Chapter 2

Generalized Linear Mixed Effects Models

2.1 Modeling with Generalized Linear Mixed Models

2.1.1 Continuous Data Models

– In generalized linear models, the expected value μ of the response variable Y depends on the explanatory variables $X_1, X_2, \ldots, X_p$ through the link function g:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{2.1}$$

– In generalized linear mixed models, the expected value μ of the response variable Y depends on the set of the explanatory variables $X_1, X_2, \ldots, X_p$ and $Z_1, Z_2, \ldots, Z_q$ through the link function g:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + b_1 Z_1 + b_2 Z_2 + \cdots + b_q Z_q. \tag{2.2}$$

– All the random effects $b_1, b_2, \ldots, b_q$ are assumed to follow normal distributions, which implies that marginal distributions for the random effects are

$$b_k \sim N(0, \sigma^2_{z_x}), \quad k = 1, 2, \ldots, q. \tag{2.3}$$

– From the applied point of view, linear mixed models and generalized linear mixed models are very close to each others. Generalized linear mixed model can also be a variance component model or a generalized linear mixed model for repeated measurements.

– For continuous data, possible distributions and generalized linear mixed models are

Normal distribution:

$$\begin{aligned} \mathbf{y}|\mathbf{b} &\sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I}), \\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \\ \mathbf{b} &\sim N(\mathbf{0}, \mathbf{G}), \qquad \text{in glmer with structure } \mathbf{G} = \sigma^2\mathbf{F}. \end{aligned}$$

Gamma distribution:

$$\begin{aligned} \mathbf{y}|\mathbf{b} &\sim Gamma(\boldsymbol{\mu}, \phi), \\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \\ \mathbf{b} &\sim N(\mathbf{0}, \mathbf{G}), \qquad \text{in glmer with structure } \mathbf{G} = \phi\mathbf{F}. \end{aligned}$$

Inverse Gaussian distribution:

$$\begin{aligned} \mathbf{y}|\mathbf{b} &\sim IG(\boldsymbol{\mu}, \phi), \\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \\ \mathbf{b} &\sim N(\mathbf{0}, \mathbf{G}), \qquad \text{in glmer with structure } \mathbf{G} = \phi\mathbf{F}. \end{aligned}$$

– In practice, possible link functions for all of these distributions are identity link $g(\mu_i) = \mu_i$, log link $g(\mu_i) = \log(\mu_i)$, and inverse link $g(\mu_i) = \frac{1}{\mu_i}$

Example 2.1.

Consider the data of the file retinal.txt where it has been studied how the amount of gas C3F8 used in eye surgery remains in eye relative to time when the amount of gas used is varied.

```
    ID Level Time   Gas
1    1   20%    1 0.990
2    1   20%    2 0.950
3    1   20%    3 0.950
.
.
180 31   25%   42 0.125
181 31   25%   49 0.125
```

```
The outcome variable was the gas (Gas) left in the eye.
The gas, with three different concentration levels,
15%, 20% and 25% (Level), was injected into the eye
before surgery for 31 patients. They were then followed
three to eight (average of 5) times over a three-month
period, and the volume of gas in the eye at the follow-up
times were recorded.
```

Denote the variables as following $Y = \mathsf{Gas}, X_1 = T = \mathsf{time}, X_2 = \mathsf{Level}$. Additionally, the variable ID identifies the sampling units which are repeatedly measured at different time points with respect to the variable Y.

Consider the generalized mixed effects model with interaction

$$\mathcal{M}: \quad g(\mu_{it}) = \beta_0 + \beta_1 t_i + \alpha_j + \gamma_j t_i + b_{i0} + b_{i1} t_i,$$

where for the observation i, the index t is related to variable $T = \mathsf{Time}$, and the index j denotes the values of the variable $X_2 = \mathsf{Level}$. For each subject i, the random effects $\mathbf{b}_i = (b_{i0}, b_{i1})'$ are assumed to follow normal distribution $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$.

(a) Investigate which link function g and distributional assumption fit best to data.

(b) Under the model $\mathcal{M}$, calculate the maximum likelihood estimate $\hat{\mu}_{i_*t}$ for the expected value μ_{i_*t} when

```
ID Level Time
12   15%   20
```

Also, calculate the maximum likelihood prediction $\tilde{\mu}_{i_*t}$ for the expected value μ_{i_*t}.

