BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 17

Generalized Linear Mixed Models

Generalized Linear Mixed Models

Previous lecture discussed *marginal models* for longitudinal data.

Next, we consider a second type of model, *generalized linear mixed models* (GLMMs).

We describe how these models extend the conceptual approach represented by the linear mixed effects model already considered for continuous outcomes.

We also highlight their greater degree of conceptual and analytic complexity relative to marginal models.

Generalized Linear Mixed Models

Postulate unobserved latent variables (random effects) shared by the repeated measures on the same subject.

The basic premise is that we assume natural heterogeneity across individuals in a subset of the regression coefficients.

That is, a subset of the regression coefficients (e.g., intercepts and slopes) are assumed to vary across individuals according to some distribution.

Then, conditional on the random effects, it is assumed that the responses for a single individual are independent observations from a distribution belonging to the exponential family.

Generalized Linear Mixed Models

The generalized linear mixed model can be considered in two steps:

First Step: Assumes that the conditional distribution of each Y_{ij}, given individual-specific effects b_i, belongs to the exponential family with conditional mean,

$$g(E[Y_{ij}|b_i]) = X'_{ij}\beta + Z'_{ij}b_i$$

where $g(\cdot)$ is a known link function and Z_{ij} is a known design vector, a subset of X_{ij}, linking the random effects b_i to Y_{ij}.

The particular subset of the regression parameters β that vary randomly is determined by components of X_{ij} that comprise Z_{ij}.

Second-Step: The b_i are assumed to vary independently from one individual to another and $b_i \sim N(0, G)$.

Here, G is the covariance matrix for the random effects.

Note: There is an additional assumption of 'conditional independence'.

That is, given b_i, the responses $Y_{i1}, Y_{i2}, ..., Y_{in_i}$ are assumed to be mutually independent.

Example 1:

Binary logistic model with random intercepts:

$$\text{logit}(E[Y_{ij}|b_i]) = \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} + b_i$$

$Var(Y_{ij}|b_i) = E[Y_{ij}|b_i](1 - E[Y_{ij}|b_i])$ (Bernoulli variance),

and $b_i \sim N(0, \sigma_b^2)$.

Example 2:

Random coefficients (random intercepts and slopes) Poisson regression model:

$$\log(E[Y_{ij}|b_i]) = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij}$$

$Var(Y_{ij}|b_i) = E[Y_{ij}|b_i]$ (Poisson variance),

and $b_i \sim N(0, G)$.

Note: G is the covariance matrix for b_{1i} and b_{2i}.

Recall that marginal models consider the consequences of dependence among the repeated measures on the same subject, via a "working" covariance.

In contrast, GLMMs provide a potential explanation for the sources of dependence among the repeated measures on the same subject, via the introduction of random effects.

However, the introduction of random effects also has important implications for the interpretation of the regression parameters in GLMMs.

Interpretation of Fixed Effects

GLMMs are most useful when the scientific objective is to make inferences about individuals rather than population averages.

Main focus is on the individual and the influence of covariates on a *typical* ($b_i = 0$) individual's responses.

Regression parameters, β, measure the change in expected value of response while holding constant other covariates and the random effects.

For example, consider the following logistic model,

$$\text{logit}(E[Y_{ij}|b_i]) = \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} + b_i$$

with $b_i \sim N(0, \sigma^2)$.

Each element of β measures the change in the log odds of a 'positive' response per unit change in the respective covariate, for an individual with propensity to respond positively, b_i.

The interpretation of any component of β, say β_k, is in terms of changes in a specific *individual's* log odds of response for a unit change in the corresponding covariate, say X_{ijk}.

Note: This is not always directly observable from the data.

