Longitudinal Data Analysis: Generalized repeated measures

Arnab Maity

NCSU Department of Statistics ~ 5240 SAS Hall ~ 919-515-1937 ~ amaity[at]ncsu.edu

Contents

Introduction

In the previous chapters we focused on methods for analyzing longitudinal data where the response variable is continuous and modeled using a multivariate normal distribution. In this chapter, we consider the case where the vector of subject-level responses cannot be modeled using a normal distribution. Examples of such cases are when the response variable is binary, taking the values 0 or 1 (e.g., logistic or probit regression), response is count (e.g., Poisson or negative binomial regression) and so on. We refer to these types of responses by the name *generalized* responses. The models used to analyze generalized responses, analogous to linear models, are called *generalized linear models*[1].

[1] Not to be confused with "general" linear model.

We begin with a review of the generalized linear models for scalar responses and then discuss these class of models for repeated measures.

Review of generalized linear models (GLMs)

Generalized linear models extend the methods of regression analysis to settings where the outcome is dichotomous (binary) variable, count etc. They share many of the characteristics of linear models; most notably the fact that a linear combination of the covariates is related to the mean response. They differ from the linear model in couple of ways including the fact that the distribution of the response may not be normal. Instead the distribution of the response is assumed to be in the *exponential family* of distributions. The exponential family class is very wide and includes normal distribution as a special case, but also Bernoulli distribution (binary data), Poisson (count data), Gamma (strictly positive outcome) etc.

Prussian army horse kick data

Let us take the example of horse kick data[2], containing the numbers of Prussian militiamen killed by being kicked by a horse in separate corps of militiamen between 1875 – 1894. Here is a snapshot of the data:

[2] Hand et al. , 1994; data available in the `pscl` package in R as `prussian`.

```
head(pscl::prussian)
```

```
##   y year corp
## 1 0   75    G
## 2 2   76    G
## 3 2   77    G
## 4 1   78    G
```

```
## 5 0   79    G
## 6 0   80    G
```

```
tail(pscl::prussian)
```

```
##     y year corp
## 275 2   89   XV
## 276 2   90   XV
## 277 0   91   XV
## 278 0   92   XV
## 279 0   93   XV
## 280 0   94   XV
```

Here we can ask: are the differences in the numbers of men killed attributed to systematic effects of year or corps? What are the main challenges with this data?

Generalized linear models with scalar response

Suppose we observe outcome Y_i and a set of covariates $X_{i1}, \ldots, X_{iK}$ for $i = 1, \ldots, n$. The generalized linear models are specified by the following main parts:

- *distributional assumption* of Y_i
- modeling the *systematic component* (i.e. how the covariates are modeled)
- specification of the *link function* (i.e. links the mean response to the systematic component).

The Distributional assumption. We assume that the distribution of Y_i is in the exponential family, and that Y_i's are independent over i. Below, we show a few examples (Normal, Bernoulli and Poisson).

- $Y_i \sim N(\mu_i, \sigma^2)$. This means
$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu_i)^2}.$$
In this case $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma^2$.
- $Y_i \sim Bernoulli(p_i)$ for $p_i > 0$. This means
$$P(Y_i = y) = p_i^y (1-p_i)^{1-y}.$$
In this case $E(Y_i) = p_i$ and $V(Y_i) = p_i(1-p_i)$.
- $Y_i \sim Poisson(\lambda_i)$. This means
$$P(Y_i = y) = \frac{e^{-\lambda_i}\lambda_i^y}{y!}.$$
In this case $E(Y_i) = \lambda_i$ and $V(Y_i) = \lambda_i$.

In general, the exponential family models are denoted by $EF(\eta_i, \phi)$, where η_i is related to the *mean* of Y_i and ϕ_i is called *dispersion parameter* or *scale parameter* and is related to the variance of Y_i.

The systematic component. The systematic component specifies that the effect of the covariates $X_{i1}, \ldots, X_{iK}$ on the mean response Y_i can be expressed in terms of the following linear predictor

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK}$$

or in vector format $\eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta}$, where $\boldsymbol{X}_i = [1, X_{i1}, \ldots, X_{iK}]^T$, and $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$ is the $(K+1)$-dimensional column vector of β_l's. The parameter η_i is called the *linear predictor*. The parameter $\boldsymbol{\beta}$ is called *regression parameter*.

The link function. The *link function* applies a transformation to the mean response and then links the transformed mean to the covariates (through the linear predictor). Denote the mean response $\mu_i = E[Y_i]$. Then the common notation for the link function is

$$g(\mu_i) = \eta_i.$$

The link function is known and assumed monotone and differentiable over the domain of μ_i. Some examples of commonly used link functions are given below.

- *Identity link* (common for normal responses), $g(x) = x$:
- *Logit link* (common for binary responses $0/1$), $g(x) = log\left(\frac{x}{1-x}\right)$.
- *Probit link* (used for binary responses) $g(x) = \Phi^{-1}(x)$ where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal variable $N(0,1)$.
- *Log link* (used for counts responses) $g(x) = log(x)$.

Combining the components. Let us now continue with the previous examples, and see how to combine the three components.

- Linear regression, $Y_i \sim N(\mu_i, \sigma^2)$. This means

  $$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu_i)^2}.$$

  In this case $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma^2$. Here, if we specify the systematic component as

  $$\eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta},$$

and use the *identity link*

$$\mu_i = \eta_i,$$

we arrive as the usual linear regression model:

$$E(Y_i) = \boldsymbol{X}_i\boldsymbol{\beta}.$$

- Logistic regression, $Y_i \sim Bernoulli(p_i)$ for $p_i > 0$. This means

$$P(Y_i = y) = p_i^y(1-p_i)^{1-y}.$$

In this case $E(Y_i) = p_i$ and $V(Y_i) = p_i(1-p_i)$. Here we specify the systematic component as

$$\eta_i = \boldsymbol{X}_i^T\boldsymbol{\beta},$$

and use the logit link

$$log\left(\frac{p_i}{1-p_i}\right) = \eta_i.$$

Thus we have

$$p_i = \frac{exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+exp(\boldsymbol{X}_i\boldsymbol{\beta})}.$$

- Poisson regression, $Y_i \sim Poisson(\lambda_i)$. This means

$$P(Y_i = y) = \frac{e^{-\lambda_i}\lambda_i^y}{y!}.$$

In this case $E(Y_i) = \lambda_i$ and $V(Y_i) = \lambda_i$. Here we again specify the systematic component as

$$\eta_i = \boldsymbol{X}_i^T\boldsymbol{\beta},$$

and use the log link

$$log\,(\lambda_i) = \eta_i.$$

Thus we have

$$\lambda_i = exp(\boldsymbol{X}_i^T\boldsymbol{\beta}).$$

Such a formulation can be applied to the horse kick data where Y_i denotes the number of death during a year, and $\boldsymbol{X}_i$ contains dummy variables indicating which corp the i-th data point belongs to.
In general, if the observed counts Y_i are measured at different time duration, T_i, the we model the *rates* using the log link:

$$Y_i \sim Poisson(\lambda_i) \text{ and } log(\lambda_i/T_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}.$$

The "covariate" T_i is the *time at risk*, and is known as *offset*.

