Regression Models for Clustered and Longitudinal Data

Introduction to Mixed Logit Models and GEE Logistic Regression

June 28, 2001

Steve Gregorich

Examples of Multilevel/Clustered/Hierarchical Data Structures

Clustered Data

A three-level data structure

Schools, classrooms with schools, students within classrooms.

"Level-1" ~ students within classrooms.
"Level-2" ~ classrooms within schools.
"Level-3" ~ schools.

Example two-level data structures.

sex-partner couples, individuals within couples.

primary sampling units (e.g., area codes), households within PSUs.

Examples of Multilevel/Clustered/Hierarchical Data Structures

Longitudinal Data

A two-level data structure.

Repeated measures "clustered" or "nested" within individuals.

"Level-1" ~ Repeated measures within individuals.
"Level-2" ~ Individuals.

Combinations of Clustered and Longitudinal Data

Schools, students within schools, repeated measures on students.

"Level-1" ~ repeated measures nested within students.
"Level-2" ~ students within schools.
"Level-3" ~ schools.

Examples of Multilevel/Clustered/Hierarchical Data Structures

Notes

Outcome data is measured at level-1

Covariates can be measured at any level

Interactions possible between covariates measured at different levels

Obs. nested w/in higher-level units, not assumed independent

Repeated measures on the same individual not assumed independent

Highest-level units are assumed to be independent

Contrasting Fixed and Mixed Logistic Regression

Plain logistic regression

Fixed effects only

All observations are independent
A single unit of analysis, e.g., the respondent

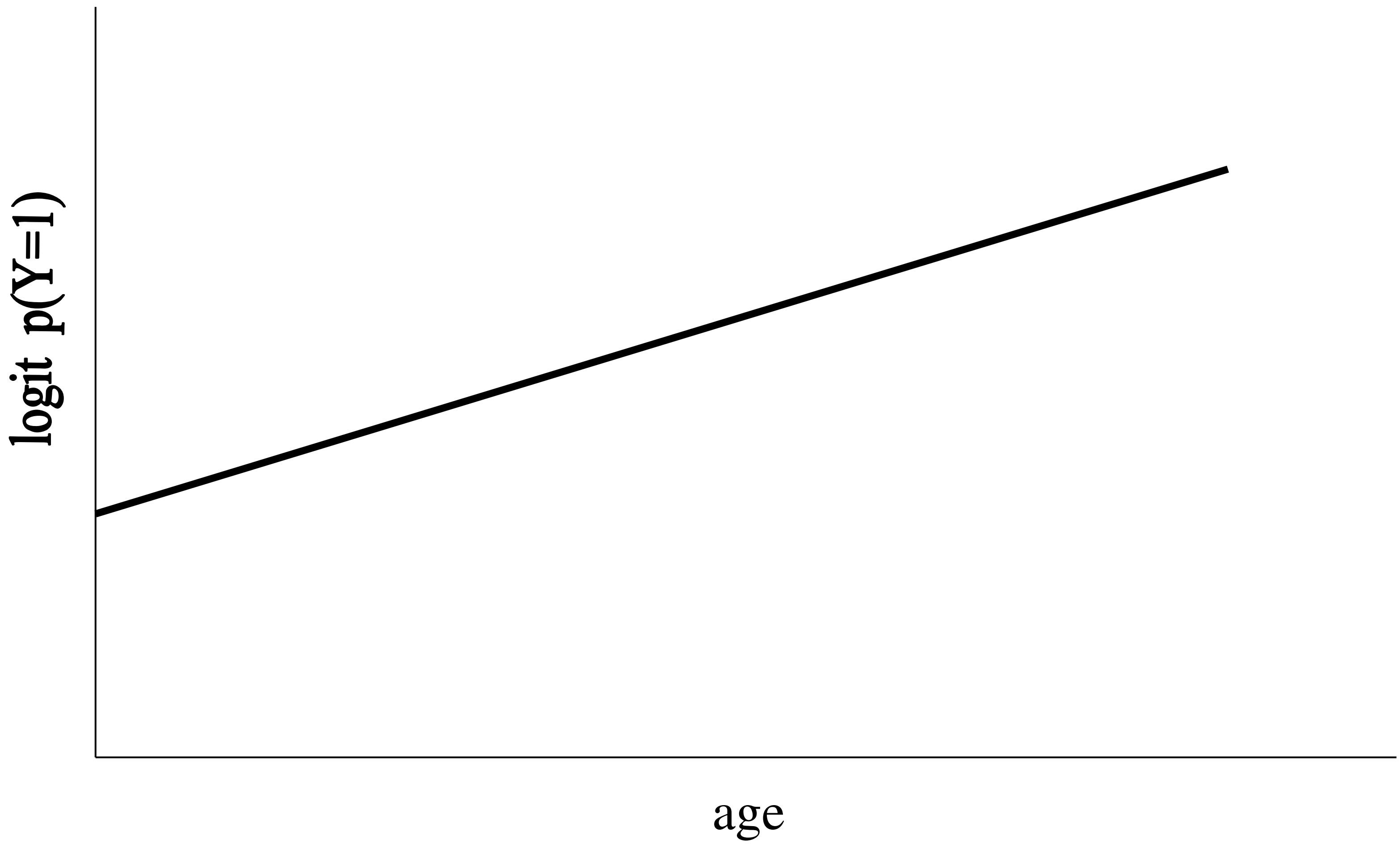
Fixed parameters: marginal, population averaged, unit-generic

