

Longitudinal Data Analysis

Peter Song

*Department of Biostatistics, School of Public Health,
University of Michigan, 1420 Washington Heights,
Ann Arbor, MI 48109-2029, U.S.A.*

pxsong@umich.edu

June, 2008

1

TOPICS

- I. Introduction to Longitudinal Data and Modeling Strategies
- II. Marginal models and Generalized Estimating Equations (GEE)
- III. Mixed-Effects Models and Inferences
- IV. Vector Generalized Linear Models for Correlated Data

Refer to

Song (2007, Ch 1–8) “Correlated Data Analysis: Modeling, Analytics and Applications.” Springer.

University of Michigan

2

Peter Song

Part I

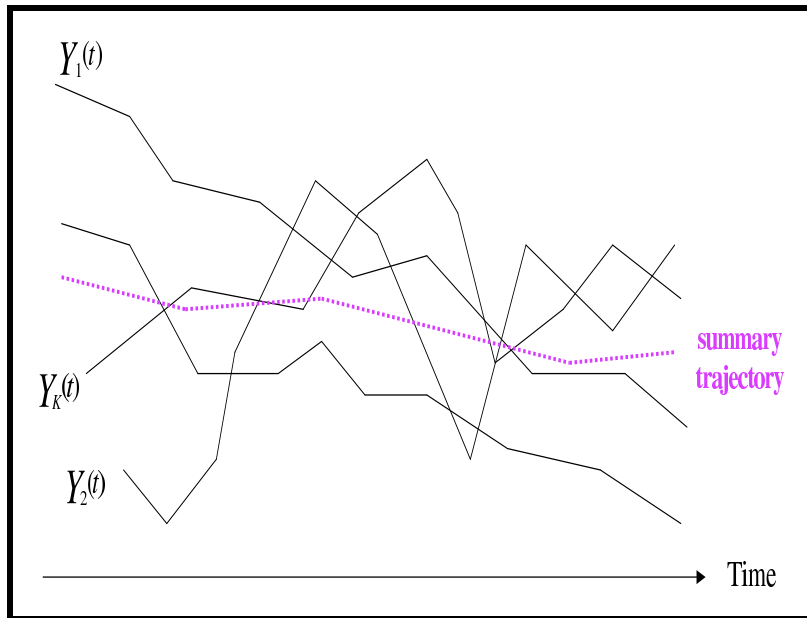
Introduction to Longitudinal Data and Modeling Strategies

3

What Is Longitudinal Data?

- *Longitudinal Data*: Sequentially observed over time, *longitudinal data* may be collected either from an observational study or a designed experiment, in which response variables pertain to a sequence of events or outcomes recorded at certain time points during a study period.
- Longitudinal data may be regarded as a collection of many time series, each for one subject.

4

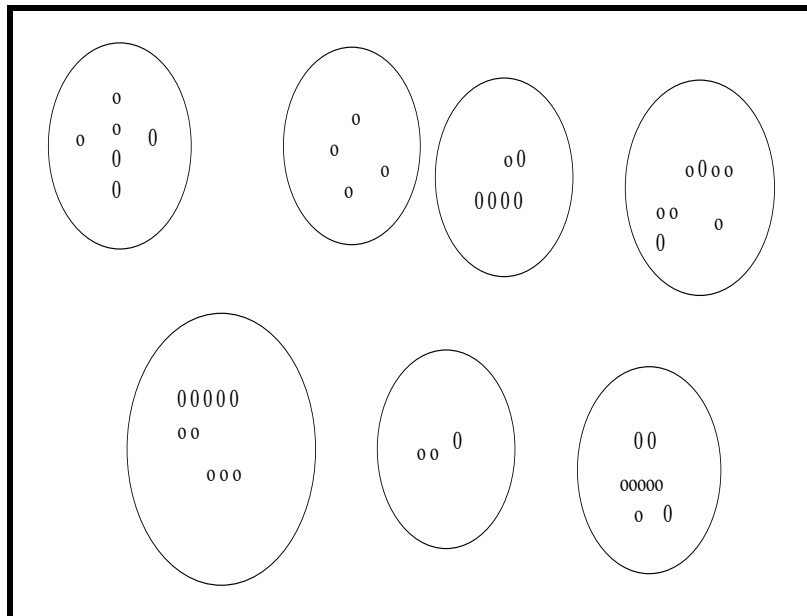


5

- *Clustered data* refers to a set of measurements collected from subjects that are structured in clusters, where a group of related subjects constitutes a cluster, such as a group of genetically related members from a familial pedigree.

6

L



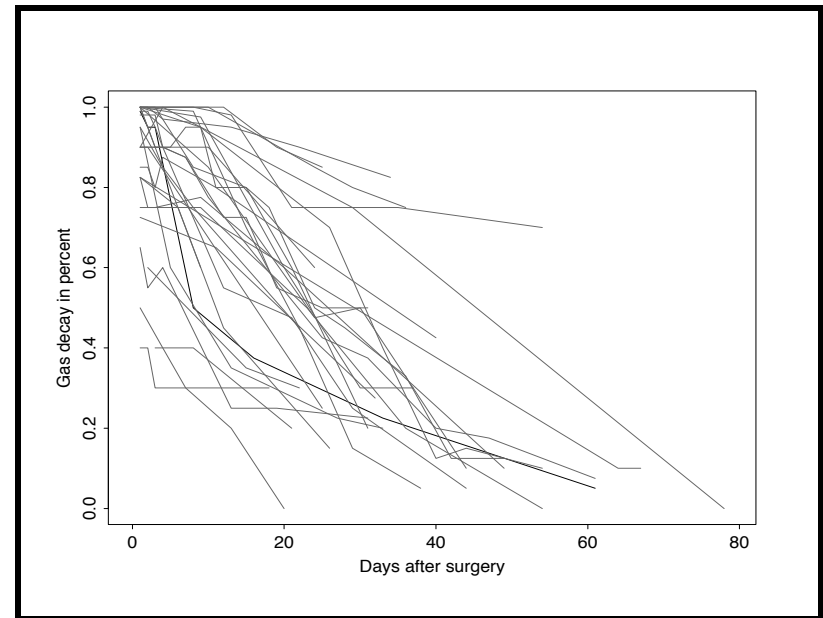
- *Spatial data* are collected from spatially correlated clusters, where correlation structures appear to be 2- or 3-dimensional, as opposed to 1-dim in time for longitudinal data.
- *Multilevel data* are collected from clusters in multi-level hierarchies, such as spatio-temporal data.
- This short course focuses on longitudinal data, and related methodology may be applied to analyze other types of correlated data such as clustered data.

8

Visualize Longitudinal Data I: Spaghetti Plot

- Plotting the time series of all subjects on one graph.
- Useful to observe population-average patterns.

9

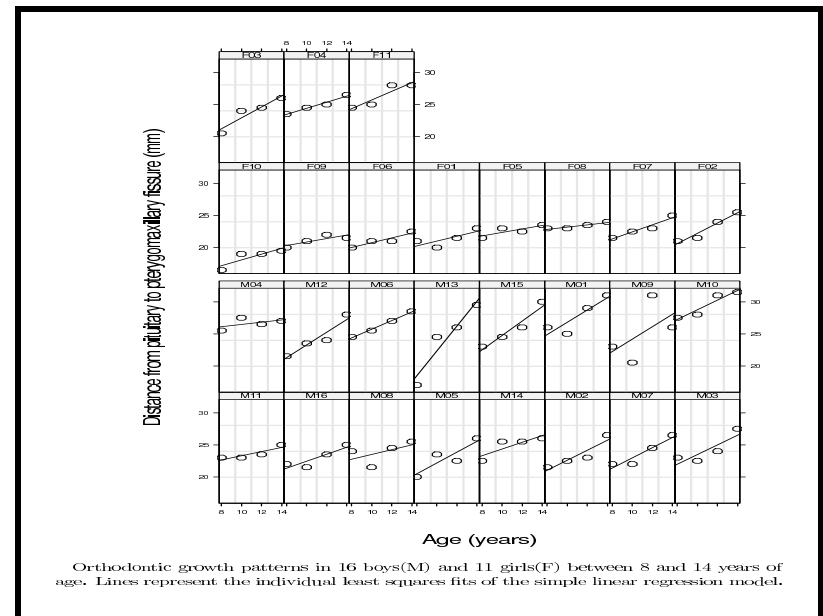


10

Visualize Longitudinal Data II: Trellis Plot

- Plotting the individual time series of subjects, each on one panel.
- Useful to observe subject-specific characteristics.

11



Orthodontic growth patterns in 16 boys(M) and 11 girls(F) between 8 and 14 years of age. Lines represent the individual least squares fits of the simple linear regression model.

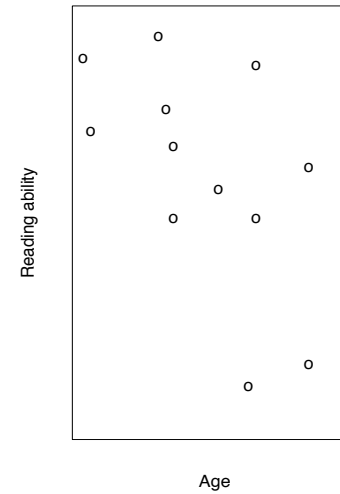
12

Analysis of Longitudinal Data

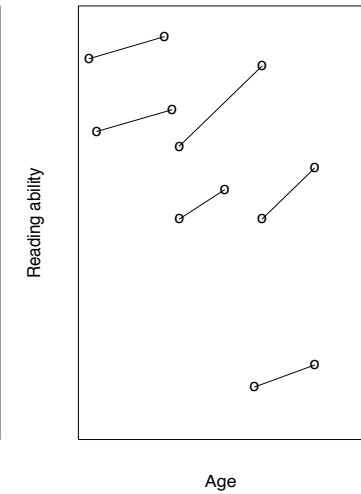
- Primary interest lies in the mechanism of change over time, including growth, time profiles or effects of covariates.
- Main advantages of a longitudinal study:
 - (1) To investigate how the variability of the response varies in time with covariates. For instance, to study time-varying drug efficacy in treating a disease, which cannot be examined by a cross-sectional study.
 - (2) To separate the so-called *cohort* and *age* (or time) effects. From the figure, we learn:
 - (a) Importance of monitoring individual trajectories;
 - (b) Characterize changes within each individual in the reference to his baseline status.
- To help the recruitment of subjects, especially in studies of rare diseases.

13

(a) Cross-sectional study



(b) Longitudinal study



14

Challenges

- Complexity of the underlying probability mechanism of data generation. Likelihood inference is either unavailable or numerically too intricate to be implemented.
- Difficulty of dealing with missing data. (a) Partial information is available to hopefully “recover” the full data; (b) Constraint of preserving the same correlation structure.
- Expectation of dealing nuisance parameters in correlation structures; when time series is long, modeling the transitional behavior (or correlation structure) become a primary task.

15

Main Features

- The presence of repeated measurements for each subject implies that data are autocorrelated or serially correlated. Thus, statistical inference needs to take this serial correlation into account.
- The length of time series determines how much we like to learn about the correlation structure of the data.
- In many practical studies, outcomes are not normally distributed.
- Outcomes are vector-values at give a time point.
- Data contain missing values.

16

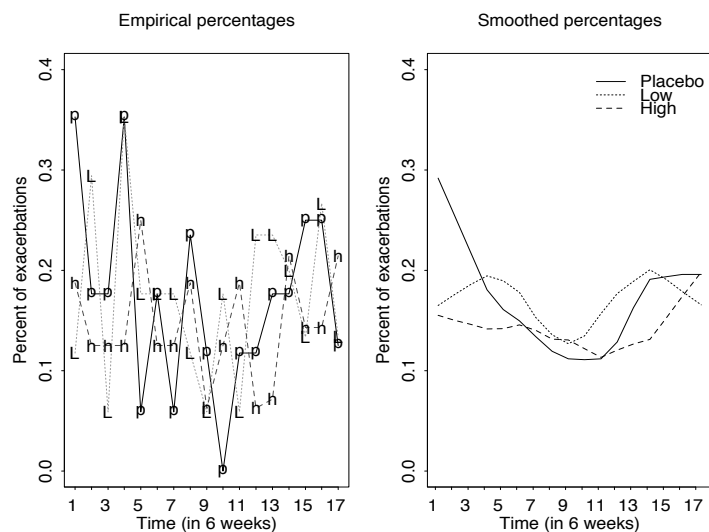
Example 1: Multiple Sclerosis Trial (MST)

- A longitudinal clinical trial to assess the effects of neutralizing antibodies on interferon beta-1 (IFNB) in relapsing-remitting multiple sclerosis (MS), a disease that destroys the myelin sheath surrounding the nerves (Petkau et al, 2004).
- Six-weekly frequent Magnetic Resonance Imaging (MRI) sub-study involving 52 patients, randomized into 3 treatment groups; 17 in placebo, 17 in low dose and 16 in high dose.
- At each of 17 scheduled visits, a binary outcome of *exacerbation* was recorded at the time of each MRI scan, according to whether an exacerbation began since the previous scan.
- Baseline covariates include age, duration of disease (in years), sex, and initial EDSS (expanded disability status scale) scores.
- Does the IFNB help to reduce the risk of exacerbation?

17

- A collection of $N = 52$ short time series, which are equally spaced at 17 time points.