(c) Test at 5% significance level, is the explanatory variable $X_2 = \mathsf{Level}$ statistically significant variable in the model $\mathcal{M}$. Calculate the value of the test statistic.

(d) Find the estimate for the covariance matrix

$$\mathrm{Cov}(\mathbf{b}_i) = \mathrm{Cov}\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} = \begin{pmatrix} \sigma_{b_0}^2 & \sigma_{b_0,b_1} \\ \sigma_{b_0,b_1} & \sigma_{b_1}^2 \end{pmatrix}.$$

(e) Under the model $\mathcal{M}$, construct 80% prediction interval for the random variable Y_{it}, when

```
ID Level Time
12   15%   20
```

2.1.2 Count Data Models

– For count data, possible distributions and generalized linear mixed models are

Poisson distribution:

$$\begin{aligned}\mathbf{y}|\mathbf{b} &\sim Poi(\boldsymbol{\mu}),\\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b},\\ \mathbf{b} &\sim N(\mathbf{0}, \mathbf{G}).\end{aligned}$$

Negative binomial distribution:

$$\begin{aligned}\mathbf{y}|\mathbf{b} &\sim NegBin(\boldsymbol{\mu}, \Theta),\\ g(\boldsymbol{\mu}) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b},\\ \mathbf{b} &\sim N(\mathbf{0}, \mathbf{G}).\end{aligned}$$

– Usual link functions for these distributions are identity link $g(\mu_i) = \mu_i$ and log link $g(\mu_i) = \log(\mu_i)$.

– Note that in count data situation, often interest is to model the expected value of the ratio $z_i = \frac{y_i}{t_i}$, where t_i in known nonrandom index variable.

– The (conditional) expected value of the ratio $z_i = \frac{y_i}{t_i}$ is usually modeled by the log link function

$$\mathrm{E}(z_i|\mathbf{b}) = \frac{\mu_i}{t_i} = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{b}). \tag{2.4}$$

Example 2.2.

Consider the data in the file ratescancer.txt, where lung cancer cases occur in certain cities at certain ages. In dataset, the response variable is the $Y =$ cases and the index variable is $t =$ pop. The explanatory variables are $X =$ age and $Z =$ city.

```
> ratescancer
         city   age  pop cases
1  Fredericia 40-54 3059    11
2     Horsens 40-54 2879    13
3     Kolding 40-54 3142     4
4       Vejle 40-54 2520     5
5  Fredericia 55-59  800    11
6     Horsens 55-59 1083     6
7     Kolding 55-59 1050     8
8       Vejle 55-59  878     7
9  Fredericia 60-64  710    11
10    Horsens 60-64  923    15
11    Kolding 60-64  895     7
12      Vejle 60-64  839    10
13 Fredericia 65-69  581    10
14    Horsens 65-69  834    10
15    Kolding 65-69  702    11
16      Vejle 65-69  631    14
17 Fredericia 70-74  509    11
18    Horsens 70-74  634    12
19    Kolding 70-74  535     9
20      Vejle 70-74  539     8
21 Fredericia   75+  605    10
22    Horsens   75+  782     2
23    Kolding   75+  659    12
24      Vejle   75+  619     7
```

Consider the mixed effects ratio model

$$\mathcal{M}: \quad \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \alpha_j + b_h,$$

where $b_h \sim N(0, \sigma_z^2)$ are random effects associated to the values of the variable $Z = \mathsf{city}$.

(a) Calculate the maximum likelihood estimate for the ratio $\frac{\mu_i}{t_i}$ when $x_i = 70-74$. Also calculate the maximum likelihood prediction for the ratio $\frac{\mu_i}{t_i}$ when $x_i = 70-74$ and $z_i = \mathsf{Kolding}$.

(b) Test the hypotheses

$$H_0: \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + b_h,$$
$$H_1: \log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \alpha_j + b_h.$$

(c) Create 80 % prediction interval for the ratio $\frac{y_{i*}}{t_{i*}}$ when $x_{i_*} = 70-74$ and $z_{i_*} = \mathsf{Kolding}$.