When X_{ijk} takes on some value x, the log odds of a positive response is,

$$\log\left[\frac{\Pr(Y_{ij}=1|b_i,X_{ij1},\ldots,X_{ijk}=x,\ldots,X_{ijp})}{\Pr(Y_{ij}=0|b_i,X_{ij1},\ldots,X_{ijk}=x,\ldots,X_{ijp})}\right] =$$

$$b_i + \beta_1 X_{ij1} + \cdots + \beta_k x + \cdots + \beta_p X_{ijp}.$$

Similarly, when X_{ijk} now takes on some value $x+1$,

$$\log\left[\frac{\Pr(Y_{ij}=1|b_i,X_{ij1},\ldots,X_{ijk}=x+1,\ldots,X_{ijp})}{\Pr(Y_{ij}=0|b_i,X_{ij1},\ldots,X_{ijk}=x+1,\ldots,X_{ijp})}\right] =$$

$$b_i + \beta_1 X_{ij1} + \cdots + \beta_k (x+1) + \cdots + \beta_p X_{ijp}.$$

$\longrightarrow$ β_k is change in log odds for individual with propensity to respond, b_i.

This *subject-specific* interpretation of β_k is more appealing when X_{ijk} is a *time-varying* covariate.

That is, when it is possible to hold b_i (and remaining covariates) fixed and also change the value of the covariate, X_{ijk}.

Recall: Time-varying covariate is one whose value can change over time, e.g., time since baseline, smoking status, and environmental exposures.

When X_{ijk} is *time-invariant* the interpretation of β_k is less transparent.

With a time-invariant covariate (e.g., gender), changing the value of the covariate requires also a change in the index i of X_{ijk}, say $X_{i'jk}$.

When X_{ijk} takes on some value x, the log odds of a positive response is,

$$\log\left[\frac{\Pr(Y_{ij}=1|b_i,X_{ij1},\ldots,X_{ijk}=x,\ldots,X_{ijp})}{\Pr(Y_{ij}=0|b_i,X_{ij1},\ldots,X_{ijk}=x,\ldots,X_{ijp})}\right] =$$

$$b_i + \beta_1 X_{ij1} + \cdots + \beta_k x + \cdots + \beta_p X_{ijp}.$$

Similarly, when $X_{i'jk}$ now takes on some value $x+1$,

$$\log\left[\frac{\Pr(Y_{i'j}=1|b_{i'},X_{i'j1},\ldots,X_{i'jk}=x+1,\ldots,X_{i'jp})}{\Pr(Y_{i'j}=0|b_{i'},X_{i'j1},\ldots,X_{i'jk}=x+1,\ldots,X_{i'jp})}\right] =$$

$$b_{i'} + \beta_1 X_{i'j1} + \cdots + \beta_k (x+1) + \cdots + \beta_p X_{i'jp}.$$

Even when we consider two subjects with identical covariates except for the k^{th}, the difference in log odds is

$$\beta_k + (b_i - b_{i'}).$$

That is, β_k has become confounded with $b_i - b_{i'}$.

This dilemma can only be resolved by assuming same value for the unobserved random effects, $b_i = b_{i'}$; however, this contrast is not directly observable.

Estimation

The joint probability density function is given by:

$$f(Y_i|X_i, b_i)f(b_i)$$

Estimation using maximum likelihood (ML) involves two steps:

First, ML estimation of β (and possibly ϕ) and G is based on the marginal or integrated likelihood of the data

$$L(\beta, \phi, G) = \prod_{i=1}^{N} \int f(Y_i|X_i, b_i)f(b_i)db_i$$

obtained by averaging over the distribution of the unobserved random effects, b_i.

However, simple analytic solutions are rarely available.

In general, computations are difficult.

- maximization of the likelihood is iterative
- likelihood evaluation requires many integrations

In general, ML estimation requires numerical or Monte Carlo integration techniques that can be computationally quite intensive.

1. Numerical integration techniques, known as Gaussian quadrature, simply approximate the integral as a weighted sum,

$$L(\beta, \phi, G) \approx \prod_{i=1}^{N} \sum_{k=1}^{K} f(Y_i | b_i = v_k) w_k,$$

where the known quadrature points (the weights, w_k, and the evaluation points, v_k) are chosen to provide an accurate numerical approximation.