Parameter Estimation and Inference

We estimate the regression coefficients using maximum likelihood estimation approach. This is possible since we explicitly specify distribution of Y_i. The maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is obtained by setting the first derivative of the log-likelihood function to zero and solving for a root. These equation are also also known as *estimating equations* for the regression parameter $\boldsymbol{\beta}$. In general, the estimating equation have the form,

$$\sum_{i=1}^{n} \frac{1}{v(\mu_i)}(Y_i - \mu_i)\left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right) = 0,$$

where

- The term $(Y_i - \mu_i)$ represents the deviation of Y_i from its mean, μ_i,
- The term $v(\mu_i)$ is the variance of Y_i.

Note that the regression parameter $\boldsymbol{\beta}$ can appear in both the mean, μ_i and the weight $(1/v(\mu_i))$. Even when $\partial \mu_i / \partial \boldsymbol{\beta}$ is constant as function of $\boldsymbol{\beta}$ it may be still very complicated to get close form expression for the estimate $\widehat{\boldsymbol{\beta}}$. This optimization problem is not easy to solve, because it involves (in general) a nonlinear system of equations in $\boldsymbol{\beta}$. For that reason, an iterative method is needed. One simple way to understand this procedure is from the sequence of iterations called iterative re-weighted least squares (IRWLS). We will omit the details of IRWLS in this course.

For inference of $\boldsymbol{\beta}$ we need to know its distribution. For large n, the MLE $\widehat{\boldsymbol{\beta}}$ satisfies

$$\widehat{\boldsymbol{\beta}} \sim N_K\{\boldsymbol{\beta}, V\}.$$

Hypothesis tests of the form $H_0 : L\boldsymbol{\beta} = h$ can be carried using Wald test. In particular we notice that

$$L\widehat{\boldsymbol{\beta}} \sim N(L\boldsymbol{\beta}, LVL^T).$$

Thus the construction of the test statistics is similar to that discussed in the previous chapters.

Fitting GLM in R

Generalized linear models are fit in R using the function `glm()`. The command is:

```
glm(formula, family, data, offset)
```

For example, we can fit a Poisson regression model to the horse kick data using corps as covariates as follows.

```
kick_pois <- glm(y ~ corp, family=poisson, data=pscl::prussian)
summary(kick_pois)
```

```
##
## Call:
## glm(formula = y ~ corp, family = poisson, data = pscl::prussian)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5811  -1.0955  -0.8367   0.5438   2.0079
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.231e-01  2.500e-01  -0.893   0.3721
## corpI        4.072e-09  3.535e-01   0.000   1.0000
## corpII      -2.877e-01  3.819e-01  -0.753   0.4512
## corpIII     -2.877e-01  3.819e-01  -0.753   0.4512
## corpIV      -6.931e-01  4.330e-01  -1.601   0.1094
## corpIX      -2.076e-01  3.734e-01  -0.556   0.5781
## corpV       -3.747e-01  3.917e-01  -0.957   0.3387
## corpVI       6.062e-02  3.483e-01   0.174   0.8618
## corpVII     -2.877e-01  3.819e-01  -0.753   0.4512
## corpVIII    -8.267e-01  4.532e-01  -1.824   0.0681 .
## corpX       -6.454e-02  3.594e-01  -0.180   0.8575
## corpXI       4.463e-01  3.202e-01   1.394   0.1633
## corpXIV      4.055e-01  3.227e-01   1.256   0.2090
## corpXV      -6.931e-01  4.330e-01  -1.601   0.1094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 323.23  on 279  degrees of freedom
## Residual deviance: 297.09  on 266  degrees of freedom
## AIC: 630.17
##
## Number of Fisher Scoring iterations: 5
```

Here the intercept represents the corp `G`. Here we are fitting the model

$$log(\lambda_i) = \beta_0 + X_{i1}\beta_1 + \ldots + X_{i,13}\beta_{13},$$

where each X_{ik} is a dummy variable for a specific corp, with the corp `G` being the baseline. Thus, $log(\lambda_i) = \beta_0$ when the i-th data point is from `corpG`, $\beta_0 + \beta_1$ when the i-th data point is from `corpI`, …, $\beta_0 +$

β_{13} when the i-th data point is from `corpXV`. Thus the parameters $\beta_1, \ldots, \beta_{13}$ are the differences in the log-rate between corp G and the other corps. Since all p-values (for $\beta_1 - \beta_{13}$) are non-significant at $\alpha = 0.05$, we can say that no corp show higher death rate than others.

Generalized repeated measures

Now we move onto the case where we have repeated measures/longitudinal data with generalized outcomes. Let us start with an data example.

Example: Epileptic seizures and chemotherapy study.

The dataset and the following description are taken from `https://content.sph.harvard.edu/fitzmaur/ala2e/`. The data are from a placebo-controlled clinical trial of 59 epileptics. Patients with partial seizures were enrolled in a randomized clinical trial of the anti-epileptic drug, progabide. Participants in the study were randomized to either progabide or a placebo, as an adjuvant to the standard anti-epileptic chemotherapy. Progabide is an anti-epileptic drug whose primary mechanism of action is to enhance gamma-aminobutyric acid (GABA) content; GABA is the primary inhibitory neurotransmitter in the brain. Prior to receiving treatment, baseline data on the number of epileptic seizures during the preceding 8-week interval were recorded. Counts of epileptic seizures during 2-week intervals before each of four successive post-randomization clinic visits were recorded.

```
epidata <- read.table("../data/epileptic_seizures.txt")
names(epidata) <-  c("id", "seizures", "time" , "trt", "seizure_base", "age")
head(epidata)
```

```
##    id seizures time trt seizure_base age
## 1 104       11    0   0           11  31
## 2 104        5    1   0           11  31
## 3 104        3    2   0           11  31
## 4 104        3    3   0           11  31
## 5 104        3    4   0           11  31
## 6 106       11    0   0           11  30
```

```
tail(epidata)
```

```
##      id seizures time trt seizure_base age
## 290 232        0    4   1           13  36
## 291 236       12    0   1           12  37
## 292 236        1    1   1           12  37
```

```
## 293 236        4    2   1          12  37
## 294 236        3    3   1          12  37
## 295 236        2    4   1          12  37
```

Here `time = 0` denotes the baseline data, that is, number of epileptic seizures during the preceding 8-week interval. Then `time = 1` to `time = 4` denote counts of epileptic seizures during 2-week intervals before each of four successive post-randomization clinic visits. Thus the column `seizure_base` is same as data from `time = 0` for each individual. The two treatments are recorded in the `trt` column (0=Placebo, 1=Progabide).