18



19

Example 2: Epileptic Seizures Data

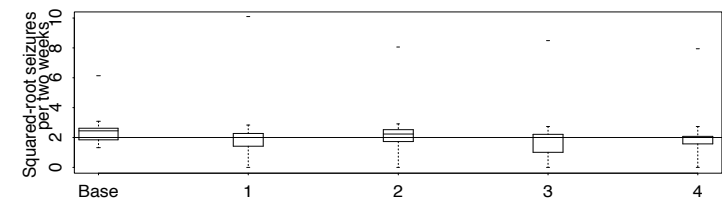
- Data were collected from a clinical trial of 59 epileptics.
- It aimed to examine the effectiveness of the drug **progabide** in treating epileptic seizures.
- For each patient, the number of epileptic seizures was recorded during a baseline period of 8 weeks.
- Patients were then randomized to two treatment arms, one with progabide, and the other with a placebo, in addition to a standard chemotherapy.
- The number of seizures was recorded in 4 consecutive two-week periods after the randomization.
- The scientific question: whether the drug progabide helped to reduce the rate of epileptic seizures.

20

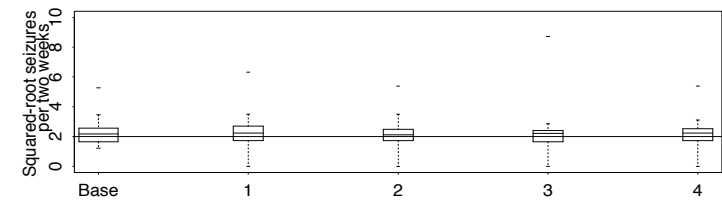
- A collection of 59 short time series, which are equally spaced at 4 time points after randomization.
- ID 207 (in the treatment arm) is a possible outlier, with unduly large counts of epileptic seizures.
- Covariates: Baseline count of seizures (Disease severity), age, treatment, and interaction between age and treatment.

21

Progabide treatment



Placebo treatment

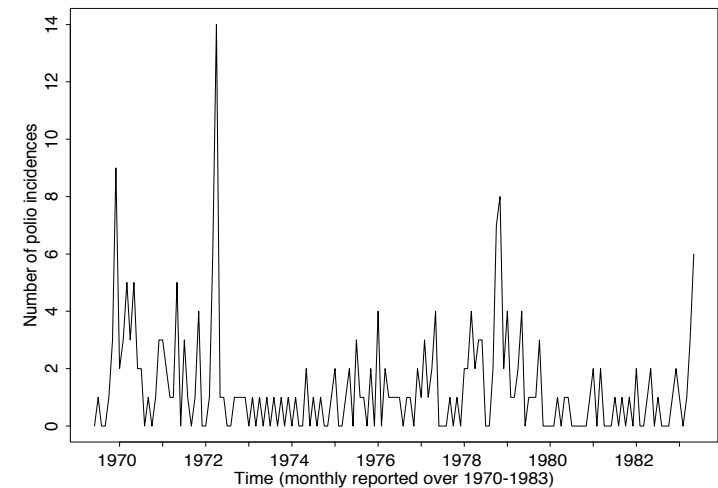


22

Example 3: Polio Incidences in USA

- An example of long time series: monthly counts of polio incidences in USA from 1970 - 1983 (14 years).
- Objective: to assess whether the data provide evidence a decreasing trend in the rate of polio infections over time, after a nationwide anti-polio vaccination policy in early 1970s.
- One time series of 168 repeated measurements.

23



24

Modeling Longitudinal Data

- Express the data in matrix notation, $(\mathbf{y}_i, \mathbf{X}_i, \mathbf{t}_i)$, $i = 1, \dots, N$, where

$$\begin{aligned}\mathbf{y}_i &= (y_{i1}, \dots, y_{in_i})' \\ \mathbf{X}_i &= (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}) \\ \mathbf{t}_i &= (t_{i1}, \dots, t_{in_i})'\end{aligned}$$

- For example of Epileptic Seizures Data: For subject ID 104 (placebo, 31 yrs old, 11 seizures during the 8 weeks prior to the randomization)

25

$$\begin{aligned}\mathbf{y}_1 &= (5, 3, 3, 3)' \\ \mathbf{X}_1 &= \begin{bmatrix} 1 & 0 & 1 & 31 & 11 \\ 1 & 0 & 2 & 31 & 11 \\ 1 & 0 & 3 & 31 & 11 \\ 1 & 0 & 4 & 31 & 11 \end{bmatrix} \\ \mathbf{t}_1 &= (2, 4, 6, 8)'\end{aligned}$$

26

- A parametric modeling framework assumes that \mathbf{y}_i is a realization of \mathbf{Y}_i drawn from a certain population of the form,

$$\mathbf{Y}_i | (\mathbf{X}_i, \mathbf{t}_i) \stackrel{ind.}{\sim} p(\mathbf{y} | \mathbf{X} = \mathbf{X}_i, \mathbf{t} = \mathbf{t}_i; \boldsymbol{\theta}), \quad i = 1, \dots, N,$$

where $\boldsymbol{\theta}$ is the parameter of interest.

- What is $\boldsymbol{\theta}$? Typically, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \Gamma)$, where
 - $\boldsymbol{\beta}$ is the parameter vector involved in a regression model for the mean of the population
 - Γ represents the other model parameters needed for the specification of a full parametric distribution $p(\cdot | \cdot)$, including those in the correlation structure.
- Explicitly specifying such a parametric distribution for nonnormal data is not trivial.
- Multivariate normal! Multivariate binomial? Multivariate Poisson? Multivariate Multinomial? ...

27

- We know how to handle marginals very well from the GLM theory.

$$Y_{ij} | \mathbf{x}_{ij}, t_{ij} \sim \text{GLM}(\mu_{ij}, \sigma_{ij}^2)$$

- The mean μ_{ij} follows a GLM,

$$g(\mu_{ij}) = \eta(\mathbf{x}_{ij}, t_{ij}; \boldsymbol{\beta}), \quad j = 1, \dots, n_i$$

- The dispersion σ_{ij}^2 follows

$$\log(\sigma_{ij}^2) = \zeta(\mathbf{x}_{ij}, t_{ij}; \varsigma).$$

- $\sigma_{ij}^2 = 1$ in Poisson and binary data, unless overdispersion (underdispersion) occurs.

28

Specification of the Mean Structure

- Several commonly used marginal models (specification of η function) in the literature.

(a) Marginal GLM Model takes (the most popular one)

$$\eta(\mathbf{x}_{ij}, t_{ij}; \boldsymbol{\beta}) = \mathbf{x}'_{ij} \boldsymbol{\beta},$$

Parameter $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$ is interpreted as the population-average effects of covariates. They are constant over time as well as across subjects.

(b) Marginal Generalized Additive Model takes

$$\eta(\mathbf{x}_{ij}, t_{ij}; \boldsymbol{\beta}) = \theta_0 + \theta_1(x_{ij1}) + \dots + \theta_p(x_{ijp}),$$

$\boldsymbol{\beta}$ denotes the set of nonparametric regression functions $\theta_l, l = 0, 1, \dots, p$. When one covariate is time t_{ij} , the resulting model characterizes a nonlinear time-varying profile of the data, particularly desirable in longitudinal data analysis.

29

(c) Semi-Parametric Marginal Model includes both parametric and nonparametric predictors, for example,

$$\eta(\mathbf{x}_{ij}, t_{ij}; \boldsymbol{\beta}) = \theta_0(t_{ij}) + \mathbf{x}'_{ij} \boldsymbol{\Upsilon},$$

$\boldsymbol{\beta}$ contains both function $\theta_0(\cdot)$ and coefficients $\boldsymbol{\Upsilon}$. The population-average effect of a covariate (e.g. drug treatment) is adjusted by a nonlinear time-varying baseline effect. Note that it makes no sense to specify a nonlinear function for the covariate of drug treatment.

(d) Time-Varying Coefficient Marginal Model follows a GLM with time-varying coefficients,

$$\eta(\mathbf{x}_{ij}, t_{ij}; \boldsymbol{\beta}) = \mathbf{x}'_{ij} \boldsymbol{\beta}(t_{ij}),$$

$\boldsymbol{\beta} = \boldsymbol{\beta}(t)$ represents a vector of regression coefficient functions in time. This model characterizes time-varying effects of covariates, which is of great interest in longitudinal data analysis. Time-varying

30

effects of covariates, rather than population-average constant effects, are more realistic.

(e) Single-Index Marginal Model is specified

$$\eta(\mathbf{x}_{ij}, t_{ij}; \boldsymbol{\beta}) = \theta_0(t_{ij}) + \theta_1(\mathbf{x}'_{ij} \boldsymbol{\Upsilon}),$$

$\boldsymbol{\beta}$ includes functions $\theta_0(\cdot)$ and $\theta_1(\cdot)$ and the vector of coefficients $\boldsymbol{\Upsilon}$. It is particularly useful for dimension reduction in the presence of a large number of covariates.

(f) A certain combination of models (a)-(e).

- Regression coefficient $\boldsymbol{\beta}$ may be specified as subject-specific, $\boldsymbol{\beta}_i$, leading to another class of so-called mixed-effects models.

31

Longitudinal Data Analysis

WHY BOTHER DEPENDENCE?

Question: what happens if we ignore the dependence in the analysis? (or what if we just used standard regression methods anyway?)

- incorrect assessment of precision of regression coefficient estimates
⇒ incorrect scientific conclusion
- inefficient estimation
OLS versus WLS
- Biased estimates if some obs. were missing *non-trivially*
- Dependence may be of primary interest

University of Michigan

32

Peter Song

IMPACTS OF IGNORING THE DEPENDENCE

Consider $Y_i = (Y_{i1}, \dots, Y_{in}), i = 1 \dots, N$ with

$$E(Y_{ij}) = \beta_0 + \beta_1 x_{ij}, \text{var}(Y_{ij}) = \sigma^2$$

and exchangeable correlation:

$$\text{cov}(Y_{ij}, Y_{ik}) = \sigma^2 \alpha, j \neq k.$$

Ignoring the correlation leads to the OLS estimator $\hat{\beta}_1$ and its variance is

$$V_1 = \text{var}(\hat{\beta}_1; \alpha = 0) = \frac{\sigma^2}{\sum_{i,j}(x_{ij} - \bar{x})^2} = \frac{\sigma^2}{V_T}$$

INCORRECT PRECISION ASSESSMENT

The correct variance of $\hat{\beta}_1$ is

$$V_2 = \text{var}(\hat{\beta}_1; \alpha) = V_1 \{1 + \alpha(n\phi - 1)\}$$

$$\phi = n \sum_{i=1}^N (\bar{x}_i - \bar{x})^2 / V_T$$

Fraction of between subject variation in covariate x 's

Two extreme cases of ϕ :

- $\phi = 0$: $\bar{x}_1 = \dots = \bar{x}_N$
e.g. $x = \text{Time}$: longitudinal study in which every subject is measured at the same set of times
- $\phi = 1$: $x_{i1} = \dots = x_{in}$ (the same value of x for all subjects in cluster i).
e.g. Subject-specific covariate (same drug treatment for a subject)

$$\log \frac{V_1}{V_2} = -\log\{1 + \alpha(n\phi - 1)\}$$

- $\phi = 0$: Within-subject comparison
 - V_1 is wider than should be
 - the discrepancy between V_1 and V_2 increases with α
- $\phi = 1$: between-subject comparison
 - V_1 is too narrow
 - the discrepancy also increases with α

Invalid scientific conclusion may be drawn if V_1 is used as the variance estimate.

EFFICIENCY LOSS

The WLS estimator which properly accounts for the correlation gives the variance

$$V_3 = V_1 \frac{(1 - \alpha)\{1 + (n - 1)\alpha\}}{1 - \alpha + n\alpha(1 - \phi)}$$

Fraction of efficiency loss

$$\frac{V_3}{V_2} = \left\{ 1 + \frac{n^2 \alpha^2 \phi (1 - \phi)}{(1 - \alpha)\{1 + (n - 1)\alpha\}} \right\}^{-1}$$

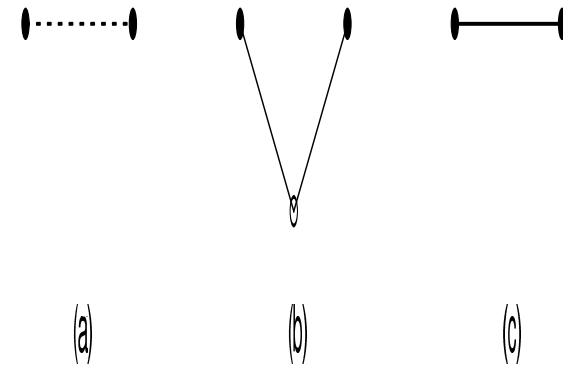
- fully efficient if $\phi = 0$ or 1
- least efficient when ϕ approaches to 0.5

Ignoring correlation leads to a loss of power.

Strategies of Joining Marginal Models

- Not enough to only specify the marginal first moments of the distribution $p(\cdot)$.
- A much harder task is to specify higher moments of the joint distribution $p(\cdot)$ or even the joint distribution itself.
- The marginals have to be joined by a certain suitable correlation structure.
- Three popular strategies of modeling: (a) **Quasi-likelihood Modeling**, (b) **Conditional Modeling**, and (c) **Joint Modeling**.

37



38

Quasi-likelihood (QL) Modeling Approach

- Do not fully specify the joint distribution $p(\cdot)$, but only specify its first two moments, including a correlation structure.
- The minimal set of model conditions required to make a valid statistical inference.
- The QL approach explicitly specifies the covariance of the data, $V_i = \text{cov}(\mathbf{Y}_i | \mathbf{X}_i, \mathbf{t}_i)$:

$$V_i = \text{diag} \left[\sqrt{\text{var}(Y_{ij})} \right] R \text{diag} \left[\sqrt{\text{var}(Y_{ij})} \right]$$

where the key component is the correlation matrix $R = [\alpha_{ts}]$ of \mathbf{Y}_i .

