Statistical Methods for Correlated Data

General Regression Models

Michele Guindani

Department of Biostatistics
UCLA

- GLMs can be extended to incorporate dependencies in obs on the same unit
- Two ways to do this:
  - conditional models: the dependences are introduced through modeling of unit-specif random effects:
    - Generalized linear mixed models (GLMMs)
    - Bayesian methods
  - marginal models: first- and second- moment assumptions only
    - Generalized estimating equations

Ex: Contraception Data

- A randomized longitudinal contraception trial
- 1,151 women received a low and a high dose of DepoProvera, a drug use for contraceptive purposes, on the day of randomization and three additional injections at 90-day intervals. There was a final follow-up 3 months after the last injection (a year after the initial injection).
- Outcome: Amenorrhea yes/no, the absence of menstrual bleeding for a specified number of days, during each of the four 3-months intervals.
- Different # of measurements at different occasions
- Individual spaghetti-plot not informative for binary data; better to plot averages over time (approx probability of amenorrhea over time for the two treatment groups)

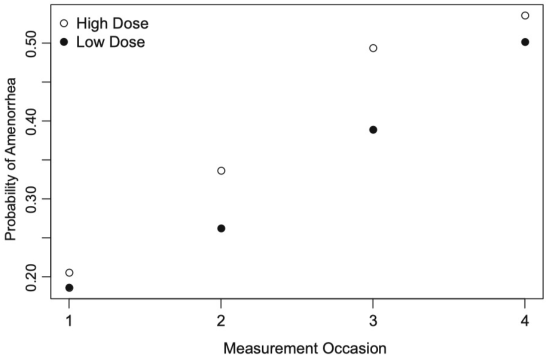


Fig. 9.1 Probability of amenorrhea over time in low- and high-dose groups, in the contraception data

Increasing probabilities of amenorrhea in both groups, with the probabilities in the high dose group being greater than in the low dose group.

Ex: Contraception Data

- With binary data, no clear measure of dependence

Table 9.1 Empirical variances (on the *diagonal*) and correlations (on the *upper diagonal*), between measurements on the same woman at different observation occasions (1–4), in the low- (*left*) and high- (*right*) dose groups of the contraception data

	1	2	3	4		1	2	3	4
1	0.15	0.40	0.28	0.27	1	0.16	0.31	0.25	0.29
2		0.19	0.45	0.35	2		0.22	0.43	0.43
3			0.24	0.13	3			0.25	0.47
4				0.25	4				0.25

- Correlation between observations on the same woman, with a suggestion that the correlations decrease on measurements taken further apart

- Multivariate binary data are needed to take into account within-subject correlation

Seizure data

- Thall and Vail (1990) describe data on epileptic seizures in 59 individuals.

- For each patient, the number of epileptic seizures was recorded during a baseline period of 8 weeks, after which patients were randomized to one of two groups: treatment with either the antiepileptic drug progabide or with placebo.
- The number of seizures was recorded in four consecutive 2-week periods

- $m_{placebo} = 28$ and $m_{progabide} = 31$
- Let Y_{ij} represent the number of counts for patient $i, i = 1, \ldots, 59$ at occasion j, with $j = 0$ the baseline period and $j = 1, \ldots, 4$ the subsequent set of four 2 -week measurement periods.
- Let T_0 be the baseline period and $T_j = 2$ for $j = 1, 2, 3, 4$.
- See R

Generalized Linear Mixed Models (GLMMs)

- A modeling framework that allows the introduction of random effects in the GLM framework (Breslow and Clayton, 1993)
- A GLMM is defined by the following two-stage model:
- Stage One (conditional): The distribution of the data is $Y_{ij}|\theta_{ij}, \phi \sim p(\cdot)$ where $p(\cdot)$ is a member

$$p\left(y_{ij}|\theta_{ij}, \phi\right) = \exp\left\{\left[y_{ij}\theta_{ij} - b\left(\theta_{ij}\right)\right]/a(\phi) + c\left(y_{ij}, \phi\right)\right\}$$

for $i = 1, \ldots, m$ units and $j = 1, \ldots, n_i$, measurements per unit. The variance is

$$\operatorname{var}\left(Y_{ij}|\theta_{ij}, \phi\right) = \phi v\left(\mu_{ij}\right)$$

with ϕ a dispersion parameter and $v\left(\mu_{ij}\right)$ indicating how the variance is functionally related to the mean response