We might be interested in asking: Is the mean trend of seizures different across the two treatment groups? The primary objective of the study was to determine whether progabide reduces the rate of seizures in subjects like those in the trial. Additionally, does age affect the mean trend over time?

In general, we have the following setting:

- Outcome of interest: $\{Y_{i1}, \ldots, Y_{im_i}\}$ for unit/subject i
- Y_{ij} either binary, or counts, or rates, etc
- Times of the repeated measurements: $\{t_{i1}, \ldots, t_{im_i}\}$ unit/subject i
- Covariates associated with the jth measurement of the ith subject

$$\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijK});$$

for example $X_{ij1} = t_{ij}$, $X_{ij2} = t_{ij}^2$, $X_{ij3} = Treatment_i$ etc. The covariates describe by X_{ij} could change over time, or not change over time ('time stationary'). Let $\boldsymbol{X}_i$ be $m_i \times K$ matrix obtained by row-stacking X_{ij}.

Our objective is to study the effect of the covariates on the mean response trend. The main challenge is that the responses within each subject are correlated.

We will discuss two approaches:

(1) Extending the ideas of general linear model where we first model the marginal distribution of the response and then model the association between the repeated measures in some way. The models developed in this way are called *population-average models* or *marginal models*.

(2) The second approach is to extent the ideas in linear mixed models (for normal responses) to the generalized response case. We do this by including random effects in formulation.

We discuss these approaches in the next few sections.

Population-average models: specification of marginal models

To simplify our discussion, we will assume that the model matrix $\boldsymbol{X}_i$ is fixed. The marginal models describe separately the distribution for each individual response Y_{ij} and the within subject correlation. The marginal models for longitudinal data have the following three part specification:

1. **mean** for each response Y_{ij}, $E[Y_{ij}] = \mu_{ij}$ depends on the explanatory variables X_{ij} through the link function

$$g(\mu_{ij}) = X_{ij}^T \beta;$$

   $g(\cdot)$ is the known monotone link function and β is the regression parameter that quantifies the (linear) effect of the covariates on the transformed mean response;

2. **variance** of Y_{ij} is assumed to depend on the mean function μ_{ij}

$$\text{Var}(Y_{ij}) = \phi v(\mu_{ij})$$

   where $v(\cdot)$ is the (known) variance function and ϕ is the unknown dispersion parameter. For balanced longitudinal designs a separate scale parameter could be estimated at each occasion ϕ_j; alternatively the scale parameter could depend on the time at which the repeatedly observed outcome Y_{ij} is collected, t_{ij};

3. **within-subject pairwise association** among the repeated responses (i.e. the dependence between Y_{ij} and $ Y_{ij'})$. It is assumed that this association is a function of possibly the mean μ_{ij} and another unknown parameter ω. For example, the components of ω might represent the pairwise correlations between the repeated responses. By analogy, when the responses were continuous and normal assumption was appropriate we described the association by Pearson correlations: $cor(Y_{ij}, Y_{ik}) = \rho_{ijk}(\omega)$, where $\rho_{ijk}(\cdot)$ is used to indicate a correlation function that is known up to the parameter ω.

The three part specification of the marginal models makes the extension of generalized linear models to longitudinal data transparent. The first two parts describe the effects of the covariates on the mean and variance, and they are straightforward extensions from the generalized linear models with scalar response. The last part describes the association among the responses measured on the same unit/subject (recognizes the lack of independence among these responses) and represents the main extension.

The correlation is a natural measure of the linear dependence for continuous responses; however it is not the common measure to describe the association, otherwise. For example, for continuous responses the correlation can be any value between -1 and 1, and it is independent of the means; this is not the case for discrete response. For example, consider the case where Y_1, Y_2 are binary variables such that $P(Y_1 = 1) = 0.2$ and $P(Y_2 = 1) = 0.8$. Then $corr(Y_1, Y_2) \leq 0.25$. Check![3] As a result with discrete responses, correlation is not the common way to describe association. Instead, the odds ratio (or log odds ratio) is a preferable metric to describe association among binary responses.

[3] $corr(Y_1, Y_2) = [P(Y_1 = 1, Y_2 = 1) - P(Y_1 = 1)P(Y_2 = 1)]/[sd(Y_1)Sd(Y_2)] \leq [0.2 - 0.2 * 0.8]/[0.2 * 0.8] = 0.25$. The inequality is due to $E[Y_1Y_2] = P(Y_1 = 1, Y_2 = 1) \leq min(P(Y_1 = 1), P(Y_2 = 1))$.

A few examples of marginal models are shown below.

Normal repeated responses:

- $E[Y_{ij}] = \mu_{ij}$; (identity link function) $\mu_{ij} = X_{ij}\beta$
- $\text{Var}(Y_{ij}) = \phi v(\mu_{ij})$ with $v(\mu_{ij}) = 1$, Caution: this model assumes homogeneous variance.
- $corr(Y_{ij}, Y_{ij'}) = \omega^{|j-j'|}$ if regular design ($t_{ij} = t_j$ for all i).

The marginal model discussed here is an example of the linear regression models for longitudinal studies.

Count repeated responses:

- $E[Y_{ij}] = \mu_{ij}$; (log link function) $\log(\mu_{ij}) = X_{ij}\beta$
- $\text{Var}(Y_{ij}) = \phi v(\mu_{ij})$ with $v(\mu_{ij}) = \mu_{ij}$. Here ϕ is an overdispersion parameter and accounts for the extra variability of the model. Often in medical applications it is necessary to account for this extra variability in order to have accurate inferences about the mean effects.
- Assume unstructured for the pairwise correlation to describe the pairwise association: $corr(Y_{ij}, Y_{ij'}) = \omega_{jj'}$ if regular design. And ω is the vector containing all the pairwise correlations $\omega_{jj'}$

Binary repeated responses $Y_{ij} = 1$ (success) /0 (failure):