Cross-sectional OK, but not clustered or longitudinal data

Contrasting Fixed and Mixed Logistic Regression

Plain logistic regression

Population averaged effects from cross-sectional data



Contrasting Fixed and Mixed Logistic Regression

GEE logistic regression

Fixed effects only

Not all observations are independent

Data can be represented by 2 nested levels

Each level represents a unit of analysis

Clustered sampling *OR* repeated measures

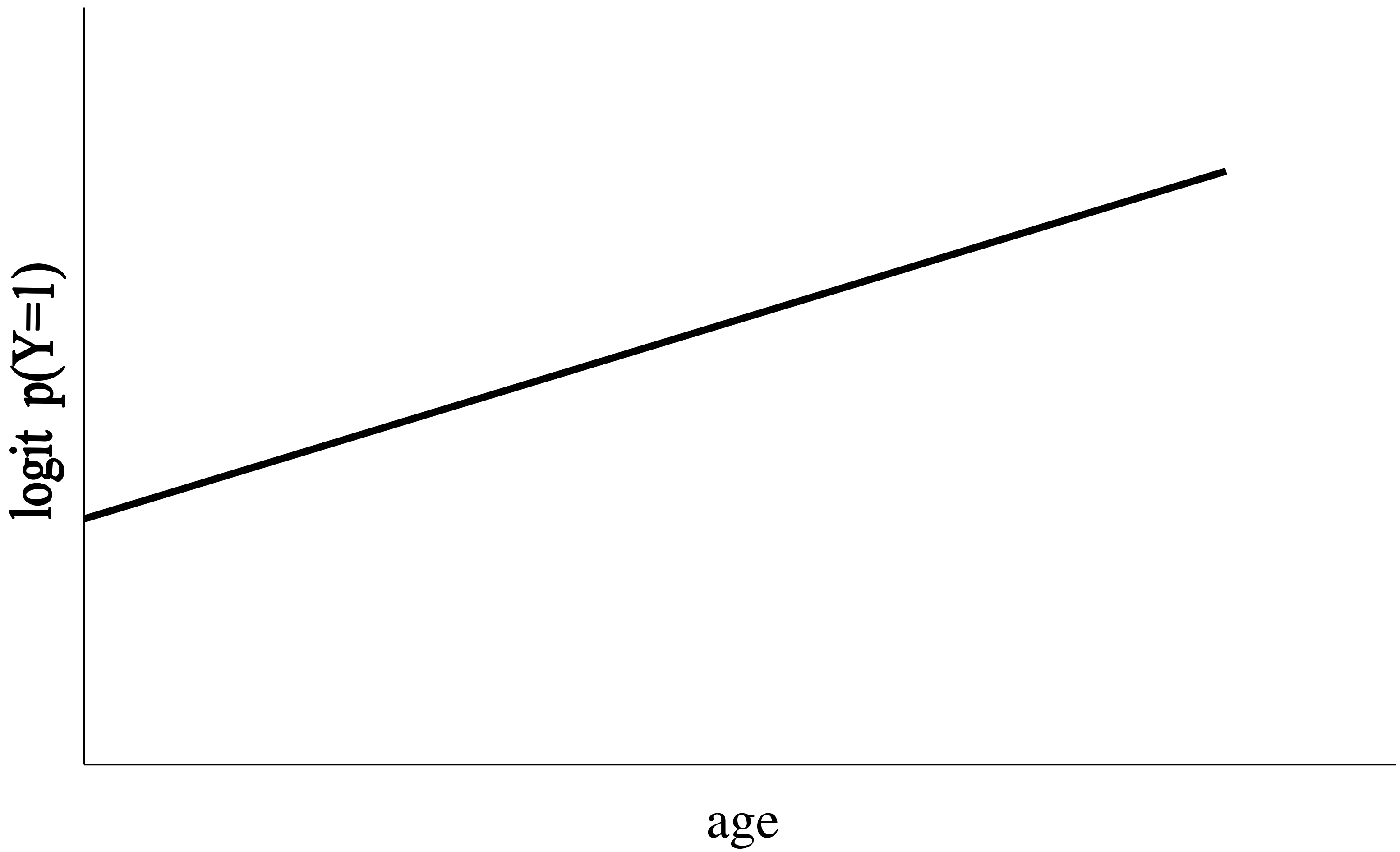
Fixed effects: marginal, population averaged, unit-generic