- How to specify R ?
 - Pearson correlation of linear dependency
 - Odds ratio for association between categorical outcomes
 - Nonlinear dependency: Kendall's τ , Spearman's ρ , Normal scoring ν

39

Common Types of Correlation Structures

- (1) (*Independence*) Assumes all pairwise correlation coefficients are zero:

$$\gamma(Y_{it}, Y_{is}) = 0, t \neq s,$$

- (2) (*Unstructured*) Assumes all pairwise correlation coefficients are different parameters:

$$\gamma(Y_{it}, Y_{is}) = \alpha_{st}, t \neq s,$$

- (3) (*Interchangeability, Exchangeable, Compound symmetry*) Assumes pairwise correlation coefficients are equal

$$\gamma(Y_{it}, Y_{is}) = \alpha, t \neq s,$$

40

- (4) (*AR-1*) Assumes the correlation coefficients decay exponentially over time

$$\gamma(Y_{it}, Y_{is}) = \alpha^{|t-s|}, t \neq s,$$

- (5) (*m-dependence*) Assumes the responses are uncorrelated if they are apart more than m units in time, or $|t - s| > m$,

$$\gamma(Y_{it}, Y_{is}) = \alpha_{ts}, \text{ for } |t - s| \leq m,$$

41

Which Correlation Structure Is Suitable?

- Invoke a residual analysis with the following steps:

Step I: Fit longitudinal data by a marginal GLM under the independence correlation structure, and output fitted values $\hat{\mu}_{it}$.

Step II: Calculate the Pearson-type residuals, which presumably carry the information of correlation that was originally ignored in Step I:

$$r_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{V(\hat{\mu}_{it})}}, t = 1, \dots, n_i, i = 1, \dots, N,$$

where $V(\cdot)$ is the variance function chosen according to the marginal model.

Step III: Compute the pairwise Pearson correlations $\hat{\alpha}_{ts}$ of the residuals for each pair of fixed indices (t, s) , which produces a sample correlation matrix $\hat{R} = (\hat{\alpha}_{ts})$.

42

Step IV: Examine the pattern of matrix \hat{R} , to match with one of those listed above.

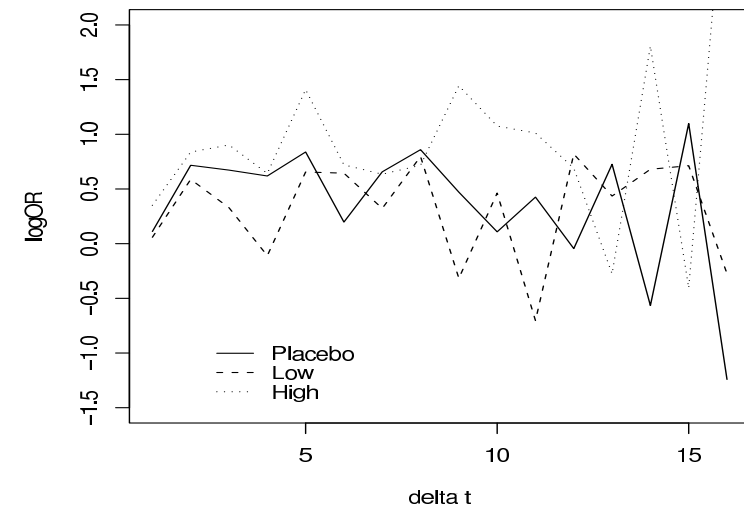
- Step III may be modified as sample log-odds ratios for categorical responses, called *lorelogram* (Heagerty and Zeger, 1998):

$$\text{LOR}(t_j, t_k) = \log \text{OR}(Y_j, Y_k).$$

- For the example of Multiple Sclerosis Trial data, the lorelograms of the observed exacerbation incidences across the 3 treatment groups: (i) LOR is constant over time for each treatment arm; (ii) LOR seems different across different treatment arms; (iii) the variation of LOR increases over time.
- More rigorous decision may be made via a certain model selection criterion.

43

Lorelogram for exacerbation



44

Conditional Modeling Approach

- **Latent Variable Approach:** Conditional on a latent variable \mathbf{b} , $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ are independent,

$$\mathbf{Y} = (Y_1, \dots, Y_n)' | \mathbf{b} \sim p(y_1 | \mathbf{b}) \cdots p(y_n | \mathbf{b}).$$

where conditional distributions are 1-dimensional, so the GLM theory can be applied.

- The joint distribution $p(\cdot)$ is obtained by

$$\begin{aligned} p(\mathbf{y} | \mathbf{X}, \mathbf{t}) &= \int_{\mathcal{B}} p(\mathbf{y}, \mathbf{b} | \mathbf{X}, \mathbf{t}) d\mathbf{b} \\ &= \int_{\mathcal{B}} \prod_{i=1}^n p(y_i | \mathbf{b}, \mathbf{X}, \mathbf{t}) p(\mathbf{X}, \mathbf{t}) d\mathbf{b}, \end{aligned}$$

- The correlation structure is induced from the specification of the latent variables and their distributions.

45

- How many latent variables, *one or ten*, say, and in which form, would be appropriate to capture the underlying true correlation structure of the data? Little study has been done for answers to the questions.

46

- **Transitional Model Approach:** For subject i ,

$$\begin{aligned} p(y_{i1}, \dots, y_{in_i} | \mathbf{X}_i, \mathbf{t}_i) &= f(y_{in_i} | y_{i, n_i-1}, \dots, y_{i,1}, \mathbf{X}_i, \mathbf{t}_i) \times \cdots \\ &\cdots \times f(y_{i2} | y_{i1}, \mathbf{X}_i, \mathbf{t}_i) f(y_{i1} | \mathbf{X}_i, \mathbf{t}_i) \end{aligned}$$

- For example, the transitional logistic model

$$\text{logit} P[Y_{it} = 1 | y_{it-1}, y_{it-2}, \dots, y_{it-q}] = \mathbf{x}'_{it} \boldsymbol{\beta} + \sum_{j=1}^q \psi_j y_{it-j}.$$

- No convenient software packages available to fit transitional models.

47

Joint Modeling Approach

- Directly specify the joint distribution $p(\cdot)$ of the data.
- Mostly *ad hoc* methods, but few general frameworks such as Song's (2000) proposal based on Gaussian copulas.
- The resulting models are called the vector generalized linear models (VGLM).

48

Part II

Marginal Models and Generalized Estimating Equations (GEE)

- OUTLINE
- Marginal models
 - Generalized estimating equation (GEE)
 - SAS codes (PROC GENMOD)

- NOTATION
- A collection of time series (repeated measurements over time), one for each of a number of subjects

$$(y_{ij}, t_{ij}), j = 1, \dots, n_i, i = 1, \dots, N.$$

y_{ij} is observed at time t_{ij} .
 - General assumption:

$$\mathbf{y}_i = (y_{i1}, \dots, y_{in_i}), i = 1, \dots, N$$

$\mathbf{y}_1, \dots, \mathbf{y}_N$ are independent.
 - Quite often in practice, (y_{ij}, t_{ij}) are observed associated with covariates (explanatory variables) \mathbf{x}_{ij} , either time-dependent (age, weight) or time-independent (sex, smoking). As a result, data are presented by triplet

$$(y_{ij}, \mathbf{x}_{ij}, t_{ij})$$

- MARGINAL GENERALIZED LINEAR MODELS (MGLMS)
- Consider longitudinal (or cluster) data $(Y_{ij}, \mathbf{x}_{ij})$ available at time $t_{ij}, j = 1, \dots, n_i; i = 1, \dots, N.$
- Suppose $Y_{ij} \sim DM(\mu_{ij}, \sigma^2)$, where dispersion parameter σ^2 is constant and location parameter μ_{ij} follows a GLM

$$h(\mu_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta}$$
 - ▷ $h(\cdot)$ is a suitable link function, chosen according to the data types.
 - ▷ Y_{ij} and $Y_{ij'}$ are dependent, $j \neq j'$.
 - population-average cause-and-effect mechanism.
 - Statistical tasks: estimation and inference on $\boldsymbol{\beta}$, as well as σ^2 , with incorporation of correlation.
 - Liang and Zeger's (1986, Bka) GEE
 - Qu et al.'s (2000, Bka) quadratic inference function (QIF)

DISPERSION MODELS

- Jørgensen's definition: $Y \sim \text{DM}(\mu, \sigma^2)$ if

$$p(y; \mu, \sigma^2) = a(y; \sigma^2) \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}, y \in \mathcal{S},$$
 where $a \geq 0$ is a suitable (normalizing) function and d is a given regular unit *deviance* satisfying:
 - $d(y; \mu) \geq 0$ with equality when $y = \mu$
 - $\frac{\partial^2 d}{\partial \mu^2}(y; \mu) > 0$
- μ is the location parameter and σ^2 is the dispersion parameter.
- d is a kind of distance measuring the discrepancy between observation y and the expected μ .
- Both conditions assure that d is locally convex around μ .
- The unit variance function $V(\mu)$ is $2 / \frac{\partial^2 d}{\partial \mu^2}(\mu; \mu)$, where the denominator is the curvature of d function at μ .

- The GLMs (also called exponential dispersion (ED) models) are the special cases of the dispersion models because

$$d(y; \mu) = 2 \left[\sup_{\theta \in \Theta} \{y\theta - \kappa(\theta)\} - y\tau^{-1}(\mu) + \kappa\{\tau^{-1}(\mu)\} \right].$$
- The ED family distributions include: normal, Poisson, binomial, negative binomial, gamma, inverse Gaussian, ...
- Important property: Mean-variance relation:

$$\text{var}(Y) = \sigma^2 V(E(Y)) = \sigma^2 V(\mu).$$
- Some DM but not ED family distributions: Simplex distribution and von Mises distribution

- The von Mises distribution for angular (circular) data (Fisher, 93): $Y \sim \text{vM}(\mu, \sigma^2)$

$$\frac{e^\lambda}{2\pi I_0(\lambda)} \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}, \lambda = 1/\sigma^2,$$
 where $I_0(\cdot)$ is the modified Bessel function, and

$$d(y, \mu) = 2\{1 - \cos(y - \mu)\}, y, \mu \in [0, 2\pi),$$
 and $V(\mu) = 1$.
- Popular for modeling angular or circular data
- Analog to the normal distribution on the manifold of the unit circle

- The simplex distribution for proportional data (Barndorff-Nielsen and Jørgensen, 91, JMA): $Y \sim S^-(\mu, \sigma^2)$

$$[2\pi\sigma^2 \{y(1-y)\}^3]^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} d(y; \mu) \right\}$$
 where

$$d(y, \mu) = \frac{(y - \mu)^2}{y(1-y)\mu^2(1-\mu)^2}, y, \mu \in (0, 1)$$
 and $V(\mu) = \mu^3(1-\mu)^3$.
- Useful to model continuous proportional data
- Better than the beta distribution from the GLM perspective

NAIVE ESTIMATING EQUATION

Assume all Y_{it_j} independent. Then log-likelihood function

$$\sum_{i=1}^N \sum_{j=1}^{n_i} \left\{ \log a(y_{it_j}; \sigma^2) - \frac{1}{2\sigma^2} d(y_{it_j}; \mu_{ij}(\boldsymbol{\beta})) \right\}$$

Let $\mathbf{u}_i = (u_{i1}, \dots, u_{in_i})^\top$ be the score vector of the form

$$u_{ij} = -\frac{1}{2} d'(y_{it_j}; \mu_{ij})$$

For the ED models,

$$u_{ij} = \frac{y_{it_j} - \mu_{ij}}{V(\mu_{ij})}$$

But in general u_{ij} is a non-linear function in both arguments, e.g. for the Simplex,

$$u = \frac{y - \mu}{\mu(1 - \mu)} \left\{ d(y; \mu) + \frac{1}{\mu^2(1 - \mu)^2} \right\}.$$

- The (naive) score function w.r.t. $\boldsymbol{\beta}$ is

$$\Psi_I(\boldsymbol{\beta}; \mathbf{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^N \left(\frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \right) \mathbf{u}_i(\mathbf{y}_i; \boldsymbol{\mu}_i)$$

- It is shown to be unbiased, $E(\Psi_I(\boldsymbol{\beta}; \mathbf{Y})) = \mathbf{0}$, under the regularity condition of the exchangeability between integration and differentiation.

- The naive estimator, $\hat{\boldsymbol{\beta}}_I$, of $\boldsymbol{\beta}$ is defined as the solution to the estimating equation,

$$\Psi_I(\boldsymbol{\beta}; \mathbf{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^N \mathbf{D}_i^\top \mathbf{u}_i(\mathbf{y}_i; \boldsymbol{\mu}_i) = \mathbf{0}$$

- Note that $\hat{\boldsymbol{\beta}}_I$ is a GLM estimator. In particular, for the case of cross-sectional study, $N = 1$, $\hat{\boldsymbol{\beta}}_I$ is the MLE.

OPTIMAL ESTIMATING FUNCTION

- Consider a class of unbiased estimating functions of the form

$$\Psi_c(\boldsymbol{\beta}) = \sum_{i=1}^N \mathbf{C}_i(\boldsymbol{\beta}) \mathbf{u}_i(\mathbf{y}_i; \boldsymbol{\mu}_i)$$

where \mathbf{C}_i is a nonstochastic weighting matrix. Clearly the naive estimating function Ψ_I belongs to this class.