- $E[Y_{ij}] = \mu_{ij}$; (logit link function) $logit(\mu_{ij}) = X_{ij}\beta$
- $\text{Var}(Y_{ij}) = v(\mu_{ij})$ with $v(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$.
- Assume unstructured pairwise log odds ratio (OR) pattern to describe the pairwise association: $logOR(Y_{ij}, Y_{ij'}) = \omega_{jj'}$ where

$$OR(Y_{ij}, Y_{ij'}) = \frac{P(Y_{ij} = 1, Y_{ij'} = 1)/P(Y_{ij} = 1, Y_{ij'} = 0)}{P(Y_{ij} = 0, Y_{ij'} = 1)/P(Y_{ij} = 0, Y_{ij'} = 0)}$$

There is an implicit assumption made by the marginal models: mean of Y_{ij} only depends on X_{ij}, that is, mean of the response at j-th time point only depend on the covariates measured at that time point. This assumption holds for time-invariant covariates (X_{ij} does not vary over j), or for time-varying covariates that are set a priori by study design in a manner completely unrelated to the longitudinal response. However when a time-varying covariate varies over time, this assumption may not hold. For example such assumption would be violated when the current value of Y_{ij} given the current covariate X_{ij} predicts the subsequent value of $X_{i(j+1)}$.[4]

[4] For example this may arise in observational studies to assess the effect of physical exercise on reducing the blood glucose level. In this case the subjects with elevated blood glucose level increase the amount of exercise; while the ones with normal level maintain their usual level of exercise. Because this implies a dependence among the covariates.

The correlation model implied by this model specification is popularly referred to in the context of these models as *working correlation model*. This is because this correlation model carries still a lot of uncertainty. The model is considered only a 'working model' rather than necessarily representing what is probably a very complex truth. The model for the pairwise correlation is attempting to represent all sources of variation that could lead to associations among the observations:

- correlation due to within-subject fluctuations (including measurement error)
- correlation due to the between subjects variation

To represent the overall correlation, one can use familiar models that we discussed in the modeling of normally distributed data:

- unstructured correlation
- compound symmetry (exchangeable)
- one dependent correlation (only adjacent observations are correlated)
- AR(1) correlation among observations on the same subject tails off
- Markov models (generalization of AR(1) to unbalanced data)

Working correlation models are very popular in the context of longitudinal data. Let $\Gamma_i(\omega)$ be $m_i \times m_i$ pairwise correlation matrix (that describes the pairwise associations among repeated responses on the same unit), and also let $V(\mu_i)$ be a diagonal matrix with diagonal elements $v(\mu_{ij})$. Then the covariance among the repeated observations is represented as $\Sigma_i = \Sigma_i(\mu_i, \omega)$:

$$\Sigma_i = \phi\{V(\mu_i)\}^{1/2}\Gamma_i(\omega)\{V(\mu_i)\}^{1/2}.$$

Estimation of model components

With discrete response data there is no analogue of the multivariate, normal distribution. Thus there is no 'convenient' likelihood function. Furthermore there is no unified likelihood based framework for

marginal models. The estimation is based on an alternative approach called *Generalized Estimating Equations (GEE)*.

Liang and Zeger (1986) proposed a method for estimating β based on the concept of estimating equations. This provide a general and unified approach for analyzing discrete and continuous responses with marginal models. The key idea is to generalize the unusual univariate likelihood based estimating equations to the case where the response per subject is vector, by introducing the covariance matrix of the vector of responses Y_i.

Recall that for GLMs for univariate response[5], the estimating equations are

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} \frac{1}{v(\mu_i)} (Y_i - \mu_i) = 0,$$

with $\mu_i = X_i^T \boldsymbol{\beta}$. For longitudinal data, these estimating equations are modified to incorporate covariance matrix of the vector of response. We omit the mathematical details.

[5] $Y_i \sim EF(\eta_i, \phi)$ and $E[Y_i] = \mu_i$ and $g(\mu_i) = \eta_i = X_i^T \beta$

It is important to note that GEE estimator is not an ML estimator, as it does not rely on the distributional assumption of Y_i; instead it was derived from an ad-hoc procedure. Nevertheless we can establish theoretical properties of this estimator. Assuming that the estimators of ω (covariance parameters) and ϕ (overdispersion parameter) are consistent, then $\widehat{\beta}$ (regression coefficients, the solution of the GEE) has the following properties:

- $\widehat{\beta}$ is a consistent estimator of β, even when the covariance of Y_i is misspecified,
- for large samples (large n), $\widehat{\beta}$ has approximately multivariate normal distribution,

  $$\widehat{\beta} \sim N\left\{\beta, V_{\widehat{\beta}}\right\}.$$

  In practice we estimate $V_{\widehat{\beta}}$; denote the estimator as $\widehat{V}_{\widehat{\beta}}$.

Like GLS, we can get a model based estimator of $V_{\widehat{\beta}}$ that assumes that the covariance specification of Y_i is correct. Alternatively, we can get a robust estimator (called *empirical sandwitch estimator*) of $V_{\widehat{\beta}}$ to accommodate for possible misspecification of covariace of Y_i. The empirical sandwich estimator is not appealing in the following situations:

- the number of subjects is small comparative to the number of repeated observations per subject/unit
- the sampling design is unbalanced;
- subj/units cannot be grouped on the basis of having identical covariate design matrices

Intuitively there is not sufficient information in the data for the sample covariance to be well estimated. For all these situations, the model based covariance is more suited.

We can then carry out hypotheses tests on $\boldsymbol{\beta}$. As before, we can reformulate the hypothesis testing as $H_0 : L\beta = h$ for some known L and h. We can then use Wald testing procedures as in hypothesis testing for the ordinary generalized linear models. Specifically use the test statistic $\chi^2 = (L\widehat{\beta} - h)^T (L\widehat{V}_{\widehat{\beta}} L^T)^{-1} (L\widehat{\beta} - h)$. Under the null hypothesis, this test statistic has $\chi^2_{\text{number of rows in L}}$.

Model selection for GEE

Since we do not have a proper likelihood function, we can not directly compute criterion like AIC and BIC. Instead, *Quasi-likelihood information criterion (QIC)* was developed by Pan (2001) as a modification of the AIC to apply to models fit by GEE.