2.1.3 Binary Data Models

– For binary data, the assumed distribution is Bernoulli distribution $y_i \sim Ber(\mu_i)$, i.e., $P(y_i = 1) = \mu_i$, with most often used link function being the logit link:

$$\mathbf{y}|\mathbf{b} \sim Ber(\boldsymbol{\mu}),$$
$$\text{logit}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \qquad \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$$
$$\mathbf{b} \sim N(\mathbf{0}, \mathbf{G}).$$

– Alternative link functions are based cumulative distribution functions of, e.g., normal distribution - probit link or Cauchy distribution - cauchy link.

– Confidence interval estimation and prediction of μ_i needs to constructed by the bootstrap methods.

– In binary situation, it is meaningful to predict the sum of new unobserved random variables $Y_S = \sum_{i=n+1}^{N} y_{if}$. The prediction interval for the sum Y_S can be based on the interval

$$\left[\widehat{Y}_S - z_{\alpha/2}\sqrt{\widehat{\text{Var}}_b(e_{Y_S})}, \widehat{Y}_S + z_{\alpha/2}\sqrt{\widehat{\text{Var}}_b(e_{Y_S})}\right], \tag{2.5}$$

where $\widehat{Y}_S$ is the point prediction of the Y_S, $\widehat{\text{Var}}_b(e_{Y_S})$ is the bootstrap estimate of the variance of the prediction error $e_{Y_S} = Y_S - \widehat{Y}_S$, and $z_{\alpha/2}$ is $1 - \alpha/2$ quantile of the standard normal distribution.

Example 2.3.

Consider the following dataset:

```
     N Th Age Sex R1 R2 R3 R4
1    1  1  28   1  4  4  4  4
2    2  1  32   1  4  4  4  4
3    3  1  41   1  3  3  3  3
4    4  2  21   1  4  3  3  2
5    5  2  34   1  4  3  3  2
6    6  1  24   1  3  3  3  2
7    7  2  28   1  4  3  3  2
8    8  2  40   1  3  2  2  2
.
```

```
A data frame with 127 observations on the following 8 variables.
N
Patient's number
Th
Therapy ( placebo = 1, treatment = 2)
Age
Age in years
Sex
Gender (male = 0, female = 1)
R1 -Pain before treatment (no pain = 1, severe pain = 5)
R2 -Pain after three days of treatment
R3 -Pain after seven days of treatment
R4 -Pain after ten days of treatment
```

```
In a clinical study n=127 patients with sport related injuries have been treated with
two different therapies (chosen by random design).
After 3,7 and 10 days of treatment the pain occuring during knee movement was observed.
```

Let us model the probability of felt pain being at level 4 or 5 by the longitudinal logistic mixed effects models

$$\mathcal{M}_1: \quad \text{logit}(\mu_{it}) = \beta_0 + \beta_1 t_i + \alpha_j + \gamma_h + b_{i0},$$

$$\mathcal{M}_2: \quad \text{logit}(\mu_{it}) = \beta_0 + \beta_1 t_i + \alpha_j + \gamma_h + b_{i0} + b_{i1} t_i,$$

where j and h are related to the categories of the variables $X_1 = \mathsf{Th}$ and $X_2 = \mathsf{Sex}$. The random effects b_{i0}, b_{i1} are assumed to follow joint normal distribution.

(a) Calculate the prediction $\tilde{\mu}_{i_* t}$ for the expected value $\mu_{i_* t}$ when

```
  N Th Sex  T
127  2   0 11
```

(b) Test at 5% significance level, is the explanatory variable $X_1 = \mathsf{Th}$ statistically significant variable in the model. Calculate the value of the test statistic.

(c) Find the estimate for the covariance matrix

$$\mathrm{Cov}(\mathbf{b}_i) = \mathrm{Cov}\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} = \begin{pmatrix} \sigma^2_{b_0} & \sigma_{b_0,b_1} \\ \sigma_{b_0,b_1} & \sigma^2_{b_1} \end{pmatrix}$$

in the model $\mathcal{M}_2$.

(d) Suppose that there are extra 100 patients outside the data with all being females and not getting any real treatment to their knee pain. Predict how many of these extra 100 patients are feeling knee pain at the level 4 or 5 at time $\mathsf{T} = 3$. Create 80% prediction interval for the number of patients feeling high pain at the time $\mathsf{T} = 11$.