Non-independence is considered a nuisance

Contrasting Fixed and Mixed Logistic Regression

GEE logistic regression

Population averaged effects from clustered or longitudinal data



Contrasting Fixed and Mixed Logistic Regression

Mixed logit models:

Fixed and random parameters

Fixed parameters: marginal, pop averaged, unit-generic

Random parameters are unit-specific

Not all observations are independent

Data can be represented by 2 *or more* nested levels

Each level represents a unit of analysis

Clustered sampling *AND/OR* repeated measures

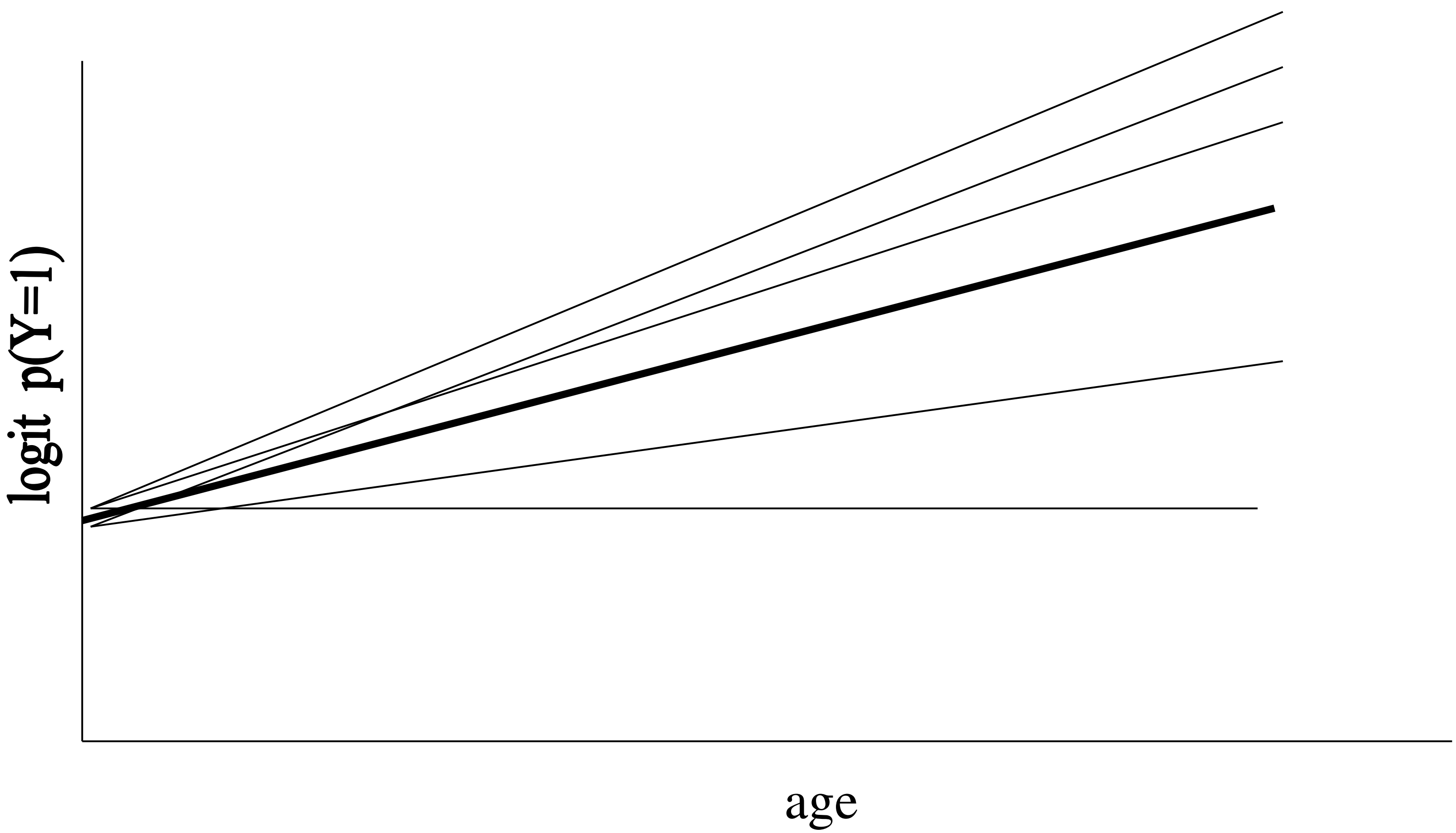
Non-independence is substantively interesting and is modeled