- According to Crowder (1987), the optimal function of this form is given by taking

$$\mathbf{C}_i(\boldsymbol{\beta}) = E \left\{ \frac{\partial \mathbf{u}_i^\top(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\} \text{var}^{-1}(\mathbf{u}_i).$$

- Therefore the optimal estimating function is obtained as

$$\Psi_o(\boldsymbol{\beta}) = - \sum_{i=1}^N \mathbf{D}_i^\top \mathbf{A}_i \text{var}^{-1}(\mathbf{z}_i) \mathbf{z}_i$$

where

$$\mathbf{z}_i = \text{diag} \{ V(\mu_{ij}) \} \mathbf{u}_i,$$

called as “working” score vector, and

$$\mathbf{A}_i = \sigma^{-2} \text{diag} \{ \text{var}(u_{ij}) V(\mu_{ij}) \}.$$

- Note that $\mathbf{z}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ in the case of ED models.

ESTIMATING EQUATION FOR β

- Following Liang and Zeger (1986), the working covariance matrix

$$\mathbf{V}_i = \text{diag}^{1/2} \{ \text{var}(z_{ij}) \} \mathbf{R}(\alpha) \text{diag}^{1/2} \{ \text{var}(z_{ij}) \}$$
 where $\mathbf{R}(\alpha)$ is a correlation matrix and α is a $q \times 1$ vector which fully characterizes $\mathbf{R}(\alpha)$.
 Define GEE as

$$\Psi(\beta, \alpha) = - \sum_{i=1}^N \mathbf{D}_i^\top \mathbf{A}_i \mathbf{V}_i^{-1} \mathbf{z}_i = \mathbf{0}$$
 and hence define the estimate, $\hat{\beta}$, of β as the solution to this equation.
- SAS PROC GENMOD implements GEE1: $\Psi(\beta, \hat{\alpha}) = \mathbf{0}$, where the estimate $\hat{\alpha}$ (estimated externally) is plugged in the equation.
- α has to be properly estimated (consistency); if not, the GEE1 estimator of β is problematic!

ESTIMATING EQUATION FOR α

- Define residuals by

$$r_{ij} = \frac{u_{ij}}{\sqrt{\frac{1}{2} \text{Ed}''(y_{ij}; \mu_{ij})}}$$
- They becomes Pearson residuals for ED models.

$$\text{E}(r_{ij}) = 0, \text{var}(r_{ij}) = \sigma^2,$$

$$\text{E}(r_{ij}r_{ij'}) = \sigma^2 \text{corr}(u_{ij}, u_{ij'}) = \sigma^2 \text{corr}(z_{ij}, z_{ij'}).$$
- Following Prentice (1988), form an additional estimating equation for α ,

$$\Phi(\beta, \alpha) = \sum_{i=1}^N \left(\frac{\partial \boldsymbol{\eta}_i^\top}{\partial \alpha} \right) \mathbf{W}_i^{-1} (\mathbf{r}_i - \boldsymbol{\eta}_i) = \mathbf{0}$$
 where $\mathbf{r}_i = (r_{i1}r_{i2}, r_{i1}r_{i3}, \dots, r_{in_i-1}r_{in_i})^\top$, \mathbf{W}_i is a working covariance matrix and $\boldsymbol{\eta}_i = \text{E}(\mathbf{r}_i)$.

GEE2: AN EXTENDED VERSION

- A joint generalized estimating equations for β and α are obtained,

$$\Upsilon(\boldsymbol{\theta}; \mathbf{Y}) = \Upsilon(\alpha, \beta; \mathbf{Y}) = \begin{bmatrix} \Psi(\alpha, \beta; \mathbf{Y}) \\ \Phi(\alpha, \beta; \mathbf{Y}) \end{bmatrix} = \mathbf{0}, \quad (1)$$
 and simultaneously solving the joint equations for $\hat{\beta}$ and $\hat{\alpha}$.
- $\Upsilon(\boldsymbol{\theta}; \mathbf{Y})$ is unbiased, namely $\text{E}\Upsilon(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{0}$.
- Under some mild regularity conditions, the estimator $\hat{\boldsymbol{\theta}} = (\hat{\beta}, \hat{\alpha})$ is consistent and $N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically multivariate Gaussian with zero mean and covariance matrix of the form $\lim_N N\mathbf{J}^{-1}(\boldsymbol{\theta})$ where $\mathbf{J}(\boldsymbol{\theta})$ is the Godambe information matrix given by

$$\mathbf{J}(\boldsymbol{\theta}) = \mathbf{S}^\top \mathbf{V}^{-1} \mathbf{S}.$$
- Therefore the (asymptotic) standard errors of $\hat{\boldsymbol{\theta}}$ are given by the square-root of diagonal elements of $\mathbf{J}(\hat{\boldsymbol{\theta}})$.

- The sensitivity matrix \mathbf{S} is a block diagonal matrix, $\text{diag}(\mathbf{S}_1, \mathbf{S}_2)$, with

$$\mathbf{S}_1 = \text{E}\Psi'_\beta(\boldsymbol{\theta}) = - \sum_{i=1}^N \mathbf{D}_i^\top \mathbf{A}_i \mathbf{V}_i^{-1} \mathbf{A}_i \mathbf{D}_i,$$

$$\mathbf{S}_2 = \text{E}\Phi'_\alpha(\boldsymbol{\theta}) = - \sum_{i=1}^N \left(\frac{\partial \boldsymbol{\eta}_i^\top}{\partial \alpha} \right) \mathbf{W}_i^{-1} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \alpha^\top} \right).$$
- The variability matrix can be written in block matrix as follows,

$$\mathbf{V} = \text{E}\Upsilon(\boldsymbol{\theta})\Upsilon^\top(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix},$$
 with

$$\mathbf{V}_{11} = \sum_{i=1}^N \mathbf{D}_i^\top \mathbf{A}_i \mathbf{V}_i^{-1} \text{cov}(\mathbf{z}_i) \mathbf{V}_i^{-1} \mathbf{A}_i \mathbf{D}_i$$

$$\mathbf{V}_{22} = \sum_{i=1}^N \left(\frac{\partial \boldsymbol{\eta}_i^\top}{\partial \boldsymbol{\alpha}} \right) \mathbf{W}_i^{-1} \text{cov}(\mathbf{r}_i) \mathbf{W}_i^{-1} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\alpha}^\top} \right),$$

$$\mathbf{V}_{12} = \sum_{i=1}^N \mathbf{D}_i^\top \mathbf{A}_i \mathbf{V}_i^{-1} \text{cov}(\mathbf{z}_i, \mathbf{r}_i) \mathbf{W}_i^{-1} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\alpha}^\top} \right),$$

and $\mathbf{V}_{21} = \mathbf{V}_{12}^\top$.

- Note that

$$\text{cov}(\mathbf{z}_i) = \text{diag} \{V(\mu_{ij})\} \text{cov}(\mathbf{u}_i) \text{diag} \{V(\mu_{ij})\},$$

and an estimate of $\text{cov}(\mathbf{z}_i)$ is obtained by plugging the estimates of $\hat{\mu}_{ij}$ and replacing $\text{cov}(\mathbf{u}_i)$ by $\hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top$ in the above expression. The same approach is used to calculate the other blocks of \mathbf{V} .

- The dispersion parameter σ^2 is estimated by the so-called Jørgensen estimator:

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^N n_i} \sum_{i=1}^N \sum_{j=1}^{n_i} d(y_{ij}; \hat{\mu}_{ij})$$

or

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^N n_i - p} \sum_{i=1}^N \sum_{j=1}^{n_i} d(y_{ij}; \hat{\mu}_{ij})$$

NEWTON-SCORING ALGORITHM

- Find the solution of GEE iteratively

$$\begin{aligned} \boldsymbol{\theta}^{(n+1)} &= \boldsymbol{\theta}^{(n)} - \mathbf{S}^{-1} \Upsilon \left(\boldsymbol{\theta}^{(n)} \right) \\ &= \text{diag} (\mathbf{S}_1^{-1}, \mathbf{S}_2^{-1}) \Upsilon \left(\boldsymbol{\theta}^{(n)} \right). \end{aligned}$$

- Equivalently, iteratively solve the two equations till convergence.

SOME EXAMPLES OF $\hat{\boldsymbol{\alpha}}$

- Exchangeability structure: $\text{corr}(z_{ij}, z_{ij'}) = \alpha$, and $\mathbf{W}_i = I$,

$$\hat{\alpha} = \hat{\sigma}^{-2} \sum_{i=1}^N \sum_{j>j'} \hat{r}_{ij} \hat{r}_{ij'} / \left\{ \sum_{i=1}^N \frac{1}{2} n_i (n_i - 1) - p \right\}$$

- AR(1) structure: $\text{corr}(z_{ij}, z_{ij'}) = \alpha^{|t_{ij} - t_{ij'}|}$, $|\alpha| < 1$, or the exponential correlation model (ECM) $\text{corr}(z_{ij}, z_{ij'}) = \exp(-\alpha |t_{ij} - t_{ij'}|^k)$, for $\alpha > 0$ and certain $k = 1, 2, \dots$

- In particular, for ECM with $k = 1$ and $\mathbf{W}_i = I$, then

$$\Phi(\alpha) = \sum_{i=1}^N \mathbf{c}_i^\top (\mathbf{r}_i - \boldsymbol{\eta}_i) = 0$$

where \mathbf{c}_i is given by

$$[|t_{i1} - t_{i2}| \exp(-\alpha |t_{i1} - t_{i2}|), \dots, |t_{in_i-1} - t_{in_i}| \exp(-\alpha |t_{in_i-1} - t_{in_i}|)]^\top.$$

MARGINAL MODEL ANALYSIS OF RETINAL DATA

- Measure of proportions:

$$Y_t = \frac{\text{Volume}_t}{\text{Volume}_0},$$
- Background: From a prospective study in ophthalmology where intraocular gas was used in complex retinal surgeries to provide internal tamponade of retinal breaks in the eye. 31 patients were visited 3 to 8 times over a three-month period after gas injection into their eyes, and the volume of the gas in their eyes at the follow up time was recorded as a *percentage* to the initial gas volume in their eyes.
 - y_{ij} = percentage of gas volume for patient i at time t_{ij} .
 - \mathbf{x}_{ij} = [Time after surgery (days), gas concentration]
- Objective: To estimate the kinetics (*e.g.*, decay rate, half-life, and so on) of the disappearance of the gas.

- Y_{it} 's are percents confined (0, 1)
- Assume $Y_{it} \sim S^-(\mu_{it}, \sigma^2)$, Simplex distribution
- $\mu_{it} = E(Y_{it}) \in (0, 1)$
- Marginal model takes the form

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 \log(\text{Time}_{ij}) + \beta_2 \log^2(\text{Time}_{ij}) + \beta_3 \text{GAS}_i.$$

- Newton-scoring algorithm produces:

| <i>Results of the Ophthalmology Study</i> | | | | | | |
|---|--------|------|-----------|------|-------|------|
| Variable | Indep. | | Exchange. | | AR(1) | |
| | Est. | s.e. | Est. | s.e. | Est. | s.e. |
| Intercept | 2.69 | 0.30 | 2.65 | 0.29 | 2.73 | 0.27 |
| $\log(\text{TIME})$ | 0.06 | 0.25 | 0.16 | 0.20 | 0.10 | 0.20 |
| $\log^2(\text{TIME})$ | -0.34 | 0.07 | -0.38 | 0.05 | -0.35 | 0.04 |
| <i>GAS</i> | 0.33 | 0.19 | 0.25 | 0.18 | 0.30 | 0.17 |
| α | - | - | 0.25 | 0.04 | 0.49 | 0.14 |
| σ | 14.16 | - | 14.24 | - | 14.22 | - |

Merits of the GEE Method

- It is useful to evaluate the population-average effects of covariates.
- It is simple, as it only requires to correctly specify the first two moments of the underlying distribution of the data.
- It is robust against the model misspecification on the correlation structure.
- It is easy to implement numerically using available software packages such as SAS and R. This is really under the framework of Weighted Least Squares.

Caveats of the GEE Method

- (1) First underlying assumption is that data are relatively homogeneous, in the sense that the variation in the response is mostly due to different levels of covariates (not due to subject-specific variation).
- (2) Second underlying assumption is that the first moment mean model is correct,

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$$
- (3) Third underlying assumption is that the nuisance correlation parameter α is properly estimated.
- (4) Fourth underlying assumption is that missing data are missing completely at random (MCAR).
- (5) No way of performing model selection because of the lack of an objective function in the estimation procedure.
 - QIF can help to deal with (2), (3), and (5).

73

```

title "UNSTRUCTURED CORRELATION";
proc genmod data=exacerb;
class id;
model rel= dose dur t1 t2 / dist=bin
link=logit;
repeated subject=id / type=un corrw covb modelse;
run;

title "EXCHANGEABLE CORRELATION (type=cs)";
proc genmod data=exacerb;
class id;
model rel= dose dur t1 t2 / dist=bin
link=logit;
repeated subject=id / type=exch corrw covb modelse;
run;

```

75

SAS CODE: ILLUSTRATED BY MULTIPLE SCLEROSIS DATA

- Only applicable for ED family distributions (binomial, Poisson, gamma) not for general DM distributions
- SAS code for multiple sclerosis data:
 y_{it} : exacerbation (yes/no)
 \mathbf{x}_{it} : (1, dose levels, duration, time, time²)

$$\log \frac{\pi_{it}}{1 - \pi_{it}} = \beta_0 + \beta_1 DOSE_i + \beta_2 DUR_i + \beta_3 TIME_t + \beta_4 TIME_t^2$$

University of Michigan

74

Peter Song

```

title "AR(1) CORRELATION";
proc genmod data=exacerb;
class id;
model rel= dose dur t1 t2 / dist=bin
link=logit;
repeated subject=id / type=ar corrw covb modelse;
run;

```

76

Part III

Generalized Linear Mixed Models (GLMM) and Inferences

GLMM FOR CONTINUOUS RESPONSE

- A GLMM is specified by a hierarchical structure.