The number of quadrature points determines the degree of accuracy of the approximation involved in replacing the integral by a weighted sum.

2. Another approximation is "Penalized Quasi-Likelihood" (PQL) based on a Laplace approximation of the integral is **faster** but **not as accurate**.

In the second step, given ML estimates of β, ϕ and G, the random effects can be predicted as follows,

$$\hat{b}_i = E(b_i | Y_i; \hat{\beta}, \hat{\phi}, \hat{G})$$

(Posterior mean)

Note that $E(b_i | Y_i; \hat{\beta}, \hat{\phi}, \hat{G})$ involves integrating (or averaging) over the distribution of the unobserved random effects, b_i.

However, simple analytic solutions are rarely available and numerical or Monte Carlo integration techniques are also required.

Case Study 1

Oral Treatment of Toenail Infection

Randomized, double-blind, parallel-group, multicenter study of 294 patients comparing 2 oral treatments (denoted A and B) for toenail infection.

Outcome variable: Binary variable indicating presence of onycholysis (separation of the nail plate from the nail bed).

Patients evaluated for degree of onycholysis (separation of the nail plate from the nail-bed) at baseline (week 0) and at weeks 4, 8, 12, 24, 36, and 48.

Interested in the effect of treatment on changes in an individual's risk of onycholysis over time?

Assume that the conditional probability of onycholysis follows a logistic model,

$$\text{logit}\left(E[Y_{ij}|b_i]\right) = \beta_1 + \beta_2 \text{Month}_{ij} + \beta_3 \text{Trt}_i * \text{Month}_{ij} + b_i$$

where $Trt = 1$ if treatment group B and 0 otherwise.

Here, we assume that $\text{Var}(Y_{ij}) = E(Y_{ij}|b_i)\left[1 - E(Y_{ij}|b_i)\right]$.

We also assume $b_i \sim N(0, \sigma_b^2)$.

Table 1: ML estimates and standard errors from random effects logistic regression model for onycholysis data.

PARAMETER	ESTIMATE	SE	Z
INTERCEPT	-1.697	0.330	-5.15
Month	-0.389	0.043	-8.97
Trt × Month	-0.142	0.065	-2.19
σ_b^2	16.034	3.039	5.28

ML based on 100-point adaptive Gaussian quadrature.

Results

From the output above, we would conclude that:

1. There is a significant difference in the rate of decline of risk for individuals in the two treatment groups ($P < 0.05$).

2. Over 12 months, the odds of infection decreases by a factor of 0.01 [or exp(-0.389*12)] for an individual receiving treatment A.

3. Over 12 months, the odds of infection decreases by a factor of 0.002 [exp(-0.531*12)] for an individual receiving treatment B.

4. Odds ratio comparing 12 month decreases in risk between treatments A and B is approx 5.5 (or $e^{12*0.142}$).

5. Estimated variance of the random intercepts, $\widehat{\sigma}_b^2 = 16.03$ is relatively large.

   For example, the estimated variance implies that 95% of patients have a baseline risk of infection between

$$\frac{exp(-1.697 \pm 1.96 \times \sqrt{16.034})}{1 + exp(-1.697 \pm 1.96 \times \sqrt{16.034})}$$

   (or between 0 and 0.997).

   This suggests substantial heterogeneity of risk.