Contrasting Fixed and Mixed Logistic Regression

Mixed logit models (random intercepts)

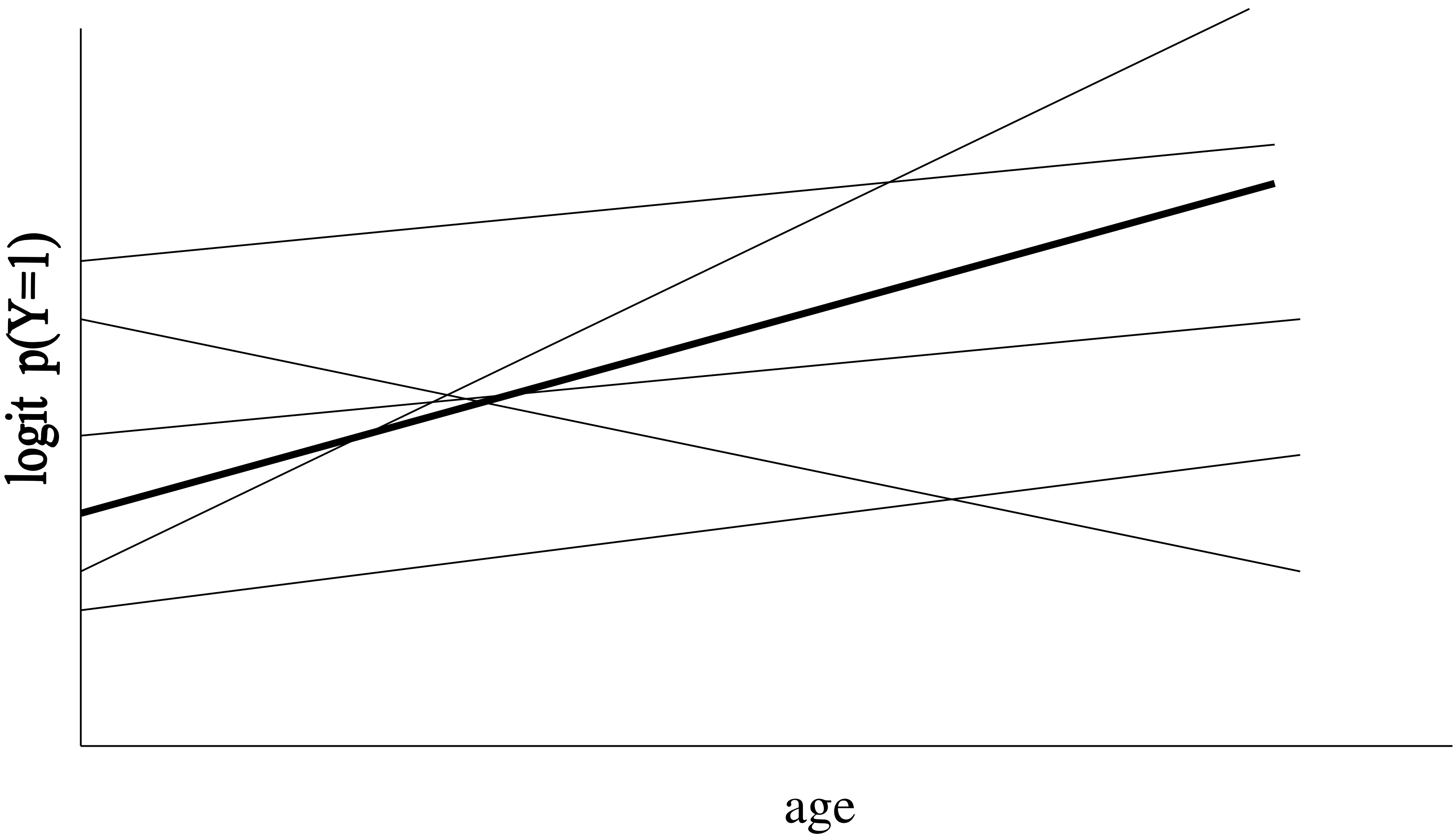
Contrasting Fixed and Mixed Logistic Regression