- **Stage I:**

$$Y_{ij} = \beta_{i0} + \beta_{i1}t_j + \varepsilon_{ij}, j = 0, 1, \dots, n, i = 1, \dots, N$$

- $t_0 = 0$ (baseline)
- β_{i0} : average effect for subject (or cluster) i
- β_{i1} : rate of change for subject (or cluster) i
- $\varepsilon_{ij} | \beta_{i0}, \beta_{i1} \stackrel{iid}{\sim} N(0, \sigma^2)$

OUTLINE

- GLMM specification and interpretation
- Approximate inference
- Inference via EM-algorithm
- Bayesian inference using MCMC

- **Stage II:**

$$\beta_{i0} = \alpha_0 + \alpha_1^\top \mathbf{x}_i + b_{i0}$$

$$\beta_{i1} = \gamma_0 + \gamma_1 z_i + b_{i1}$$

- $z_i = 1$ if new treatment or 0 if placebo
- \mathbf{x}_i : the baseline covariates (may include z_i)

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \stackrel{iid}{\sim} BVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D} = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} \right)$$

- **Parameter Interpretation:**
 - γ_0 : average rate of change in placebo group ($z_i = 0$)
 - $(\gamma_0 + \gamma_1)$: average rate of change in treatment group ($z_i = 1$)
 - α 's characterize the dependence of Y on baseline covariates
 - Combining the stages:

$$Y_{ij} = \alpha_0 + \alpha_1^\top \mathbf{x}_i + \gamma_0 t_j + \gamma_1 z_i t_j + b_{i0} + t_j b_{i1} + \varepsilon_{ij}$$
 - b_{i0} and b_{i1} vary from subject to subject reflecting underlying heterogeneity caused by unmeasured factors
 - Heterogeneity across subjects leads to within-subject dependence

- $E(Y_{ij}) = \alpha_0 + \alpha_1^\top \mathbf{x}_i + \gamma_0 t_j + \gamma_1 z_i t_j$
- $\text{var}(Y_{ij}) = D_{11} + 2t_j D_{12} + t_j^2 D_{22} + \sigma^2$
- $\text{cov}(Y_{ij}, Y_{ij'}) = D_{11} + (t_j + t_{j'}) D_{12} t_j t_{j'} D_{22}$

When only random intercept b_{i0} is present, $D_{12} = D_{21} = D_{22} = 0$, exchangeable within-cluster correlation

- SAS PROC MIXED
- The marginal (population-average) interpretation of the fixed effects is lost in the GLMM. For example, a simple random intercept logistic model:

$$\mu_{ij} = E\{E(Y_{ij}|b_{i0})\} = \int \frac{\exp(\beta_0^* + \beta_1^* z_i + b_{i0})}{1 + \exp(\beta_0^* + \beta_1^* z_i + b_{i0})} \phi(b_{i0}; \sigma_b^2) db_{i0}$$

$$\neq \frac{\exp(\beta_0^* + \beta_1^* z_i)}{1 + \exp(\beta_0^* + \beta_1^* z_i)}$$

- Neuhaus et al. (1991) found
 - $|\beta_l| \leq |\beta_l^*|, l = 1, \dots, p;$
 - the equality holds if and only if $\beta_l^* = 0$; and
 - the difference between β_l and β_l^* increases as $D = \text{var}(\mathbf{b}_i)$ (or the main diagonal elements) increases.
- Zeger et al. (1988) found in the random intercept logistic model,

$$\beta_l \approx \frac{\beta_l^*}{\sqrt{1 + 0.346 D_{11}}}, l = 1, \dots, p.$$

MODEL SPECIFICATION

(A) Given \mathbf{b}_i 's, the responses Y_{i1}, \dots, Y_{in_i} are mutually independent, and

$$Y_{ij} | \mathbf{b}_i \sim \text{DM}(\mu_{ij}^b, \sigma^2)$$

$$\eta_{ij}^b = g(\mu_{ij}^b) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{b}_i$$

(B) The random effects, $\mathbf{b}_1, \dots, \mathbf{b}_m$ are *i.i.d* with a multivariate density $f(\mathbf{b}; \mathbf{D})$.

- Typically, $f = \text{MVN}$ and $\mathbf{D} = \text{cov}$.
- Typically, $\mathbf{D} = \mathbf{D}(\boldsymbol{\tau})$ with an unknown vector $\boldsymbol{\tau}$ of variance components

Goal: Estimate fixed effect $\boldsymbol{\beta}$, random effects \mathbf{b}_i 's and variance components $\boldsymbol{\tau}$.

LIKELIHOOD FUNCTION

$$\mathbf{b} = (\mathbf{b}_1^\top, \mathbf{b}_2^\top, \dots, \mathbf{b}_N^\top)^\top$$

$$\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \dots, \mathbf{y}_N^\top)^\top, \text{ with } \mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})^\top$$

Augmented likelihood based on (\mathbf{y}, \mathbf{b}) , $L(\boldsymbol{\beta}, \boldsymbol{\tau}) \propto$

$$|\mathbf{D}(\boldsymbol{\tau})|^{-N/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^N \sum_{j=1}^{n_i} d(y_{ij}, \mu_{ij}^{b_i}) - \frac{1}{2} \sum_{i=1}^N \mathbf{b}_i^\top \mathbf{D}^{-1}(\boldsymbol{\tau}) \mathbf{b}_i \right\}.$$

Marginal log-likelihood based on the \mathbf{Y}

$$\ell(\boldsymbol{\beta}, \boldsymbol{\tau}) \propto -\frac{N}{2} \log |\mathbf{D}(\boldsymbol{\tau})| + \log \int e^{-\kappa(\mathbf{b})} d\mathbf{b},$$

$$\kappa(\mathbf{b}) = \frac{1}{2\sigma^2} \sum_{i=1}^N \sum_{j=1}^{n_i} d(y_{ij}, \mu_{ij}^{b_i}) + \frac{1}{2} \sum_{i=1}^N \mathbf{b}_i^\top \mathbf{D}^{-1}(\boldsymbol{\tau}) \mathbf{b}_i.$$

Technical issue: How to evaluate the integral (may be high-dimensional)

- Need to evaluate the integral

$$\begin{aligned} \int_{\mathcal{R}} h(b) \frac{e^{-b^2/(2\tau)}}{\sqrt{2\pi\tau}} db &= \int_{\mathcal{R}} h(\sqrt{2\tau}v) \frac{e^{-v^2}}{\sqrt{\pi}} dv \\ &= \int_{\mathcal{R}} h^*(v) e^{-v^2} dv, \\ &\approx \sum_{k=1}^Q h^*(v_k) w_k \end{aligned}$$

where

$$\begin{aligned} h^*(v) &= h(\sqrt{2\tau}v)/\sqrt{\pi}, \text{ with} \\ h(v) &= \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} d(y_{ij}; \mu_{ij}^{b_i}) \right\}. \end{aligned}$$

- v_k and w_k need to be determined by a certain numerical evaluation method.

MLE BASED ON NUMERICAL INTEGRATION

- For one-dimensional random effects, the GLMM is

$$\begin{aligned} Y_{ij}|b_i &\overset{i.i.d.}{\sim} \text{MD}(\mu_{ij}^b, \sigma^2) \\ b_i &\overset{i.i.d.}{\sim} N(0, \tau) \end{aligned}$$

$$\text{with } g(\mu_{ij}^b) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + b_i.$$

- Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau, \sigma^2)$. The resulting likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^N \int_{\mathcal{R}} \prod_{j=1}^{n_i} f(y_{ij}|b_i) f(b_i) db_i \\ &= \prod_{i=1}^N \left\{ \prod_{j=1}^{n_i} c(y_{ij}; \sigma^2) \right\} \int_{\mathcal{R}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} d(y_{ij}; \mu_{ij}^b) \right\} \frac{e^{-b_i^2/(2\tau)}}{\sqrt{2\pi\tau}} db_i \end{aligned}$$

GAUSSIAN QUADRATURE

- Gaussian Quadrature is usually used in the evaluation of integrals against certain probability measure ξ .
- With some pre-selected abscissas x_k 's and weights w_k 's,

$$\int_{-\infty}^{\infty} f(x) d\xi(x) \approx \sum_{k=1}^Q w_k f(x_k).$$

- Basically, the sum of areas of many vertical rectangles approximates the area under the curve $f(\cdot)$.

- In the above trapezoidal rule, if ξ is a uniform measure, then v_k 's are equally spaced. In general, measure ξ is not uniform, so x_k 's are not necessarily uniformly spaced. Typically, x_k 's will be clustered, with more points in regions of high probability.
- An important example in many statistical models and methods is the case ξ is the normal distribution measure. The resulting quadrature method is called *Gaussian-Hermit Quadrature*.

- Gaussian quadrature evaluates this integral:

$$\int h^*(v)e^{-v^2}\sqrt{\pi}dv \approx \sum_{k=1}^Q h^*(v_k)w_k.$$
- How to choose v_k and w_k ? How to choose Q ?
- Practical experience suggests that $Q \geq 20$ is satisfactory.
- The v_k and w_k are determined such that the approximation $\sum_{k=1}^Q h^*(v_k)w_k$ will give the exact answer to the integral with $h^*(\cdot)$ being all orthogonal polynomials up to degrees $(2Q - 1)$.
- This implies that

$$v_k = \text{kth zero (root) of Hermit polynomial } H_n(v)$$

$$w_k = \frac{2^{-1}n!\sqrt{\pi}}{n^2\{H_{n-1}(v_k)\}^2},$$
 where $H_n(\cdot)$ is the Hermit orthogonal polynomial of order n .

MATLAB gives these v_k and w_k of any order Q .

- Abramowitz and Stegun's (1964) Handbook of Mathematical Functions lists the values of v_k and w_k . For example, Table 1 lists these values for the case of $Q = 3, 4, 5$, respectively.

| Q | v_k | w_k |
|-----|-------------|------------|
| 3 | -1.22474487 | 0.29540898 |
| | 0.0 | 1.18163590 |
| | 1.22474487 | 0.29540898 |
| 4 | -1.65068012 | 0.08131284 |
| | -0.52464762 | 0.80491409 |
| | 0.52464762 | 0.80491409 |
| | 1.65068012 | 0.08131284 |
| 5 | -2.02018287 | 0.01995324 |
| | -0.95857246 | 0.39361932 |
| | 0.0 | 0.94530872 |
| | 0.95857246 | 0.39361932 |
| | 2.02018287 | 0.01995324 |

- For example,

$$\int_{-\infty}^{\infty} (1 + v^2)e^{-v^2} dv = \frac{3}{2}\sqrt{\pi} = 2.65868.$$

Using the Gauss-Hermit quadrature method with $Q = 3$, we obtain

$$\begin{aligned} \int_{-\infty}^{\infty} (1 + v^2)e^{-v^2} dv &= \\ \{1 + (-1.22474487)^2\}(0.29540898) &+ \\ +(1 + 0^2)(1.18163590) &+ \\ +(1 + 1.22474487^2)(0.29540898) &= 2.65868, \end{aligned}$$

as expected.

- Consider

$$\int_{-\infty}^{\infty} (1 + v^6)e^{-v^2} dv = \frac{23}{8}\sqrt{\pi} = 5.0958.$$

The same 3-point quadrature formula gives a result of 3.76645, a very poor approximation. However, the 4-point quadrature formula gives an almost exact answer (an exercise).
- Multi-dimensional Case: the multi-dimensional quadrature formula is based on Cartesian product Gaussian-Hermit quadrature rules (Davis and Rabinowitz, 1984), which carries out 1-dimensional Gauss-Hermit allocation. In order to do so, the multivariate normal may first be written either as a product of conditional normals through Cholesky decomposition of the variance-covariance matrix.
- Some caveats for the application of Gauss-Hermit quadrature formula:

- the integrand $h^*(\cdot)$ has to be “centered”. For example,

$$\int e^{2va-a^2} e^{-v^2} dv = \sqrt{\pi}, \text{ for all } a.$$

If we use the 5-point quadrature to evaluate this un-centered integrand, the errors are reported below:

Table 1: Errors in the 5-point quadrature evaluation.

| Error | a |
|--------------|---------------------------|
| 0.0 | $a = 0$ |
| 0.001 | $a = 1$ |
| 0.240 | $a = 2$ |
| $\sqrt{\pi}$ | $a \rightarrow \pm\infty$ |
- Integrand $h^*(\cdot)$ should be kind of smooth function. The approximation can be very poor if the integrand has jumps.

- SAS PROC NLMIXED implements this type of approach (adaptive Gaussian quadrature).
- Limited to low-dimensional random effects, $q \leq 3$.