Generalized Linear Mixed Models (GLMMs)

- A modeling framework that allows the introduction of random effects in the GLM framework (Breslow and Clayton, 1993)
- A GLMM is defined by the following two-stage model:
- Stage One (conditional): The distribution of the data is $Y_{ij}|\theta_{ij}, \phi \sim p(\cdot)$ where $p(\cdot)$ is a member

$$p\left(y_{ij}|\theta_{ij}, \phi\right) = \exp\left\{\left[y_{ij}\theta_{ij} - b\left(\theta_{ij}\right)\right]/a(\phi) + c\left(y_{ij}, \phi\right)\right\}$$

for $i = 1, \ldots, m$ units and $j = 1, \ldots, n_i$, measurements per unit. The variance is

$$\text{var}\left(Y_{ij}|\theta_{ij}, \phi\right) = \phi v\left(\mu_{ij}\right)$$

with ϕ a dispersion parameter and $v\left(\mu_{ij}\right)$ indicating how the variance is functionally related to the mean response

Generalized Linear Mixed Models (GLMMs)

- Let $\mu_{ij} = \mathrm{E}\left[Y_{ij} | \theta_{ij}, \alpha\right]$ and, for a link function $g(\cdot)$, suppose

$$g\left(\mu_{ij}\right) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_{i}$$

  (sistematic component) $g(\cdot)$ is called the link function

Generalized Linear Mixed Models (GLMMs)

- Let $\mu_{ij} = \mathrm{E}\left[Y_{ij}|\theta_{ij}, \alpha\right]$ and, for a link function $g(\cdot)$, suppose

$$g\left(\mu_{ij}\right) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i$$

  (sistematic component) $g(\cdot)$ is called the link function

- State two: The random effects are assigned a normal distribution:

$$\boldsymbol{b}_i|\boldsymbol{D} \sim_{iid} \mathbf{N}_{q+1}(\mathbf{0}, \boldsymbol{D})$$

- Given b_i, $Y_{ij} \perp\!\!\!\perp Y_{ik}$ (conditional independence assumption

Example: Poisson GLMMs

- $Y_{iJ} \left| b_i \sim Poisson \left(\mu_{ij} \right) \quad \mu_{ij} = E \left(y_{ij} | b_i \right) \right.$
- ☞ $E \left(Y_{ij} | b_i \right) = \mu_{ij}$, $\mathrm{Var} \left(Y_{ij} | b_i \right) = \mu_{ij}$
- log-link

$$g \left(\mu_i \right) = \log \left(\mu_{ij} \right)$$

$$\log \left(\mu_{ij} \right) = \boldsymbol{x}_{ij} \boldsymbol{\beta} + \boldsymbol{z}_{ij} \boldsymbol{b}_i$$

Example: Poisson GLMMs

- $Y_{iJ}|b_i \sim Poisson(\mu_{ij}) \quad \mu_{ij} = E(y_{ij}|b_i)$
- ☞ $E(Y_{ij}|b_i) = \mu_{ij}$, $\text{Var}(Y_{ij}|b_i) = \mu_{ij}$
- log-link

$$g(\mu_i) = \log(\mu_{ij})$$

$$\log(\mu_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i$$

- How to interpret the coefficients?
- In GLMMs, the regression parameters $\boldsymbol{\beta}$ have subject-specific interpretations:
- ☞ They represent the influence of covariates on a subject-specific response
- ☞ regression coefficients represent the effect of within-subject changes in covariates and changes in an individual's transformed mean response

Example: Poisson GLMMs

- $\log\left(\mu_i\right) = \beta_1 + \beta_2 t_{ij} + b_{1i} \quad b_{1i} \sim N\left(0, \sigma_b^2\right)$
- ☞ $\mu_i = e^{\beta_1 + \beta_2 t_{ij} + b_{1i}}$
- e^{β_1} : mean response for an individual with $b_{1i} = 0$ (typical individual) ($\neq$ mean response in the population)
- e^{β_2} : multiplicative change in mean response (rate of occurrence) within an individual for a unit increase in time (e.g. $\beta_2 = 1.36$(36% increase); $\beta_2 = .75$ (25%decrease)
- ☞ β_2 represents the change in the log-expected rate for a single unit increase in the predictor (time) within an individual (or with respect to a typical individual, where "typical" typically means $\boldsymbol{b}_i = \boldsymbol{0}$).