Fitting GEE in R

There are a few ways to fit GEE in R such as the function `gee()` in the R package `gee` (Carey et al., 2012), and the function `geeglm()` in the R package `geepack`. Let us demonstrate fitting GEE using the epileptic seizures data.

```
epidata <- read.table("../data/epileptic_seizures.txt")
names(epidata) <- c("id", "seizures", "time" , "trt", "seizure_base", "age")

epidata$trt <- factor(epidata$trt,
                      levels = 0:1,
                      labels = c("placebo", "progabide"))

duration <- rep(2, nrow(epidata))
duration[epidata$time == 0] <- 8
epidata$duration = duration

epidata$baseline <- factor(epidata$time == 0,
                           levels = c(TRUE, FALSE),
                           labels = c("pre", "post"))

#epidata <- within(epidata, {
#  id <- factor(id)
#  trt <- factor(trt, levels = 0:1, labels = c("placebo", "progabide"))
#  })
#epidata = cbind(epidata, as.factor(epidata$time) )
```

```
#names(epidata)[7] = "time_fctr"
head(epidata)
```

```
##    id seizures time     trt seizure_base age duration baseline
## 1 104       11    0 placebo           11  31        8      pre
## 2 104        5    1 placebo           11  31        2     post
## 3 104        3    2 placebo           11  31        2     post
## 4 104        3    3 placebo           11  31        2     post
## 5 104        3    4 placebo           11  31        2     post
## 6 106       11    0 placebo           11  30        8      pre
```

```
#In the code above, we have converted `id` and `trt` to factor variables, added a column with time as a
```

In the code above, we have added the column `duration` to indicate the number of weeks corresponding to each measurement, and the column `baseline` indicates wheather the measurement was taken pre- or post-treatment.

Let's do some exploratory analysis. First, calculate means and variances of the number of seizures for all combinations of treatment and period.

```
itp <- interaction(epidata$trt, epidata$time)
mean_vec <- tapply(epidata$seizures, itp, mean)
var_vec <- tapply(epidata$seizures, itp, var)
cbind(mean_vec, var_vec)
```

```
##              mean_vec   var_vec
## placebo.0   30.785714 681.43386
## progabide.0 31.612903 782.97849
## placebo.1    9.357143 102.75661
## progabide.1  8.580645 332.71828
## placebo.2    8.285714  66.65608
## progabide.2  8.419355 140.65161
## placebo.3    8.785714 215.28571
## progabide.3  8.129032 193.04946
## placebo.4    7.964286  58.18386
## progabide.4  6.709677 126.87957
```

We observe that for each combination of treatment and time, mean and variances are vastly different and thus we do need the overdispersion parameter if we assume the poisson mean and variance combination for Y.

Figure 1 shows that number of seizures over time; each trajectory represents one patient. We see that in the progabide group, there is one individual (ID 207) who has over 150 seizures during the 8-week pre-treatment period, and and has over 300 seizures during the 8-week post-treatment period.

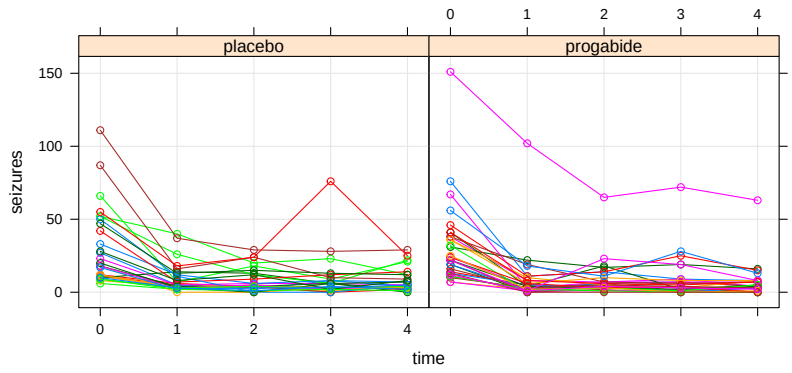

Figure 1: Number of seizures over time in both treatment groups.

```
epidata[which(epidata$seizure_base > 150),]
```

```
##      id seizures time       trt seizure_base age duration baseline
## 241 207      151    0 progabide          151  22        8      pre
## 242 207      102    1 progabide          151  22        2     post
## 243 207       65    2 progabide          151  22        2     post
## 244 207       72    3 progabide          151  22        2     post
## 245 207       63    4 progabide          151  22        2     post
```

Such an observation can easily be an influential point. We will perform our analysis with and without this data point, and compare our findings.

Let us look at mean seizures rate for different treatments and different periods (pre- and post-treatment) for the whole data. To this end we can combine the 4 measurements taken after each treatment and compute the post-treatment seizures rate.

```
itp <- interaction(epidata$trt, epidata$baseline)
mean_vec <- tapply(epidata$seizures/epidata$duration, itp, mean)
mean_vec
```

```
##    placebo.pre  progabide.pre   placebo.post progabide.post
##       3.848214       3.951613       4.299107       3.979839
```

Thus we see that for placebo group, the mean seizures rate changed from pre-treament3.85 seizures per week to 4.3. For the progabide group, the rate changes from 3.95 to 3.98. It is usual to look at the ratio of the post- vs pre-treatment rates (equivalently differences in log ratios). For placebo, the post-to-pre treatment ratio is 1.12, while for progabide, it is 1.01. It seems that for the placebo group there is a 12% increase in rate whereas for progabide group it more or less remains the same. It is usually summarized bu one quantity, called

the cross-product ratio, that is, (progabide post-to-pre ratio)/(placebo post-to-pre ratio). The cross-product ratio indicates whether the decrease in rate due to progabide is more (or less) compared to that of placebo. A value of 1 indicates that the decrease or increase of rates is the same for both treatments. A value less than 1 indicates that the rate decreases more (or increases less) in the progabide group compared to the placebo group. In our example, the cross-product ratio is 0.9, indicating a minor improvement in the progabide group.

If we omit patient ID 207 and perform the same analysis as above, the mean seizures rates are as follows.

```
## progabide   placebo
##      0.83      1.12
```

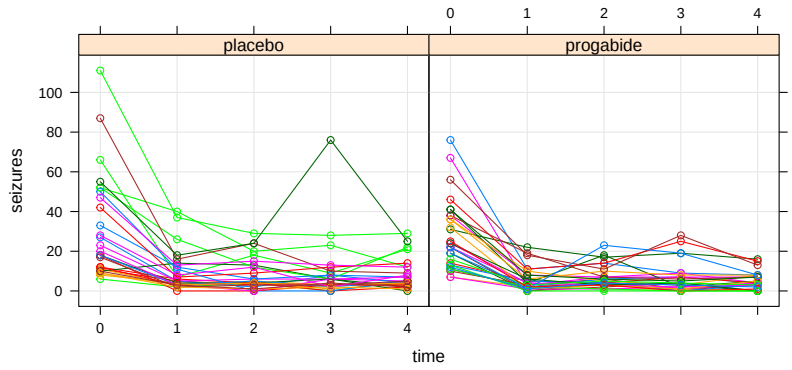

Figure 2: Number of seizures over time in both treatment groups after removing ID 207.

The cross-product ratio is 0.74, indicating a moderate improvement in the progabide group.