APPROXIMATE INFERENCE

An approximate likelihood inference: Penalized quasi-likelihood (PQL) for β and \mathbf{b}_i 's and Restricted maximum likelihood (REML) for τ

- Approximate the integral by 2nd order Laplace approximation
- Can be severely biased for binary responses of small clusters
- Allow arbitrary $q = \dim(\mathbf{b}_i)$
- Work for general DM models
- SAS PROC GLIMMIX

Stiratteli, Laird & Ware (1984) Biometrics
 Breslow & Clayton (1993) JASA
 Lin & Breslow (1996) JASA

PQL ESTIMATION

2nd order Laplace's integral approximation leads to

$$\ell(\beta, \tau) \approx -\frac{N}{2} \log |\mathbf{D}(\tau)| - \frac{1}{2} \log |\kappa''(\tilde{\mathbf{b}})| - \kappa(\tilde{\mathbf{b}}),$$

where $\tilde{\mathbf{b}}_i$ are the solutions of

$$\kappa'_{\mathbf{b}_i}(\mathbf{b}) = 0, i = 1, 2, \dots, N$$

Moreover $\ell(\beta, \tau) \approx$

$$-\frac{1}{2} \sum_{i=1}^N \log |\mathbf{I} + \mathbf{Z}_i^\top \mathbf{W}_i \mathbf{Z}_i \mathbf{D}| - \frac{1}{2\sigma^2} \sum_{i=1}^N \sum_{j=1}^{n_i} d(y_{ij}, \mu_{ij}^{b_i}) - \frac{1}{2} \sum_{i=1}^N \tilde{\mathbf{b}}_i^\top \mathbf{D}^{-1}(\tau) \tilde{\mathbf{b}}_i$$

where $\mathbf{W}_i = \text{diag}[\dots w_{ij} \dots]$ with

$$w_{ij} = \frac{\mathbb{E} \left\{ d''_{\mu_{ij}}(y_{ij}, \mu_{ij}^{b_i}) | \mathbf{b}_i \right\}}{2\sigma^2 \left\{ g'(\mu_{ij}^{b_i}) \right\}^2}$$

Breslow and Clayton suggest: estimate $(\hat{\beta}, \hat{\mathbf{b}})$ by jointly maximizing Green's (1984) penalized quasi-likelihood

$$h(\beta, \mathbf{b}) = -\frac{1}{2\sigma^2} \sum_{i=1}^N \sum_{j=1}^{n_i} d(y_{ij}, \mu_{ij}^{b_i}) - \frac{1}{2} \sum_{i=1}^N \mathbf{b}_i^\top \mathbf{D}^{-1}(\tau) \mathbf{b}_i.$$

Differentiation $h(\beta, \mathbf{b})$ with respect to β and \mathbf{b}_i leads to quasi-score equations:

$$\sum_{i=1}^N \sum_{j=1}^{n_i} \frac{\mathbf{x}_{ij}}{\sigma^2 g'(\mu_{ij}^{b_i})} u(y_{ij}; \mu_{ij}^{b_i}) = 0$$

$$\sum_{j=1}^{n_i} \frac{\mathbf{z}_{ij}}{\sigma^2 g'(\mu_{ij}^{b_i})} u(y_{ij}; \mu_{ij}^{b_i}) = \mathbf{D}^{-1}(\tau) \mathbf{b}_i, i = 1, \dots, N,$$

According to Harville (1977, JASA), solutions to above equations are

$$\left(\sum_{i=1}^N \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right) \beta = \sum_{i=1}^N \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{Y}_i,$$

where $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$ with "working" responses

$$Y_{ij} = \eta_{ij}^{b_i} + \frac{2g'(\mu_{ij}^{b_i})}{\mathbb{E} \left\{ d''_{\mu_{ij}}(y_{ij}, \mu_{ij}^{b_i}) | \mathbf{b}_i \right\}} u(y_{ij}; \mu_{ij}^{b_i}).$$

$$\mathbf{V}_i = \mathbf{W}_i^{-1} + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^\top$$

and BLUP

$$\hat{\mathbf{b}}_i = \mathbf{D} \mathbf{Z}_i^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}).$$

$$\text{var}(\hat{\beta}) \approx \left(\sum_{i=1}^N \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1}$$

REML

Breslow and Clayton version REML for τ :

$$\ell_1(\hat{\beta}(\tau), \tau) \approx -\frac{1}{2} \sum_{i=1}^N \log |\mathbf{V}_i| - \frac{1}{2} \log \left| \sum_{i=1}^N \mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i \right| - \frac{1}{2} \sum_{i=1}^N (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta})^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta}).$$

Solve, for $k = 1, \dots, q$,
 $\mathbf{S}(\tau_k) =$

$$\frac{1}{2} \sum_{i=1}^N \left[(\mathbf{Y}_i - \mathbf{X}_i \beta)^\top \mathbf{V}_i^{-1} \frac{\partial \mathbf{V}_i}{\partial \tau_k} \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \beta) - \text{tr} \left(\mathbf{P}_i \frac{\partial \mathbf{V}_i}{\partial \tau_k} \right) \right] = 0,$$

$$\mathbf{P}_i = \mathbf{V}_i^{-1} - \mathbf{V}_i^{-1} \mathbf{X}_i (\mathbf{X}_i^\top \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i^\top \mathbf{V}_i^{-1}.$$

Std err. of REML estimate computed from Fisher info matrix

ESTIMATE OF σ^2

$$\hat{\sigma}^2 = \frac{2 \sum_{i=1}^N \sum_{j=1}^{n_i} s^2(y_{ij}, \hat{\mu}_{ij})}{\sum_{i=1}^N \sum_{j=1}^{n_i} d''(y_{ij}, \hat{\mu}_{ij}) V(\hat{\mu}_{ij})},$$

$s(\cdot)$ is the score residual (Reid, 1995) given by, suppressing indices,

$$s = -d'_\mu V^{\frac{1}{2}}(\mu)/2.$$

Or,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_{ij})^2}{\sum_{i=1}^N n_i}.$$

OPHTHALMOLOGY STUDY REVISITED

- Model:

$$Y_{ij} | \mathbf{b}_i \stackrel{i.i.d.}{\sim} S^-(\mu_{ij}^b, \sigma^2), j = 1, \dots, n_i, i = 1, \dots, N,$$

$$\text{logit}(\mu_{ij}^b) = \beta_0 + b_{0i} + \beta_1 \log(\text{time}_{ij}) + \beta_2 \log^2(\text{time}_{ij}) + \beta_3 \text{gas}_{ij}.$$
- $b_{0i} \stackrel{i.i.d.}{\sim} N(0, \tau_0).$
- Also tried to fit a model with two random effects, but the second one was found insignificant.

| Method | Intercept | log(time) | log ² (time) | Gas | τ_0 | σ^2 |
|-----------------|------------|-------------|-------------------------|------------|-------------|------------|
| PQL/REML | 2.91(0.33) | 0.06(0.34) | -0.35(0.09) | 0.44(0.20) | 0.283(0.03) | 62.9 |
| <i>p</i> -value | < .0001 | .6904 | < .0001 | .0351 | < .0001 | |
| MLE(20) | 3.12(0.32) | -0.13(0.49) | -0.33(0.14) | 0.52(0.25) | 0.991(0.12) | 133.7 |
| MLE(50) | 3.13(0.33) | -0.15(0.51) | -0.32(0.15) | 0.55(0.27) | 0.996(0.12) | 133.9 |
| <i>p</i> -value | < .0001 | .7641 | .0340 | .0393 | < .0001 | < .0001 |
| LMM | 3.46(0.38) | -0.01(0.28) | -0.39(0.07) | 0.65(0.43) | 1.650(0.35) | 1.302 |
| <i>p</i> -value | < .0001 | .9624 | < .0001 | .1398 | | |

INFERENCE VIA EM-ALGORITHM

- Let $\theta = (\beta, \tau)$.

$$L(\mathbf{y}, \mathbf{b}; \theta) = \prod_{i=1}^N \prod_{j=1}^{n_i} p(y_{ij} | \mathbf{b}_i, \beta) f(\mathbf{b}_i; \theta) d\mathbf{b}_i.$$

- The r -th E-step computes

$$Q(\theta | \theta^{r-1}) = E\{\log L(\mathbf{y}, \mathbf{b}; \theta) | \mathbf{y}, \theta^{r-1}\}$$

- The r -th M-step finds θ^r , the maximizer of $Q(\theta | \theta^{r-1})$, such that

$$Q(\theta^r | \theta^{r-1}) \geq Q(\theta | \theta^{r-1})$$

- Difficulty: evaluate the expectation in the E-step.
- Use Monte Carlo approximation to $Q(\theta | \theta^{r-1})$.

MCEM ALGORITHM

MCEM algorithm: Monte Carlo E-step and M-step

Generate a random sample

$$\mathbf{b}^{r-1,1}, \dots, \mathbf{b}_{r-1,M} \sim f(\mathbf{b} | \mathbf{y}, \theta^{r-1})$$

$$Q_M(\theta | \theta^{r-1}) = \frac{1}{M} \sum_{m=1}^M \log L(\mathbf{y}, \mathbf{b}^{r-1,m}; \theta) \xrightarrow{a.s.}_{M \rightarrow \infty} Q(\theta | \theta^{r-1})$$

The rest of inference is asymptotically equivalent to the standard EM algorithm

- No software available to implement this approach
- Programming is *ad hoc*, depending on model specified
- Computational intensive
- Slow convergence and sensitive to initial values

BAYESIAN INFERENCE USING MCMC

- Alternative method of numerical evaluation
- Numerically equivalent to MLE if all priors are chosen to be flat priors
- Software BUGS and CODA (or BOA) for implementation
Free download (including users' manuals) from
www.mrc-bsu.cam.ac.uk/bugs (BUGS)
www.mrc-bsu.cam.ac.uk/bugs (CODA)
www.public-health.uiowa.edu/boa (BOA)
- Computational intensive
- Convergence diagnostics (*Burn-in*) are crucial but non-trivial

Zeger and Karim (1991) for GLMMs

POSTERIOR FOR θ

$$f(\theta | \mathbf{y}) \propto f(\theta, \mathbf{y}) = \prod_{i=1}^N \int f(y_{ij} | \mathbf{b}_i, \beta) f(\mathbf{b}_i | D) f(\delta) d\mathbf{b}_i$$

- Normalizing constant is independent of θ , so estimator of θ , e.g. posterior mode, can be derived from $f(\theta, \mathbf{y})$ alone
- If prior $f(\theta)$ is constant (flat prior), posterior is proportional to the likelihood function, and thus numerically, posterior mode is the same as MLE

PRIORS

- Priors for β : $\beta_i \sim N(0, \tau^2)$, $\tau^2 = \sigma^{-2}$, called precision, with very small τ^2 , e.g. 10^{-4} or 10^{-5}
- Priors for D :
 - (1) When $D = \text{diag}(D_{11}, \dots, D_{qq})$, use Inverse Gamma $1/G_{ii} \sim \Gamma(p, \lambda)$ with small p and λ , e.g. 10^{-3} or 10^{-4}
 - (2) When D a positive definite, use $\Sigma = G^{-1} \sim \text{Wishar}(R, k)$

GIBBS SAMPLER

- Generate samples from joint distribution $f(\mathbf{y}, \mathbf{b}, \boldsymbol{\theta})$, so inference can be drawn using the drawn samples. For example, integral is evaluated by Monte Carlo approximation (LLN)
- However it is difficult to generate samples from high-dimensional distribution
- Gibbs sampler claims that this can be done through low dimensional conditional distributions
- Easy to implement with low-dimensional distribution but computational intensive
- Consider a trivariate case, U, V, W with joint distribution $[U, V, W]$
- Assume easy to sample from each of its conditional distributions, $[U|V, W]$, $[V|U, W]$, and $[W|V, U]$
- Gibbs sampling scheme runs:

- Step 0 Give arbitrary starting values $U^{(0)}, V^{(0)}, W^{(0)}$
- Step 1.1 Draw $U^{(1)} \sim [U|V^{(0)}, W^{(0)}]$
- Step 1.2 Draw $V^{(1)} \sim [V|U^{(1)}, W^{(0)}]$
- Step 1.3 Draw $W^{(1)} \sim [W|U^{(1)}, V^{(1)}]$
- complete iteration 1 to obtain $(U^{(1)}, V^{(1)}, W^{(1)})$
- Step 1.1 Draw $U^{(2)} \sim [U|V^{(1)}, W^{(1)}]$
- Step 1.2 Draw $V^{(2)} \sim [V|U^{(2)}, W^{(1)}]$
- Step 1.3 Draw $W^{(2)} \sim [W|U^{(2)}, V^{(2)}]$
- complete iteration 2 to obtain $(U^{(2)}, V^{(2)}, W^{(2)})$
- Step ...
- Geman & Geman (1984, Annals) showed that

$$[U^{(B)}, V^{(B)}, W^{(B)}] \xrightarrow{B \rightarrow \infty} [U, V, W], \text{ at an exponential rate.}$$

Therefore, after the length of *burn-in*, the empirical distribution of the M sample values,

$$[U^{(B+k)}, V^{(B+k)}, W^{(B+k)}], k = 1, \dots, M$$

- Gelfand and Smith (1990) suggest: use every t (e.g. $t = 50$) value in the sequence to have more nearly independent sample values
- Estimate the marginal distribution of each variable, e.g.

$$\widehat{U} = \frac{1}{M} \sum_{k=1}^M [U|V^{(B+k)}, W^{(B+k)}]$$
- Obtain basic statistics such as mean, standard deviation, median, 0.025 quantile and 0.975 quantile for each of variables.