Binary GLMMs

- $y_{i_j}|b_i \sim \text{Bern}\left(\pi_{ij}\right)$, i.e. $y_{ij} = \begin{cases} 0 & \text{with prob } 1-\pi_{ij} \\ 1 & \text{with prob } \pi_{ij} \end{cases}$
- $E\left(Y_{ij}|b_i\right) = \pi_{ij}$, $Var\left(y_{ij}|b_i\right) = \pi_{ij}\left(1-\pi_{ij}\right)$
- $g(\mu_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = logit(\pi_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i$
- Example: $\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_1 + \beta_2 t_{ij} + b_{1i} + b_{2i} t_{ij}$

☞ $\underbrace{\frac{\pi_{ij}}{1-\pi_{ij}}}_{odds} = e^{\beta_1+\beta_2 t_{ij}+b_{1i}+b_{2i}t_{ij}}$

- Interpretation:
- e^{β_1} : odds of $Y_i = 1$ at baseline for a typical subj $(b_{ji} = 0)$
- e^{β_2} : multiplicative change in odds of $y_{1,j} = 1$ for a "typical subject" for one unit interval in time

- However in GLMMs it is harder to interpret fixed effects coefficients for covariates that are time-invariant (e.g., treatment/placebo; gender; etc...)
- In those cases you may want to compare two “typical individuals”, i.e. two individuals with the same random effects (e.g. a typical individual under treatment vs a typical individual under placebo)

- However in GLMMs it is harder to interpret fixed effects coefficients for covariates that are time-invariant (e.g., treatment/placebo; gender; etc...)
- In those cases you may want to compare two "typical individuals", i.e. two individuals with the same random effects (e.g. a typical individual under treatment vs a typical individual under placebo)
- In GLMMs, covariate effects are generally different at the marginal and individual level:
- GLMMs are typically parameterized in terms of covariate effects on conditional means
  - ☞ within-subject or subject-specific changes in covariates

- However in GLMMs it is harder to interpret fixed effects coefficients for covariates that are time-invariant (e.g., treatment/placebo; gender; etc...)
- In those cases you may want to compare two "typical individuals", i.e. two individuals with the same random effects (e.g. a typical individual under treatment vs a typical individual under placebo)
- In GLMMs, covariate effects are generally different at the marginal and individual level:
- GLMMs are typically parameterized in terms of covariate effects on conditional means
  - ☞ within-subject or subject-specific changes in covariates
- Differently than in LMEs, marginal effects are not easy to calculate. For example, the marginal mean is:

$$E(\boldsymbol{Y}_i) = E_{b_i}\left(g^{-1}(\boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i)\right) = \dots$$

Marginal moments in GLMMs

- The marginal variance is

$$\begin{aligned}\text{var}(Y_{ij}) &= \text{E}\left[\text{var}(Y_{ij}|\boldsymbol{b}_i)\right] + \text{var}\left(\text{E}\left[Y_{ij}|\boldsymbol{b}_i\right]\right) \\ &= \alpha \text{E}_{b_i}\left[v\left\{g^{-1}\left(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i\right)\right\}\right] + \text{var}_{b_i}\left[g^{-1}\left(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i\right)\right]\end{aligned}$$