To demonstrate a basic GEE fit, let us first only consider the baseline seizures, and compare the two groups (placebo and progabide). If we assume that the response follows a poisson distribution (where the overdispersion parameter is set to $\phi = 1$), we can attempt to simply fit a GLM, where we only have one response Y_i (number of seizures during 8 weeks at baseline), and one covariate X_i (treatment, 0=placebo and 1=progabide)

$$log(E(Y_i)/T_i) = \beta_0 + X_i\beta_1.$$

Here β_0 is the log rate for the placebo group ($X_i = 0$) at baseline, and β_1 is the difference between log rates (progabide - placebo) at baseline, or equivalently, log of the ratio of the rates.

We first fit a standard poisson GLM which assumes that there is no overdispersion, that is, $\phi = 1$.

```
base_data <- epidata[epidata$time == 0, ]

glm.out <- glm(seizures ~ trt, offset = I(log(duration)),
               family = poisson(), data = base_data)
summary(glm.out)
```

```
##
## Call:
## glm(formula = seizures ~ trt, family = poisson(), data = base_data,
##     offset = I(log(duration)))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.473  -3.868  -1.810   1.596  15.280
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.34761    0.03406  39.566   <2e-16 ***
## trtprogabide  0.02651    0.04670   0.568     0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1059.3  on 58  degrees of freedom
## Residual deviance: 1059.0  on 57  degrees of freedom
## AIC: 1358.1
##
## Number of Fisher Scoring iterations: 5
```

The estimate log ratio is 0.027 with standard error 0.047. However this model is not appropriate since we have seen before that there is strong evidence of overdispersion. Thus we now fit a GEE model allowing for such overdispersion. We will use the `gee()` function in the gee package.

```
library(gee)
gee.out <- gee(seizures ~ trt + offset(log(duration)),
               family = poisson(), data = base_data,
               scale.fix = FALSE)

summary(gee.out)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
```

```
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Independent
##
## Call:
## gee(formula = seizures ~ trt + offset(log(duration)), data = base_data,
##     family = poisson(), scale.fix = FALSE)
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
##   2.151786   8.151786  18.048387  37.048387 147.048387
##
##
## Coefficients:
##                Estimate Naive S.E.   Naive z Robust S.E.  Robust z
## (Intercept)  1.34760922  0.1651846 8.1581992   0.1573571 8.5640166
## trtprogabide 0.02651461  0.2264652 0.1170803   0.2218539 0.1195138
##
## Estimated Scale Parameter:  23.5205
## Number of Iterations:  1
##
## Working Correlation
##      [,1]
## [1,]    1
```

Now we see that the estimated overdispersion parameter is $\widehat{\phi} = 23.5205029$. Also the estimated log ratio is 0.027 with standard error 0.226. In other words, the estimated effect is the same but the standard error has been inflated by a factor of $\sqrt{\widehat{\phi}}$. Thus ignoring large overdispersion might lead to false significance.

Let us now fit a practical model:

$$log[E(Y_{ij})/T_i] = \beta_0 + X_{ij,1}\beta_1 + X_{ij,2}\beta_2 + X_{ij,1}X_{ij,2}\beta_3,$$

where $X_{ij,1}$ is the post-treatment indicator (0 if baseline visit, 1 if time is $1, 2, 3, 4$); $X_{ij,2}$ is the treatment indicator (0=placebo, 1=progabide). How to interpret the four regression parameters? Consider $log[E(Y_{ij})]$ for different combination of treatment and time.

- placebo/baseline ($X_{ij,1} = 0$, $X_{ij,2} = 0$): $log[E(Y_{ij})/T_i] = \beta_0$
- progabide/baseline ($X_{ij,1} = 0$, $X_{ij,2} = 1$): $log[E(Y_{ij})/T_i] = \beta_0 + \beta_2$
- placebo/post-treatment ($X_{ij,1} = 1$, $X_{ij,2} = 0$): $log[E(Y_{ij})/T_i] = \beta_0 + \beta_1$

- progabide/post-treatment ($X_{ij,1} = 1$, $X_{ij,2} = 1$): $log[E(Y_{ij})/T_i] = \beta_0 + \beta_1 + \beta_2 + \beta_3$.

Thus β_0 is the log-rate of seizures at baseline for placebo group; β_1 is the *difference* in the log-rate (post - pre) for placebo group; β_2 is the *difference* in log-rate at baseline between treatments (progabide - placebo). To interpret β_3 consider the difference between log-rates post- and pre-treatment for placebo (β_1) and progabide ($\beta_1 + \beta_3$). Each of these difference tells us about the impact (change in log-rate) of the treatment (placebo or progabide) of the log-rate of seizures. Thus $\beta_3 = (\beta_1 + \beta_3) - \beta_1$ represents the how much the impact of progabide (change in log rate from pre- to post treatment) is compared to that of placebo. Thus β_3 is log of the cross-product ratio discussed earlier. Recall that the value of zero for β_3 would indicate that the change in log rate for both the treatments are the same (no effect of treatment). A negative value would indicate that in the progabide group the rate decreased more (or increased less) compared to that in the placebo group.

Since we now have five measuremets per subject, we will need to specify a correlation structure. We will fit an "exchangeable" correlation model (compound symmetry) for this demonstration.

```
gee.full <- gee(seizures ~ baseline*trt + offset(log(duration)),
                id = id,
                family = poisson(), data = epidata,
                scale.fix = FALSE,
                corstr = "exchangeable")

summary(gee.full)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = seizures ~ baseline * trt + offset(log(duration)),
##     id = id, data = epidata, family = poisson(), corstr = "exchangeable",
##     scale.fix = FALSE)
##
## Summary of Residuals:
##        Min         1Q     Median         3Q        Max
```

```
##  -4.299107  -1.299107   2.020161  10.374640 147.048387
##
##
## Coefficients:
##                               Estimate Naive S.E.    Naive z Robust S.E.
## (Intercept)                1.34760922  0.1511851  8.9136359   0.1573571
## baselinepost               0.11079814  0.1547038  0.7161956   0.1160997
## trtprogabide               0.02651461  0.2072721  0.1279217   0.2218539
## baselinepost:trtprogabide -0.10368067  0.2199500 -0.4713830   0.2136100
##                               Robust z
## (Intercept)                8.5640166
## baselinepost               0.9543358
## trtprogabide               0.1195138
## baselinepost:trtprogabide -0.4853736
##
## Estimated Scale Parameter:  19.70269
## Number of Iterations:  1
##
## Working Correlation
##          [,1]     [,2]     [,3]     [,4]     [,5]
## [1,] 1.000000 0.771588 0.771588 0.771588 0.771588
## [2,] 0.771588 1.000000 0.771588 0.771588 0.771588
## [3,] 0.771588 0.771588 1.000000 0.771588 0.771588
## [4,] 0.771588 0.771588 0.771588 1.000000 0.771588
## [5,] 0.771588 0.771588 0.771588 0.771588 1.000000
```

We see that the log cross-product ratio ($\widehat{\beta}_3$) is estimate to be -0.104 with model-based se 0.22 (shown in the "Naive S.E" column) and robust se 0.214 (shown in the "Robust S.E." column). Thus the estimated cross-product ratio is $exp(\widehat{\beta}_3) = 0.9015131$, very similar to what we have seen earlier.