APPLICATION TO EPILEPTIC SEIZURE DATA

- Example available in WinBUGS Manual
- Data description: Reported by Thall and Vail (1990), the data is collected from a clinical trial of 59 epileptics, which aims to examine the effectiveness of drug progabide treating epileptic seizures. For each patient, the number of epileptic seizures was recorded during a baseline period of eight weeks. Patients were then randomized to two treatment arms, one with the anti-epileptic drug and the other with placebo, in addition to standard chemotherapy. The number of seizures was recorded in four consecutive two-week period intervals. The question is whether or not the progabide reduces the rate of epileptic seizures. The data set is attached at the end of this chapter.
- Now we analyze the epileptic seizures data using WinBUGS software. Assume that the epileptic seizure counts $Y_{ij}|\mathbf{b}_i \sim PO(\mu_{ij}^b)$,

- Log-linear random effects model:

$$\begin{aligned} \mu_{ij}^b &= \log E(y_{ij}|\mathbf{b}_i) = \beta_0 + \beta_1 \log(Base_i/4) + \beta_2 Trt_i \\ &\quad + \beta_3 (Trt_i * \log(Base_i/4)) \\ &\quad + \beta_4 \log(Age)_i + \beta_5 V4_j + \\ &\quad b_{0i} + b_{ij}, \\ &j = 1, 2, 3, 4, i = 1, \dots, 59 \\ b_{0j} &\sim Normal(0, \tau_{b_0}) \text{ with } \tau_{b_0} = 1/\eta_{b_0} \\ b_{ij} &\sim Normal(0, \tau_b) \text{ with } \tau_b = 1/\eta_b \end{aligned}$$

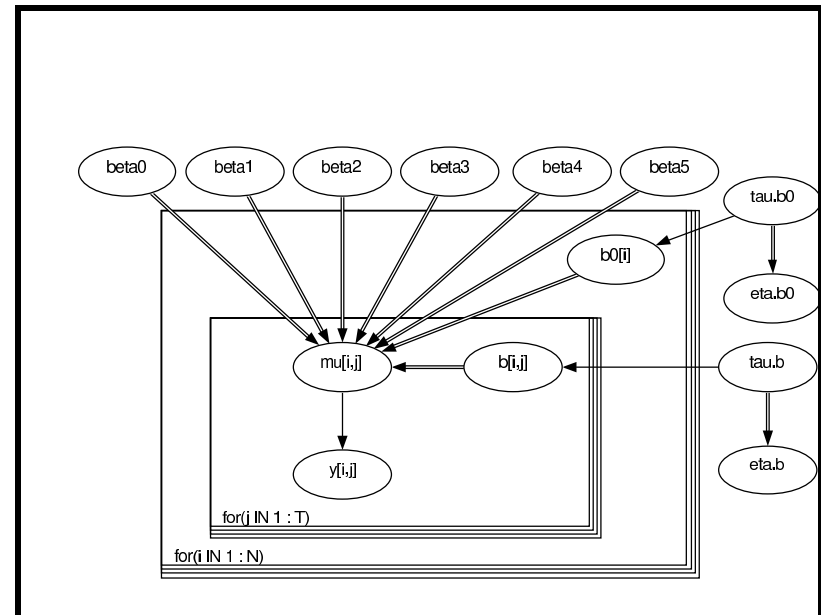
Covariate $Base_i$ is the baseline count of seizures for subject i , and

$$V4_j = \begin{cases} 0 & \text{for } j = 1, 2, 3 \\ 1 & \text{for } j = 4 \end{cases}$$

$$Trt_i = \begin{cases} 1 & \text{if } i\text{-th subject in progabide group} \\ 0 & \text{if } i\text{-th subject in placebo group.} \end{cases}$$

Here, b_{0i} and b_{ij} are assumed to be independent normal random effects with respective variance (precision) parameters η_{b_0} (or τ_{b_0}) and η_b (or τ_b). Using WinBUGS Doodle graphic tool, we may present the above model specification in the following figure.

- Doodle Graphic model specification:



- Explanation of the graph:
 - (a) The nodes outside of the outer plate represent the variables or parameters whose samples will be drawn by the Gibbs sampler.
 - (b) Different plates indicate different classes of loops.
 - (c) The double-arrowed line indicates a deterministic relation, whereas the single-arrowed line indicates a stochastic relation as long as the MCMC sampling concerns.
- Moreover, Doodle can translate this graphic representation into a BUGS programming code.
- In order to improve the poor convergence properties of burn-in due to highly correlated Markov chain, it seems necessary to standardize each covariate about its mean to ensure approximate prior independence between the regression coefficients.

- For example, covariate Age in the model will be centralized as $\text{Age} - \overline{\text{Age}}$. This will result in an adjustment on the estimate of the intercept term given by

$$\beta_0 = \beta_{new,0} - \beta_1 \overline{\log.\text{Base4}} - \beta_2 \overline{\text{Trt}} - \beta_3 \overline{\text{Base4} * \text{Trt}} - \beta_4 \overline{\text{Age}} - \beta_5 \overline{\sqrt{4}}.$$
- We ran 15,000 updates including a burn-in of 5,000 updates with and without the invocation of over-relaxation scheme, respectively.
- Discarding the first 5,000 updates, in the following table we list the estimates by the WinBUGS summary statistics of parameters for log-linear mixed models, without over-relaxation.

| Par. | Mean | Stdev | MC Err | 2.5% | Median | 97.5% |
|--------------|---------|--------|--------|---------|---------|---------|
| β_0 | -1.4210 | 1.2540 | 0.0510 | -3.9660 | -1.3990 | 1.0310 |
| β_1 | 0.8806 | 0.1396 | 0.0107 | 0.5825 | 0.8830 | 1.1530 |
| β_2 | -0.9319 | 0.4157 | 0.0365 | -1.7570 | -0.9490 | -0.1474 |
| β_3 | 0.3394 | 0.2104 | 0.0192 | -0.0454 | 0.3529 | 0.7535 |
| β_4 | 0.4921 | 0.3721 | 0.0162 | -0.2230 | 0.4843 | 1.2630 |
| β_5 | -0.1032 | 0.0857 | 0.0018 | -0.2692 | -0.1028 | 0.0647 |
| η_{b_0} | 0.4976 | 0.0716 | 0.0024 | 0.3685 | 0.4929 | 0.6504 |
| η_b | 0.3642 | 0.0439 | 0.0017 | 0.2843 | 0.3622 | 0.4548 |

- Conclusions judged by the respective 95% nominal confidence intervals:
 - (a) The 95% nominal interval is $(-1.7570, -0.1474)$, which does not contain 0. This implies that the treatment can help to reduce the number of seizures by a medium size $\exp(-0.949) = 0.39$.
 - (b) Another useful covariate to predict the average number of seizures is $\log.\text{Base4}$.
 - (c) All of the other covariates did not indicate strong impact.

Part IV

Vector Generalized Linear Models for Correlated Data

121

To proceed, we assume that we know

- (i) **data types** of the marginal components $y_{ij}, j = 1, \dots, m$, such as binary, counts, positive continuous, continuous variables.
 - m outcomes have **the same type**, or
 - m outcomes have **mixed types**
- (ii) Vector θ contains a set of parameters pertaining to **marginal characteristics** and parameters pertaining to **correlation profiles**.

MOTIVATION

Often encounter a regression data of the form

$$(\mathbf{y}_i, \mathbf{x}_i), i = 1, \dots, N$$

where \mathbf{y}_i is an m -variate response vector,

$$\mathbf{y}_i = (y_{i1}, \dots, y_{im})^T \sim F(\mathbf{y}; \theta).$$

Objectives: Formulate a regression model and establish a statistical inference for the model parameter θ .

University of Michigan

122

Peter Song

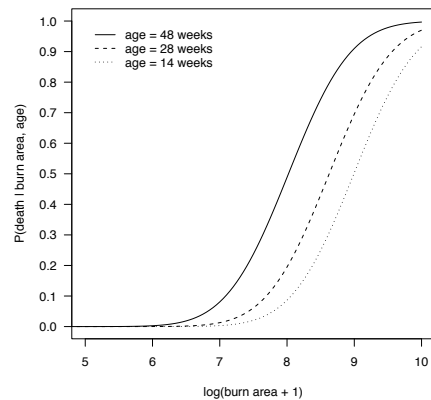
Burn Injury Data

- Reported by Fan and Gijbels (1996), 981 cases of burn injuries with two response variables: Severity of injury Y_1 measured as burn area (continuous) and disposition of death (yes/no) Y_2 .
- A common covariate **age**.
- Separate one-dimensional regression analysis:

$$\begin{aligned} E(Y_1 | \text{age}) &= \beta_{01} + \beta_{11} \text{age} \\ \text{logit}(Pr(Y_2 = 1 | \text{age})) &= \beta_{02} + \beta_{12} \text{age} \end{aligned}$$

- More powerful to fit the two models simultaneously because Y_1 and Y_2 are highly correlated.
- Advantage: Obtain conditional distribution of one variable given the other.

Figure 1: Conditional densities of death given burn severity across different age cohorts.



When data types are known, the marginal GLM is appealing. That is, for a given component,

$$y_{ij} \sim \text{GLM}_j(\mu_j(\mathbf{x}_i), \varphi_j), i = 1, \dots, N$$

with

$$h_j(\mu_j(\mathbf{x}_i)) = \eta_j(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_j.$$

Here GLM_j denotes an exponential dispersion model distribution for the j component, with mean μ_j and dispersion parameter φ_j , and the density function given by

$$g_j(y_j; \mu_j, \varphi_j) = c(y_j; \varphi_j) \exp[\{\theta_j y_j - \kappa_j(\theta_j)\} / \varphi_j], y_j \in \mathcal{R}, \theta_j \in \Theta_j.$$

Therefore, the marginal parameters include all $\boldsymbol{\beta}_j$'s and φ_j 's, $j = 1, \dots, m$.

How about correlation parameters in $F(\mathbf{y}, \boldsymbol{\theta})$?

Still an open problem!

Because we don't know what the F is.

A desired F should have some basic properties:

- P1. The marginal parameters and the correlation parameters are 'orthogonal (or unconstrained)', as in the case of multivariate normal;
- P2. It is **reproducible or marginally closed**; marginal distributions of F should have the same distribution type as the F ;
- P3. The correlation parameters permit both **positive and negative correlations**;
- P4. It accommodates **mixed outcomes**.

Some existing multivariate models in the literature:

- **Bahadur's representation**: A distributions for multivariate binary outcomes.

P1 doesn't hold

- **Log-linear representation**: A distributions for multivariate binary outcomes.

P1 and P2 don't hold

- **Latent variable model approach** for multivariate categorical outcomes, including multivariate binary outcomes.

P4 doesn't hold, e.g. mixed outcomes of binary and counts

- Several *ad hoc* ways to define multivariate Poisson distributions such as the method of **stochastic representation**.

P3 doesn't hold

Currently popular solutions: GEEs and GLMMs.

- the GEEs approach primarily detours this problem **as it does not specify the F** .
GEE improves efficiency in comparison to the independence case, yet does not address if the resultant efficiency improvement is enough.
- the GLMM approach specifies the F via a mixture model, but in general it is **hard to see and interpret the resulting correlation structure directly**.
Not an ideal model when correlation is of primary interest in data analysis.

A reason for their popularity is that both approaches are developed in a **unified framework** for various marginal distributions (data types).

MULTIVARIATE DISPERSION MODELS: GAUSSIAN COPULA APPROACH

Let the marginal CDF of ED(μ_j, φ_j) be $G_j(y_j; \mu_j, \varphi_j)$ or simply $G_j(y_j)$. Sklar (1959) suggested that a joint CDF with m different ED margins takes the form

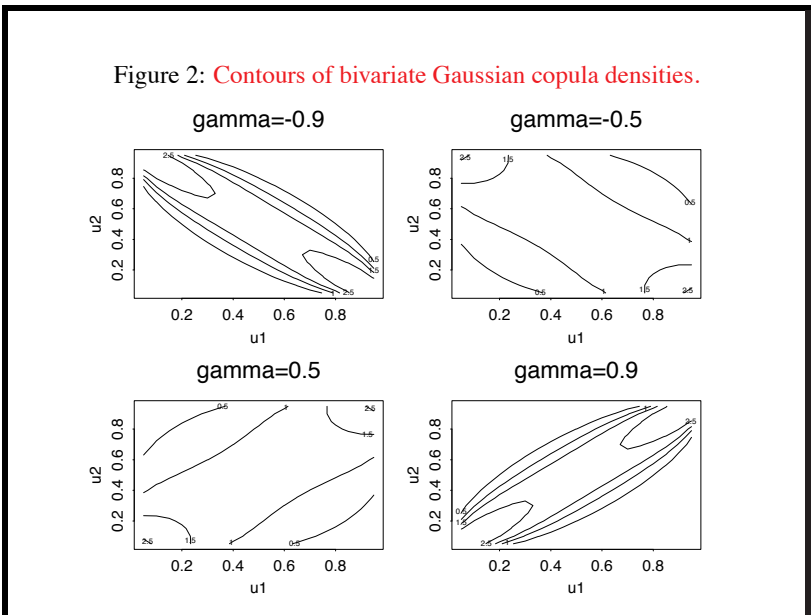
$$F(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\varphi}, \Gamma) = C \{G_1(y_1; \mu_1, \varphi_1), \dots, G_m(y_m; \mu_m, \varphi_m) | \Gamma\},$$

with $C(\cdot)$ is the m -variate Gaussian copula given by

$$C(\mathbf{u} | \Gamma) = \Phi_m \{ \Phi^{-1}(u_1), \dots, \Phi^{-1}(u_m) | \Gamma \}, (u_1, \dots, u_m)^T \in (0, 1)^m$$

where $\Phi_m(\phi_m)$ and $\Phi(\phi)$ are the respective CDFs (densities) of m -variate normal $N_m(0, \Gamma)$ with a correlation matrix Γ and the standard univariate normal $N(0, 1)$.

Such an F satisfies properties P1-P4.



Dependence matrix $\Gamma = [\gamma_{jj'}]_{m \times m}$ with

$$\gamma_{jj} = 1;$$

$$\gamma_{ij} = \text{corr} [\Phi^{-1}\{G_i(y_i)\}, \Phi^{-1}\{G_j(y_j)\}]$$

- It represents the **linear correlation of normal scores** with continuous y_i and y_j ;
- It is the **polychoric correlation** of Anderson and Pemberton (1985) with binary y_i and y_j ;
- More details, see Song (2007) “Correlated Data Analysis: Modeling, Analytics and Applications”.