Case Study 2

Clinical trial of anti-epileptic drug progabide

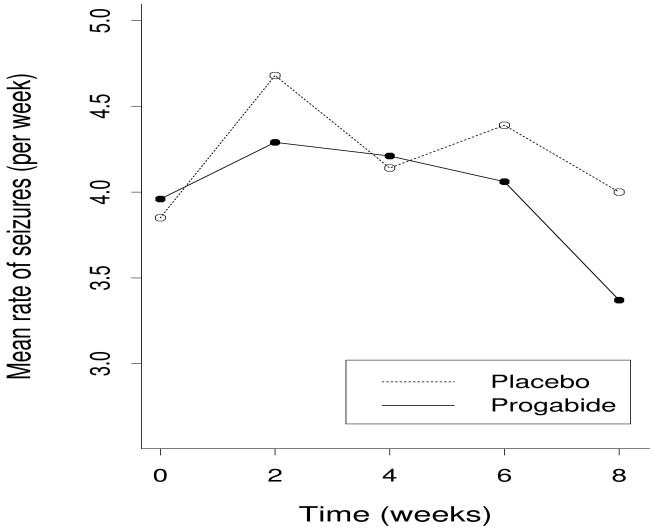
Randomized, placebo-controlled study of treatment of epileptic seizures with progabide.

Patients were randomized to treatment with progabide, or to placebo in addition to standard therapy.

Response variable: Count of number of seizures

Measurement schedule: Baseline measurement during 8 weeks prior to randomization. Four measurements during consecutive two-week intervals.

Interested in the effect of treatment with progabide on changes in an individual's rate of seizures?



Assume conditional rate of seizures follows the mixed effects loglinear model,

$$\log(E[Y_{ij}|b_i]) = \log(t_{ij}) + \beta_1 + b_{1i} + \beta_2 \text{time}_{ij} + b_{2i}\text{time}_{ij} + \beta_3 \text{trt}_i + \beta_4 \text{trt}_i * \text{time}_{ij}$$

where t_{ij} = length of period; $\text{time}_{ij} = 1$ if periods 1-4, 0 if baseline; $\text{trt}_i = 1$ if progabide, 0 if placebo.

(b_{1i}, b_{2i}) are assumed to have a bivariate normal distribution with zero mean and covariance G.

We will compare results from model that assume

1. $\text{Var}(Y_{ij}|b_i) = E[Y_{ij}|b_i]$ (No overdispersion)

2. $\text{Var}(Y_{ij}|b_i) = \phi E[Y_{ij}|b_i]$ (Overdispersion)

Table 2: Subject-specific log expected seizure rates in the two groups at baseline and during post-baseline follow-up.

Treatment Group	Period	$\log\left(\frac{E(Y_{ij}\|b_i)}{\texttt{T}_{\texttt{ij}}}\right)$
Placebo	Baseline	$\beta_1 + b_{1i}$
	Follow-up	$(\beta_1 + b_{1i}) + (\beta_2 + b_{2i})$
Progabide	Baseline	$(\beta_1 + b_{1i}) + \beta_3$
	Follow-up	$(\beta_1 + b_{1i}) + (\beta_2 + b_{2i}) + \beta_3 + \beta_4$

Parameter estimates and standard errors from mixed effects log-linear regression model **without overdispersion** for the seizure data.

Parameter	Estimate	SE	Z
Intercept	1.0707	0.1406	7.62
$\texttt{time}_{\texttt{ij}}$	−0.0004	0.1097	−0.00
$\texttt{trt}_{\texttt{i}}$	0.0513	0.1931	0.27
$\texttt{trt}_{\texttt{i}} \times \texttt{time}_{\texttt{ij}}$	−0.3065	0.1513	−2.03
$\text{Var}(b_{1i})$	0.5010	0.1010	4.96
$\text{Var}(b_{2i})$	0.2334	0.0608	3.84
$\text{Cov}(b_{1i}, b_{2i})$	0.0541	0.0559	0.97

ML based on 50-point adaptive Gaussian quadrature.

Results of the Poisson analysis (no overdispersion) suggests:

1. A patient treated with placebo has the same expected seizure rate before and after randomization $[exp(-0.0004) \approx 1]$.

2. A patient treated with progabide has expected seizure rate reduced after treatment by approximately 26% $[1 - exp(-0.0004 - 0.3065) \approx 0.26]$.