If we omit patient ID 207, and perform the analysis again, the results are shown below.

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = seizures ~ baseline * trt + offset(log(duration)),
```

```
##     id = id, data = epidata_mod, family = poisson(), corstr = "exchangeable",
##     scale.fix = FALSE)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
##  -4.2991071  -0.8541667   2.1458333  10.0345982 107.1517857
##
##
## Coefficients:
##                              Estimate Naive S.E.    Naive z Robust S.E.
## (Intercept)                 1.3476092  0.1105906 12.1855676   0.1573571
## baselinepost                0.1107981  0.1232531  0.8989482   0.1160997
## trtprogabide               -0.1080280  0.1579475 -0.6839487   0.1936732
## baselinepost:trtprogabide -0.3015995  0.1936050 -1.5578080   0.1712004
##                              Robust z
## (Intercept)                 8.5640166
## baselinepost                0.9543358
## trtprogabide               -0.5577850
## baselinepost:trtprogabide -1.7616751
##
## Estimated Scale Parameter:  10.5425
## Number of Iterations:  1
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.5941486 0.5941486 0.5941486 0.5941486
## [2,] 0.5941486 1.0000000 0.5941486 0.5941486 0.5941486
## [3,] 0.5941486 0.5941486 1.0000000 0.5941486 0.5941486
## [4,] 0.5941486 0.5941486 0.5941486 1.0000000 0.5941486
## [5,] 0.5941486 0.5941486 0.5941486 0.5941486 1.0000000
```

Now we see that the log cross-product ratio ($\widehat{\beta}_3$) is estimate to be -0.302 with model-based se 0.194 (shown in the "Naive S.E" column) and robust se 0.171 (shown in the "Robust S.E." column). Thus the estimated cross-product ratio is $exp(\widehat{\beta}_3) = 0.7396343$, again very similar to what we have seen earlier.

As a final demonstration, we now add age to the model.

```
gee.full <- gee(seizures ~ baseline*trt + age + offset(log(duration)),
                id = id,
                family = poisson(), data = epidata,
                scale.fix = FALSE,
                corstr = "exchangeable")

summary(gee.full)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = seizures ~ baseline * trt + age + offset(log(duration)),
##     id = id, data = epidata, family = poisson(), corstr = "exchangeable",
##     scale.fix = FALSE)
##
## Summary of Residuals:
##        Min          1Q     Median          3Q         Max
##  -5.460208  -1.224780   1.912426  10.821128 146.348858
##
##
## Coefficients:
##                              Estimate Naive S.E.     Naive z Robust S.E.
## (Intercept)                 2.26005271 0.44687290  5.05748443  0.43302264
## baselinepost                0.11079814 0.15186363  0.72958968  0.11609974
## trtprogabide               -0.01751109 0.20368821 -0.08597009  0.21406522
## age                        -0.03206495 0.01510322 -2.12305365  0.01474413
## baselinepost:trtprogabide -0.10368067 0.21579786 -0.48045275  0.21361002
##                              Robust z
## (Intercept)                 5.21924832
## baselinepost                0.95433583
## trtprogabide               -0.08180261
## age                        -2.17476110
## baselinepost:trtprogabide -0.48537364
##
## Estimated Scale Parameter:  18.79752
## Number of Iterations:  3
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.7620402 0.7620402 0.7620402 0.7620402
## [2,] 0.7620402 1.0000000 0.7620402 0.7620402 0.7620402
## [3,] 0.7620402 0.7620402 1.0000000 0.7620402 0.7620402
## [4,] 0.7620402 0.7620402 0.7620402 1.0000000 0.7620402
## [5,] 0.7620402 0.7620402 0.7620402 0.7620402 1.0000000
```

It seems that age has some effect (95% CI: [-0.061 -0.003] using robust se), that is, higher age seem to associated with lower log rate.

If we re-do the analysis omitting patient ID 207, the results are as follows.

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logarithm
##  Variance to Mean Relation: Poisson
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = seizures ~ baseline * trt + age + offset(log(duration)),
##     id = id, data = epidata_mod, family = poisson(), corstr = "exchangeable",
##     scale.fix = FALSE)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
##  -5.0989514  -0.9853815   2.4368955  10.3038713 107.3517957
##
##
## Coefficients:
##                              Estimate Naive S.E.    Naive z Robust S.E.
## (Intercept)                1.98864030 0.38465233  5.1699681   0.4417499
## baselinepost               0.11079814 0.12497739  0.8865455   0.1160997
## trtprogabide              -0.13384320 0.16235522 -0.8243850   0.1878251
## age                       -0.02240017 0.01301817 -1.7206855   0.0150550
## baselinepost:trtprogabide -0.30159946 0.19682837 -1.5322967   0.1712004
##                             Robust z
## (Intercept)                4.5017334
## baselinepost               0.9543358
## trtprogabide              -0.7125950
## age                       -1.4878885
## baselinepost:trtprogabide -1.7616751
##
## Estimated Scale Parameter:  11.02957
## Number of Iterations:  4
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.6136193 0.6136193 0.6136193 0.6136193
```

```
## [2,] 0.6136193 1.0000000 0.6136193 0.6136193 0.6136193
## [3,] 0.6136193 0.6136193 1.0000000 0.6136193 0.6136193
## [4,] 0.6136193 0.6136193 0.6136193 1.0000000 0.6136193
## [5,] 0.6136193 0.6136193 0.6136193 0.6136193 1.0000000
```

In this case, age effect is not significant (95% CI: [-0.052 0.007]). This is because patient ID 207 has age only 22 (younger than about 75% of the patients in the study) but much higher number of seizures.

Summary

So far we have discussed marginal models for longitudinal data that are non-normal, how to estimate the model parameters using the GEE and how to make inference about the parameters. In short we described the marginal models by specifying the marginal distribution at each timepoint and an approximate matrix of correlation that links repeated measurements on the same subject.

Notice that we have not actually fully specified the joint distributions of the repeated measurements. The estimation method, GEE, is not likelihood based; however it allows us to estimate and obtain confidence intervals for the regression parameters by accounting for dependence among the repeated measurements.

These population-average models concern mainly the modeling of the marginal parameters and use a 'working correlation' to aggregate the sources of variability in the data. Next we consider another approach *generalized linear mixed models* which will attempt to separate the sources of variability. In addition the methods allow for trajectory reconstruction by incorporating subject specific information. The methods will be close in spirit with the linear mixed models, except here the data are non-normal.