Mixed logit models (random slopes)



Contrasting Fixed and Mixed Logistic Regression

Mixed logit models (random intercepts and slopes)



Unit-Specific versus Population Averaged Effects

Longitudinal Example

Sample a group of unmarried people and follow them over time

Some become married, some never do

You want to know the impact of marital status on HH expenditures

Population-averaged approach

Assess how the *average* expenditure differed between married and unmarried groups.

No reference to observed individual changes

Unit-specific approach

Assess the *change* in expenditures at the individual level

Benefits of Modeling Non-Independence

GEE and Mixed Models

Correct standard errors

Simultaneously model effects of different units of analysis

e.g., 'contextual' analysis

Mixed Models

Useful when between-unit variation is substantial and/or of interest

Between-unit variation can be explained by additional covariates

Model more than 2 nested levels

More Formally...

Conditional mean of *Y* given X_{ij}

$\mu_{ij} = 1 / 1 + \exp -(B_0 + B_1 X_{ij})$

and

$var(Y_{ij}) = \mu_{ij} (1 - \mu_{ij})$,

$logit(\mu_{ij}) = \ln(\mu_{ij} / (1 - \mu_{ij})$.

GEE: model the population-average, logit (μ_{ij})

$logit\ (\mu_{ij}) = B_0 + B_1 X_{ij}$

$Corr(Y_{ij}, Y_{ik}) = \alpha$

Odds ratios represent the ratios of population odds.

More Formally...

GLMM: model the unit-specific, logit (μ_{ij} | U_{0j})

$$\text{logit}(\mu_{ij} \mid U_j) = B_0 + B_1 X_{ij} + U_{0j}$$

$$\text{Cov}(Y_{ij}, Y_{ik}) = \text{var}(U_{0j})$$

Odds ratios represent the ratios of individual odds.

Y_{ij} are independent, conditional in U_{0j}

Example 1: Variance Components Model

The Level-1 Model

$$\text{logit}(\mu_{ij} \mid U_{0j}) = B_{0j} + B_1 X_{ij}$$

The Level-2 Model:

$$B_{0j} = B_0 + U_{0j}$$

The Combined Model:

$$\text{logit}(\mu_{ij} \mid U_{0j}) = B_0 + B_1 X_{ij} + U_{0j}$$

$$\text{Cov}(U_{0j}, e_{ij}) = 0$$

Example 2: Random Coefficients Model

The Level-1 Model:

$$\text{logit}(\mu_{ij} \mid U_{0j}) = B_{0j} + B_{1j}X_{ij}$$

The Level-2 Model:

$$B_{0j} = B_0 + U_{0j}$$

$$B_{1j} = B_1 + U_{1j}$$

The Combined Model:

$$\text{logit}(\mu_{ij} \mid U_{0j}) = B_0 + B_1X_{ij} + U_{0j} + U_{1j}X_{ij}$$

Further extensions are possible

Estimation Procedures for GLMMs

Approximate quasi-likelihood

1st- and 2nd-order MQL and PQL

MLwiN, GLMMIX.SAS, HLM

Advantages

Fast execution.

Flexible model specification

Disadvantages

Biased parameter estimates can result when variance components are large.

Estimation Procedures for GLMMs

Gaussian quadrature

Allows numerical integration for 2-level models

MIXOR and PROC NLMIXED

Advantages

Fast execution

Unbiased parameter estimates, correct standard errors.

