sigma_{\upsilon} B_q$$

and so the conditional likelihood is

$$\ell(\boldsymbol{Y}_i \mid B_q) = \prod_{j=1}^{n_i} \Psi(\mathrm{z}_{ijq})^{Y_{ij}}[1 - \Psi(\mathrm{z}_{ijq})]^{1-Y_{ij}}$$

yielding the approximated marginal likelihood as

$$h(\boldsymbol{Y}_i) \approx \sum_{q=1}^{Q} \ell(\boldsymbol{Y}_i \mid B_q)\, A(B_q)$$

The first derivatives are then

$$\frac{\partial \log L}{\partial \boldsymbol{\eta}} \approx \sum_{i=1}^{N} h^{-1}(\boldsymbol{Y}_i) \sum_{q=1}^{Q} \sum_{j=1}^{n_i} Y_{ij} - \Psi(\mathrm{z}_{ijq}) \frac{\partial \mathrm{z}_{ijq}}{\partial \boldsymbol{\eta}} \ell(\boldsymbol{Y}_i \mid B_q) A(B_q)$$

$$\text{with} \quad \frac{\partial \mathrm{z}_{ijq}}{\partial \boldsymbol{\beta}} = \boldsymbol{x}'_{ij} \qquad \frac{\partial \mathrm{z}_{ijq}}{\partial \sigma_{\upsilon}} = B_q$$

**Empirical Bayes estimates**

$$\hat{\theta}_i = E(\theta_i \mid \boldsymbol{Y}_i) \; = \; h^{-1}(\boldsymbol{Y}_i) \int_\theta \theta_i \, \ell(\boldsymbol{Y}_i \mid \theta) \, g(\theta) \, d\theta$$

$$\approx \; h^{-1}(\boldsymbol{Y}_i) \sum_{q=1}^{Q} B_q \, \ell(\boldsymbol{Y}_i \mid B_q) \, A(B_q)$$

The variance of the empirical Bayes estimator

$$V(\hat{\theta}_i \mid \boldsymbol{Y}_i) \; = \; h^{-1}(\boldsymbol{Y}_i) \int_\theta (\theta_i - \hat{\theta}_i)^2 \, \ell(\boldsymbol{Y}_i \mid \theta) \, g(\theta) \, d\theta$$

$$\approx \; h^{-1}(\boldsymbol{Y}_i) \sum_{q=1}^{Q} (B_q - \hat{\theta}_i) \, \ell(\boldsymbol{Y}_i \mid B_q) \, A(B_q)$$

At convergence, one more round of quadrature and the converged values of

- $h(\boldsymbol{Y}_i)$ which vary by subjects
- $\ell(\boldsymbol{Y}_i \mid B_q)$ which vary by subjects and quadrature points

## Adaptive Quadrature

- adapt quadrature points and weights for each subject, and at each iteration, using EB estimates of their location $\hat{\theta}_i$ and uncertainty $s_i^2 = V(\hat{\theta}_i \mid \boldsymbol{Y}_i)$
- requires fewer points to obtain accurate solution
- especially useful if subject random effects are very spread out (*i.e.*, ICC is high)

Adapted quadrature points and weights, from original points $B_q$ and weights $A_q$, where $\phi(\cdot)$ = normal pdf

$$B_{iq} = \hat{\theta}_i + s_i B_q$$

$$A_{iq} = \sqrt{2\pi}\, s_i \exp(B_q^2/2)\, \phi(B_{iq})\, A_q$$

## Multiple Random Effects

- quadrature solution must integrate over each random effect dimension ($r$ = number of random effects)

$$\boldsymbol{B}_q = (B_{q1}, B_{q2}, \ldots, B_{qr}) = r - \text{dimension quad pt vector}$$

$$A(\boldsymbol{B}_q) = \prod_{h=1}^{r} A(B_{qh}) = \text{product of univariate weights}$$

- curse of dimensionality: $Q^r$ total points, where $Q$ is the number of points per dimension (*e.g.*, $Q = 10$ and $r = 3$ leads to evaluation at 1000 points)

- adaptive quadrature especially useful here, since $Q$ can be lower

## Random-intercepts Logistic Regression Program

*at website: http://tigger.uic.edu/~hedeker/mixbin.sas.txt*

1. form data matrices $\boldsymbol{Y}$ and $\boldsymbol{X}$

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_N \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_N \end{bmatrix}$$

2. get starting values for $\boldsymbol{\beta}$ and $\sigma_\upsilon$
   - fixed-effects estimates for $\boldsymbol{\beta}$
   - set $\sigma_\upsilon$ to some pre-assigned value, say based on ICC guess

$$\sigma_\upsilon = \sqrt{\frac{ICC\,\pi^2/3}{1 - ICC}}$$

3. Go over subjects, quad points, and repeated obs to obtain

$$h(\boldsymbol{Y}_i) \approx \sum_{q=1}^{Q} \ell(\boldsymbol{Y}_i \mid B_q)\, A(B_q)$$

$$\frac{\partial \log L}{\partial \boldsymbol{\eta}} \approx \sum_{i=1}^{N} h^{-1}(\boldsymbol{Y}_i)\, \frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\eta}}$$

$$\boldsymbol{I}(\boldsymbol{\eta}) = -E\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\eta}\, \partial \boldsymbol{\eta}'}\right] = \left[\sum_{i=1}^{N} h^{-2}(\boldsymbol{Y}_i)\, \frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\eta}} \left(\frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\eta}}\right)'\right]$$

Repeat 3 until all elements of

$$[\boldsymbol{I}(\boldsymbol{\eta})]^{-1}\, \frac{\partial \log L}{\partial \boldsymbol{\eta}} < \text{convergence criterion}$$

## Other methods for integration of $\theta$

Methods based on first- or second-order Taylor series expansions

- Marginal quasi-likelihood (MQL) involves expansion around the fixed part of the model
- Penalized or predictive quasi-likelihood (PQL) also includes the random part in its expansion
- Both are available in the SAS PROC GLIMMIX and MLwiN
- fast, but doesn't yield a likelihood for LR tests
- can yield downwardly biased estimates in certain situations (if $N$ and/or $n$ is small, or ICC is high), especially for MQL

Laplace approximation - Raudenbush et. al., (2000)

- a combination of a fully multivariate Taylor series expansion and Laplace approximation
- fast and computationally accurate
- yields a likelihood for LR tests
- available in HLM, though not for all models

Other methods

- Markov Chain Monte Carlo (MCMC) Bayesian approach (in BUGS)
- Maximum Simulated Likelihood (in some STATA programs) in econometric, transportation, political science literatures