Generalized linear mixed models (GLMM)

For non-normal responses Y_{ij}, the generalized linear mixed models that relates the response to a set of covariates X_{ij} can be described in three steps:

- **Step 1.** Assume that Y_{ij} has a distribution that depends on population specific effects β and a subject specific effect b_i. Conditional on b_i, Y_{ij}, $j = 1, \ldots, n_i$ are independent and follow a distribution from the exponential family model, $Y_{ij}|b_i \sim EF(\eta_{ij}, \phi)$
- **Step 2.** The conditional mean $\mu_{ij} = E[Y_{ij}|b_i]$ can be modeled as

$$g(\mu_{ij}) = X_{ij}\beta + Z_{ij}b_i,$$

  where $g(\cdot)$ is the known monotone link function.

- **Step 3.** The subject specific b_i are assumed to be independent and identically distributed (iid). Typically it is assumed that $b_i \sim N(0, D)$, where D is covariance matrix.

Let us see some examples below.

Normal responses with random intercept: Here we assume Y_{ij} are continuous and follows a normal distribution conditional on random intercept b_{0i}. We posit that

$$E(Y_{ij}|b_{0i}) = \beta_0 + b_{0i} + x_{ij}\beta_1.$$

Further we assume b_{0i} are IID $N(0, D)$ random variables.

Binary response/logit link with random intercept: We assume that $Y_{ij}|b_{0i}$ follows a Bernoulli distribution $Ber(p_i)$, where

$$logit(p_i) = \beta_0 + b_{0i} + x_{ij}\beta_1.$$

Further we assume b_{0i} are IID $N(0, D)$ random variables. Here the fixed effect β_1 can be interpreted as the change in any given individual's log odds of response for one unit increase in the covariate x_{ij}.

Poisson regression with random linear coefficients: We assume that $Y_{ij}|b_{0i}, b_{1i}$ follows a Poisson distribution $Poi(\lambda_i)$, where

$$log(\lambda_i) = \beta_0 + b_{0i} + x_{ij}\beta_1 + x_{ij}b_{1i}.$$

Further we assume $(b_{0i}, b_{1i}$ are IID $N(0, D)$ random variables, where D is a general 2×2 variance-covariance matrix.

Mixed effects models are most useful when we want to make inferences about individuals rather than population averages. Thus the main focus of these models are individual trajectories, and effect of the covariates on the individual. In particular, the regression parameters β's measure the direct influence of the covariates on the response of heterogeneous individuals.

Estimation and Inference for the model parameters in GLMMs

This modeling approach specifies the joint probability function of $Y_i = (Y_{i1}, \ldots, Y_{im_i})^T$ and b_i as,

$$f(Y_i, b_i) = f(Y_i|b_i) \times f(b_i).$$

Estimation and inference for β are based on the marginal or so called "integrated likelihood'' function:

$$\prod_{i=1}^{n} \int f(Y_i|b_i) \times f(b_i)\, db_i,$$

which is the marginal likelihood function averaged over the distribution of the unobserved random effects b_i. This marginal function is used as the likelihood function for β. Then MLE is used for the estimation of β as well as for the estimation of the covariance parameters.

The difficult problem for inference in GLMMs is the presence of the integrals over the random effects, which can be multi-dimensional.For the linear mixed model (normal responses), the marginal distribution could be worked out analytically (another normal). In general, numerical approximations have to be used. This is still an area of active research. Different approaches either approximate the integrand, approximate the data or approximate the integral.Estimation is done through several approaches: (i) Laplace approximation to the integrated likelihood; (ii) penalize quasi-likelihood; (iii) adaptive Gaussian quadrature; and others.

We can also perform hypothesis testing. The model is fitted using maximum likelihood. Thus, in principle, the same large sample theory apply: $\widehat{\beta}$ is asymptotically normally distributed. Specifically, for large samples $\widehat{\beta} \sim N(\beta, V_{\widehat{\beta}})$. Wald, likelihood ratio, or score tests can be used, comparing the test statistic to an appropriate chi-square distribution. However, the validity of the tests depends on the accuracy of the approximations to the likelihood used in estimation.

Fitting GLMM in R

Basic premise is that there is natural heterogeneity across individuals in the study population that is the result of unobserved covariates; random effects account for the unobserved covariates.

Generalized linear mixed models can be fitted in `R` using the `lme4` package. There are two function in the `lme4` package that fit GLMM: `lmer` and `glmer` - which are nearly interchangeable functions. If `lmer` is called with a non-default `family` argument the call is replaced by a call to `glmer` with the current arguments. If `glmer` is called with the default `family', namely the`gaussian' family with the identity link, then the call is replaced by a call to `lmer` with the current arguments. (see the `'Details" section in ?lmer). Adaptive Gaussian Quadrature can be chosen as an option in functionglmerinRpackagelme4'`. It generally is intractable, however, if $q > 2$.

Another way to fit GLMM is using the `gamm` function in `mgcv`.

Let us use the `glmer` function to fit the following model:

$$log[E(Y_{ij})/T_i] = \beta_0 + X_{ij,1}\beta_1 + X_{ij,2}\beta_2 + X_{ij,1}X_{ij,2}\beta_3 + b_{0i} + X_{ij,1}b_{1i},$$

where the random effect $(b_{0i}, b_{1i})^T$ are assumed to be IID $N(0, D)$

with a general 2×2 covariance matrix D.

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
model1 <- seizures ~ baseline*trt +  offset(log(duration)) + (1 + baseline|id)

out = glmer(model1, data = epidata, family = poisson)
summary(out)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: seizures ~ baseline * trt + offset(log(duration)) + (1 + baseline |
##     id)
##    Data: epidata
##
##      AIC      BIC   logLik deviance df.resid
##   1863.3   1889.1   -924.7   1849.3      288
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1388 -0.7118 -0.0607  0.5189  6.9652
##
## Random effects:
##  Groups Name         Variance Std.Dev. Corr
##  id     (Intercept)  0.4999   0.7070
##         baselinepost 0.2319   0.4815   0.17
## Number of obs: 295, groups:  id, 59
##
## Fixed effects:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.071299   0.140268   7.638 2.21e-14 ***
## baselinepost               -0.002394   0.109093  -0.022   0.9825
## trtprogabide                0.049481   0.192718   0.257   0.7974
## baselinepost:trtprogabide -0.307159   0.150452  -2.042   0.0412 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) bslnps trtprg
## baselinepst  0.016
## trtprogabid -0.725 -0.017
## bslnpst:trt -0.018 -0.709  0.030
```