Disadvantages

Limitations on the number of nested levels

Limitations on the number of random effects

Estimation Procedures for GLMMs

Iterated Bootstrap Bias Correction

Based upon MQL or PQL

MLwiN macros

Advantages

Unbiased parameter estimates

Flexible model specification.

Disadvantages

Computationally intensive

Desired degree of convergence may be difficult to obtain

Estimated standard errors may be questionable

Software can be unstable

Estimation Procedures for GLMMs

MCMC methods—Gibbs sampling

BUGS and MLwiN

Advantages

Unbiased parameter estimates, correct standard errors.

Flexible model specification.

Disadvantages

Judging convergence can be tricky

Computationally intensive.

Ozone Data

71 subjects, each received two doses of ozone exposure

Explanatory variable

Dose = level of ozone exposure (1=High 0=Low)

Outcome

Y = observed respiratory symptoms (1=Yes   0 = No)

Variance component model

$$\text{logit}(\mu_{ij} \mid U_{0j}) = B_{0j} + B_1 DOSE_{ij}$$

$$B_{0j} = B_0 + U_{0j}$$

$$\text{logit}(\mu_{ij} \mid U_{0j}) = B_0 + B_1 DOSE_{ij} + U_{0j}$$

Ozone Data

id	dose	y
1	0	1
1	1	1
2	0	1
2	1	1
3	0	1
3	1	1
....		
70	0	0
70	1	0
71	0	0
71	1	0

Results from different estimation methods*

	1st Order MQL	1st Order PQL	2nd Order PQL	NL-MIXED	IBBC	GEE†
B_0	-1.40 (0.32)	-1.53 (0.34)	-2.24 (0.51)	-2.68 (0.79)	-2.61 (0.58)	-1.40 (0.30)
B_1	0.86 (0.39)	0.94 (0.41)	1.42 (0.52)	1.61 (0.63)	1.56 (0.55)	0.86 (0.29)
σ^2_u	1.14	1.33	5.01	6.85	6.67	n/a

* MCMC did not converge

† Parameters are population averaged, not unit-specific, but compare to MQL.

PROC NLMIXED Syntax for a Mixed Logit Model

```
proc nlmixed method=gauss;
    eta       = beta0 + beta1*dose + u;
    expeta    = exp(eta);
    p         = expeta/(1+expeta);
    model   y ~ binomial(1,p);
    random u ~ normal(0,s2u) subject=id;
```

```
notes.  Data must be sorted by subject ID.
        Only two-level models are possible.
        Multiple random effects are possible.
        Large models and large N, a problem.
```

PROC GENMOD Syntax for a GEE Logistic Regression Model

```
proc genmod descending;
  class id;
  model y = dose /dist=bin;
  repeated subject=id / type=un corrw;


notes.  Only 2-level models are possible.
        Dependencies treated as nuisances.
        Large models & large N less of a problem
        Many different working corr structures
        CLASS, CONTRAST, ESTIMATE statements
        Type III statistics available
```

Software Links

Information

multilevel models project

http://www.ioe.ac.uk/multilevel/

multilevel listserv

http://www.jiscmail.ac.uk/lists/multilevel.html

Harvey Goldstein's papers and free book

http://www.ioe.ac.uk/hgpersonal/papers_for_downloading.htm#SectionA

JJ Hox's free book

http://www.ioe.ac.uk/multilevel/amaboek.pdf

Software Links

Free Software

MIXOR (Gaussian Quadrature)
http://tigger.uic.edu/~hedeker/mix.html

BUGS (MCMC)
http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

MAREG (for Population-Averaged Models)
http://www.stat.uni-muenchen.de/~andreas/mareg/winmareg.html

Software Links

Commercial Software

PROC NLMIXED
http://www.sas.com/rnd/app/papers/nlmixedsugi.pdf

GLMM800.SAS macro (1st-order MQL and PQL)
http://ewe3.sas.com/techsup/download/stat/glmm800.sas

MLwiN (MQL, PQL, IBBC, MCMC).
http://multilevel.ioe.ac.uk/index.html

HLM (PQL, and a Gaussian-Quadrature-like approach)
http://www.ssicentral.com/hlm/hlm.htm

GLLAMM6 (ML estimation, requires Stata)
http://www.iop.kcl.ac.uk/iop/departments/biocomp/programs/gllamm.html