When all margins are continuous, the m -variate density is

$$f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\varphi}, \Gamma) = c \{G_1(y_1), \dots, G_m(y_m) | \Gamma\} \prod_{i=1}^m g_i(y_i; \mu_i, \varphi_i)$$

where $c(\cdot)$ is the density of $C(\cdot)$ given by

$$c(\mathbf{u} | \Gamma) = |\Gamma|^{-1/2} \exp \left\{ \frac{1}{2} \mathbf{q}^T (I_m - \Gamma^{-1}) \mathbf{q} \right\}$$

with normal scores vector $\mathbf{q} = (q_1, \dots, q_m)^T$,

$$q_i = \Phi^{-1}(u_i), i = 1, \dots, m.$$

When all margins are DISCRETE, the m -variate probability function is

$$f(\mathbf{y}) = \sum_{j_1=1}^2 \dots \sum_{j_m=1}^2 (-1)^{j_1+\dots+j_m} C(u_{1,j_1}, \dots, u_{m,j_m} | \Gamma)$$

where $u_{j,1} = G_j(y_j -)$ and $u_{j,2} = G_j(y_j)$. Here $G_j(y_j -)$ is the left-hand limit of G_j at y_j .

When margins are mixed outcomes: m_1 continuous margins and the rest m_2 discrete margins, the density is

$$f(\mathbf{y}) = \left\{ \prod_{j=1}^{m_1} g_j(y_j) \right\} \sum_{j_{m_1+1}=1}^2 \dots \sum_{j_m=1}^2 (-1)^{j_{m_1+1}+\dots+j_m} \times C_1^{m_1}(G_1(y_1), \dots, G_{m_1}(y_{m_1}), u_{m_1+1, j_{m_1+1}}, \dots, u_{m, j_m} | \Gamma).$$

SIMULTANEOUS MLE

Suppose the regression data follow the multivariate ED distributions above:

$$\mathbf{y}_i | X_i \sim \text{MED}_m(\boldsymbol{\mu}_i, \boldsymbol{\varphi}_i, \Gamma), i = 1, \dots, N.$$

Want to establish a simultaneous MLE for model parameters

$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\varphi}, \Gamma).$$

- Here a common $\boldsymbol{\beta}$ for all component is assumed. In the case of mixed outcomes, different $\boldsymbol{\beta}_j$'s seem to be more reasonable.
- Possibly $\Gamma = \Gamma(\boldsymbol{\alpha})$. Then $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\varphi}, \boldsymbol{\alpha})$.

The log-likelihood function of $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}; Y, X) = \sum_{i=1}^N \ell_i(\boldsymbol{\theta}; \mathbf{y}_i, X_i),$$

and the MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; Y, X).$$

- Numerically, we find $\hat{\boldsymbol{\theta}}$ via the simplex algorithm or a Gauss-Newton type algorithm, with no need of the 2nd order derivatives.

- We estimate the observed Fisher information by

$$\hat{\mathbf{i}} = \mathbf{A}_N^{-1}(\hat{\boldsymbol{\theta}})\mathbf{B}_N(\hat{\boldsymbol{\theta}})\mathbf{A}_N^{-1}(\hat{\boldsymbol{\theta}}),$$
 where $\mathbf{A}_N(\theta)$ is the numerical Hessian and

$$\mathbf{B}_N(\theta) = \frac{1}{N} \sum_{i=1}^N \dot{\ell}_i(\hat{\boldsymbol{\theta}}; \mathbf{y}_i, X_i) \dot{\ell}_i(\hat{\boldsymbol{\theta}}; \mathbf{y}_i, X_i)^T.$$
 Consistently estimate the standard errors even if the model is misspecified.

A GAUSS-NEWTON TYPE ALGORITHM

The $(k + 1)^{th}$ iteration proceeds as:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \delta \{\mathbf{B}(\boldsymbol{\theta}^k)\}^{-1} \dot{\ell}(\boldsymbol{\theta}^k),$$

where δ is the step-halving term:

- starting at 1, it halves each time until

$$\ell(\boldsymbol{\theta}^{k+1}) > \ell(\boldsymbol{\theta}^k)$$
 holds in one iteration.
- Stops when the increase in the likelihood is no longer possible or the difference between two consecutive updates is smaller than a pre-specified precision level.

COMPARISON OF ASYMPTOTIC RELATIVE EFFICIENCY

- Compare ARE of the MLE to that of GEEs estimator for 3-variate binary and 3-variate count data.
- Simply consider exchangeable correlation structure.
- Comparison focuses only on the ARE of $\boldsymbol{\beta}$, as the GEEs treats correlation parameter α as a nuisance parameter.

- The ARE is given by

$$\text{ARE}(\boldsymbol{\beta}) = \text{diag}\{V_{\text{vglm}}\} [\text{diag}\{V_{\text{gee}}\}]^{-1}.$$
 where $\text{var}_{\text{gee}} = \mathbf{I}_0^{-1}(\boldsymbol{\theta})\mathbf{I}_1(\boldsymbol{\theta})\mathbf{I}_0^{-1}(\boldsymbol{\theta})$, with

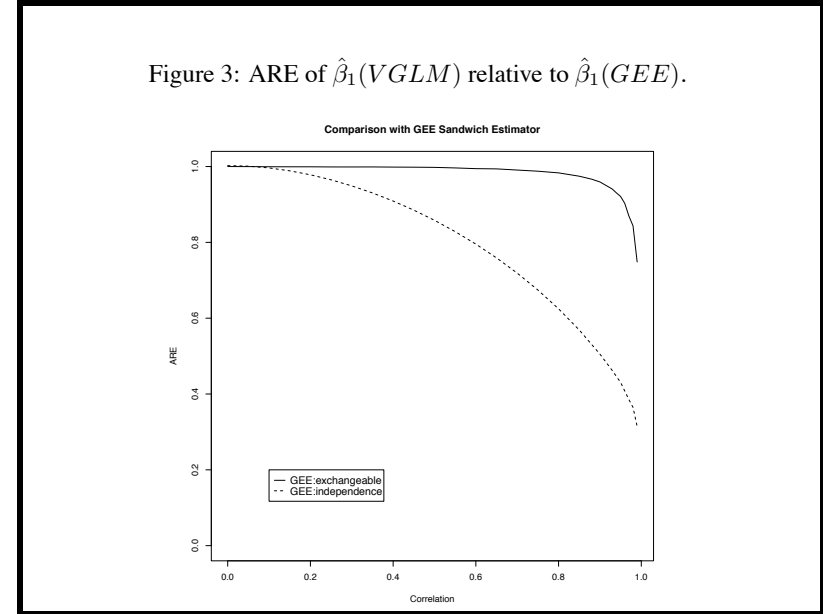
$$\mathbf{I}_0(\boldsymbol{\theta}) = \sum_{i=1}^N \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}, \mathbf{I}_1(\boldsymbol{\theta}) = \sum_{i=1}^N \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\mu}_i} \mathbf{V}_i^{-1} \text{cov}(\mathbf{y}_i) \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}.$$
- $\text{cov}(\mathbf{Y}_i)$ is the covariance matrix of \mathbf{y}_i , usually estimated by $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T$.
- Working covariance matrix

$$\mathbf{V}_i = \varphi \mathbf{A}_i^{\frac{1}{2}} R(\alpha) \mathbf{A}_i^{\frac{1}{2}},$$
 where \mathbf{A}_i is an $m \times m$ diagonal matrix $\text{diag}\{v(\mu_{ij}), j = 1, \dots, m\}$, and $R(\alpha)$ is a working correlation matrix.

LOGIT MODEL FOR 3-VARIATE BINARY DATA

- Fitzmaurice et al. (1993): A hypothetical clinical trial collects a binary response repeatedly over 3 periods.
- At each period, placebo ($x_t = 0$) or an active drug ($x_t = 1$) is randomly administered, and all 8 possible treatment configurations have equal prob of occurrence.
- Model for the marginal expectation is

$$\text{logit}(\mu_{it}) = \beta_0 + \beta_1 x_{it} + \beta_2(t - 1), \quad t = 1, 2, 3,$$
 with $\beta_0 = 0, \beta_1 = \beta_2 = 0.5$.
- No need of simulating data, the ARE can be derived analytically from respective formulae.

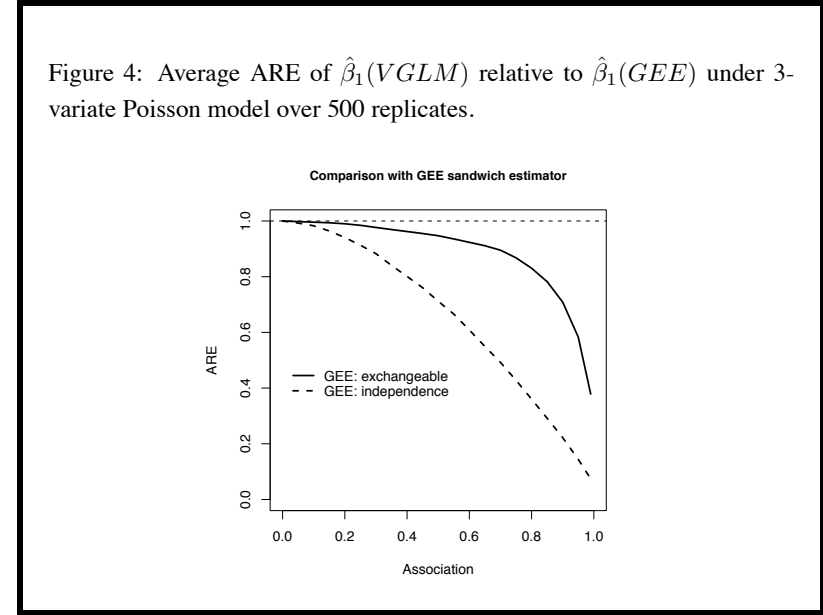


POISSON MODEL FOR 3-VARIATE COUNT DATA

Marginally, each component follows a log-linear model:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}, \quad j = 1, 2, 3, i = 1, \dots, N.$$

- True values $\beta_0 = \beta_1 = 0.5$
- Covariate x_{i1} were generated randomly according to uniform $U(0, 1)$.
- $N = 500$, so large that the var_{vglm} can achieve good precision.
- Need to estimate β 's, α , var_{vglm} and var_{gee} .
- Average ARE was yielded over 500 replicates.



DATA EXAMPLE I: 2-PERIOD CROSSOVER TRIAL

- Re-analyze the data of example 8.1 from Diggle et al. (2002).
- A cross-over trial to compare an active drug (A) and a placebo (B) on cerebrovascular deficiency.
- Response $Y_{ij} = 1$ indicates a normal electrocardiogram reading and 0 otherwise for individual i at drug administration period j .
- $N = 67$ individuals involved in the trial.
- Diggle et al. (2002) used the log-linear representation model approach; hence the std err was calculated by the model-based asymptotic covariance matrix.
- Model for $\mu_{ij} = P(Y_{ij} = 1)$:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij1} x_{ij2}, \quad j = 1, 2, i = 1, \dots, 67,$$

with treatment (x_1), period (x_2) and interaction ($x_1 x_2$).

$$\hat{\alpha} = 0.89.$$

| Variable | VGLM | | GEE | |
|---------------------------|----------------------|-------|----------------------|-------|
| | $\hat{\beta}$ (s.e.) | Z | $\hat{\beta}$ (s.e.) | Z |
| Intercept | .43 (.36) | 1.20 | .43 (.36) | 1.21 |
| Treatment (x_1) | 1.17 (.59) | 1.98 | 1.11 (.57) | 1.93 |
| Period (x_2) | .17 (.51) | .32 | .18 (.51) | .35 |
| Interaction ($x_1 x_2$) | -1.09 (.98) | -1.11 | -1.20 (.98) | -1.04 |

DATA EXAMPLE II: MIXED OUTCOMES OF BURN INJURY DATA

- Two response variables: Total burn area Y_1 and disposition of death (Y_2)
- One covariate of interest: age
- Marginal models:

$$\mu_{i1} = \beta_{01} + \beta_{11} x_i = \mathbf{x}_1^T \beta_1$$

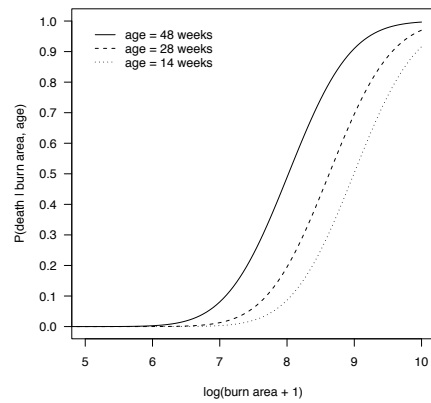
$$\text{logit}(\mu_{i2}) = \beta_{02} + \beta_{12} x_i = \mathbf{x}_2^T \beta_2,$$

with different regression coefficients (β_1 and β_2), and different link functions (the identity and the logit).

$$\hat{\alpha} = 0.80.$$

| Model | β | VGLM | | Univariate Models | |
|-------------------|-----------|---------------|--------|-------------------|--------|
| | | $\hat{\beta}$ | Z | $\hat{\beta}$ | Z |
| Linear (Severity) | Intercept | 6.6980 | 139.73 | 6.7118 | 97.24 |
| | Age | .0039 | 3.16 | .0035 | 1.97 |
| Logit (Death) | Intercept | -4.0521 | -24.44 | -3.6891 | -17.78 |
| | Age | .0527 | 19.13 | .0046 | 11.07 |

Figure 5: Conditional densities of death given burn severity across different age cohorts.



CONCLUDING REMARKS

- A unified framework to analyze nonnormal vector regression data in which a simultaneous MLE is developed and implemented.
- With the availability of the likelihood function, it is quite handy to establish likelihood ratio tests for a variety of hypothesis testing problems, which will be demonstrated in QTL analysis.
- We wrote C++ codes for all simulation studies and data analyses, and the Gauss-Newton type algorithm works very fast.
- The computational burden is mainly on the evaluation of the CDF of m -variate normal.
- As a parametric method, as always, model validation is necessary.