3. Estimated variance of the random intercepts and slopes is relatively large

4. Heterogeneity should not be ignored

Parameter estimates and standard errors from mixed effects log-linear regression model **with overdispersion** for the seizure data.

Parameter	Estimate	SE	Z
Intercept	1.12347	0.1401	8.02
$\texttt{time}_{\texttt{ij}}$	−0.0185	0.1095	−0.17
$\texttt{trt}_{\texttt{i}}$	0.0341	0.1930	0.18
$\texttt{trt}_{\texttt{i}} \times \texttt{time}_{\texttt{ij}}$	−0.2533	0.1529	−1.66
$\text{Var}(b_{1i})$	0.4649	0.1012	4.59
$\text{Var}(b_{2i})$	0.1734	0.0569	3.05
$\text{Cov}(b_{1i}, b_{2i})$	0.0835	0.0559	1.49
ϕ	1.9765	0.2016	9.79

Fit based on Penalized Quasi-Likelihood (PQL)

Results of the overdispersed Poisson analysis suggests:

1. The effect estimates do not change much when we allow for overdispersion:
   (a) A patient treated with placebo has the same expected seizure rate before and after randomization [$exp(-0.0185) \approx 0.98$].

   (b) A patient treated with progabide has expected seizure rate reduced after treatment by approximately 24% [$1 - exp(-0.0185 - 0.2533) \approx 0.24$].

2. Our estimate $\widehat{\phi} = 1.97$ suggests substantial overdispersion.

3. This overdispersed Poisson analysis suggests there is only moderate evidence of a trt * time effect ($p = 0.0994$), as opposed to strong evidence suggested by the Poisson analysis ($p = 0.0284$).

Summary of Key Points

GLMMs extend the conceptual approach represented by the linear mixed effects model.

GLMMs assume natural heterogeneity across individuals in a subset of the regression coefficients.

The focus of GLMMs is on inferences about individuals.

The regression parameters, β, have 'subject-specific' interpretations in terms of changes in the transformed mean response for a specific individual.

GLMMs using PROC GLIMMIX in SAS

PROC GLIMMIX in SAS is a procedure for fitting generalized linear mixed models longitudinal discrete response.

The syntax for PROC GLIMMIX is very similar to that for fitting linear mixed models in PROC MIXED.

PROC GLIMMIX, as with almost all software for longitudinal analyses, requires each repeated measurement in a longitudinal data set to be a separate "record".

If the data set is in a *multivariate* mode (or "wide format"), it must be transformed to a *univariate* mode (or "long format") prior to analysis.

Table 3: Illustrative commands for a generalized linear mixed model for logistic regression with random intercepts, using PROC GLIMMIX in SAS.

```
PROC GLIMMIX METHOD=QUAD(QPOINTS=50);
  CLASS id group;
  MODEL y (descending) =group time group*time /
  DIST=BINARY LINK=LOGIT S;
  RANDOM INTERCEPT / SUBJECT=id;
```

Table 4: Illustrative commands for a generalized linear mixed model for Poisson regression with random intercepts and slopes, with no overdispersion parameter, using PROC GLIMMIX in SAS.

```
PROC GLIMMIX METHOD=QUAD(QPOINTS=50);

  CLASS id group;

  MODEL y=group time group*time / DIST=POISSON LINK=LOG S;

  RANDOM INTERCEPT time / SUBJECT=id TYPE=UN;
```


Table 5: Illustrative commands for a generalized linear mixed model for overdispersed Poisson regression with random intercepts and slopes, using PROC GLIMMIX in SAS. PROC GLIMMIX requires PQL fitting when including overdispersion parameter.

```
PROC GLIMMIX;

  CLASS id group;

  MODEL y=group time group*time / DIST=POISSON LINK=LOG S;

  RANDOM INTERCEPT time / SUBJECT=id TYPE=UN;

  RANDOM _RESIDUAL_;
```