- The covariances between outcomes on the same unit are

$$\begin{aligned}\text{cov}(Y_{ij}, Y_{ik}) =& \text{E}\left[\text{cov}(Y_{ij}, Y_{ik}|\boldsymbol{b}_i)\right] + \text{cov}\left[\text{E}\left(Y_{ij}|\boldsymbol{b}_i\right], \text{E}\left[Y_{ik}|\boldsymbol{b}_i\right)\right] \\ &= \text{cov}_{b_i}\left[g^{-1}\left(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i\right), g^{-1}\left(\boldsymbol{x}_{ik}\boldsymbol{\beta} + \boldsymbol{z}_{ik}\boldsymbol{b}_i\right)\right] \\ &\neq 0\end{aligned}$$

  for $j \neq k$ due to shared random effects,

☞ β's can only be interpreted with reference to an individual subject profile

☞ Interpret the effects on "typical" subjects ($\boldsymbol{b}_i = \boldsymbol{0}$) or as effect sizes that only apply to subjects with the same $\boldsymbol{b}_i$ values

Example: Probit model

- $Y_{ij}|\mu_{ij} \sim$ Bernoulli (μ_{ij}), with $\Phi^{-1}(\mu_{ij}) = \eta_{ij}$ where $\Phi^{-1}(\cdot)$ is the inverse standard normal cumulative distribution function or probit link.

Example: Probit model

- $Y_{ij}|\mu_{ij} \sim$ Bernoulli (μ_{ij}), with $\Phi^{-1}(\mu_{ij}) = \eta_{ij}$ where $\Phi^{-1}(\cdot)$ is the inverse standard normal cumulative distribution function or probit link.
- The probit model can be written as:

$$\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \xi_{ij}, \quad \xi_{ij} = \boldsymbol{z}'_{ij}\boldsymbol{b}_i + e_{ij}$$

where ξ_{ij} is the "total residual" or random part (conditioning on $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$) of the model with variance

$$\sigma_{i,jj} \equiv \boldsymbol{z}'_{ij}\boldsymbol{D}\boldsymbol{z}_{ij} + \sigma_e^2$$

Example: Probit model

- The marginal cumulative response probabilities become

$$\begin{aligned}
\Pr\left(Y_{ij} = 1 | \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}\right) &= \Pr\left(\eta_{ij} > 0 | \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}\right) \\
&= \Pr\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \xi_{ij} > 0 | \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}\right) \\
&= \Pr\left(-\xi_{ij} \leq \boldsymbol{x}'_{ij}\boldsymbol{\beta} | \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}\right) \\
&= \Pr\left(\frac{\xi_{ij}}{\sqrt{\sigma_{i,jj}}} \leq \frac{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}{\sqrt{\sigma_{i,jj}}} | \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}\right) \\
&= \Phi\left(\frac{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}{\sqrt{\sigma_{i,jj}}}\right)
\end{aligned}$$

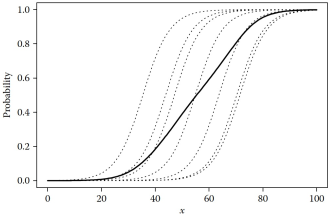
- Here the probit link is preserved for the marginal probabilities but with different regression coefficients; the marginal effects $\beta/\sqrt{\sigma_{i,jj}}$ are attenuated, or closer to zero, compared to the conditional effects β.

Remarks:

- For the log link, it can be shown that the link function is also preserved with

$$E\left(Y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}\right) = \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{D}\boldsymbol{z}_{ij}/2\right)$$

- Apart from the probit and log link, the link function for the marginal model is generally different from that for the conditional model and usually does not have a simple form.
- In logistic models, the marginal regression coefficients are attenuated compared to the conditional ones



Solid: marginal; Dashed: conditional

Remark: Bad notation

Generalized linear mixed models are often written as non-linear models with an error term. For instance, mixed logit or probit models for binary responses are written as

$$Y_{ij} = \pi_{ij} + \epsilon_{ij} z_{ij}^{(1)}, \quad \pi_{ij} = h\left(\eta_{ij}\right), \quad z_{ij}^{(1)} = \sqrt{\pi_{ij}\left(1 - \pi_{ij}\right)}$$

Constraining the variance of ϵ_{ij} to one, we obtain the required Bernoulli variance $\pi_{ij}\left(1 - \pi_{ij}\right)$. However, this formulation is awkward because it requires a very peculiar distribution of ϵ_{ij} to produce valid 0 or 1 responses. More importantly, the formulation gives the false impression that the variance of ϵ_{ij} could be estimated and should therefore be avoided.