Biostatistics Workshop 2008

$\sim$ Longitudinal Data Analysis $\sim$

Session 5

GARRETT FITZMAURICE

Harvard University

Extensions of Generalized Linear Models to Longitudinal Data (Part 1)

When the response variable is categorical (e.g., binary and count data), generalized linear models (e.g., logistic regression) can be extended to handle the correlated outcomes.

However, non-linear transformations of the mean response (e.g., logit) raise additional issues concerning the interpretation of the regression coefficients.

Different approaches for accounting for the correlation lead to models having regression coefficients with distinct interpretations.

As we will see, different models for discrete longitudinal data have somewhat different targets of inference.

MOTIVATING EXAMPLE

Oral Treatment of Toenail Infection

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

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

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

Interested in the rate of decline of the proportion of patients with onycholysis over time and the effects of treatment on that rate.

GENERALIZED LINEAR MODELS FOR LONGITUDINAL DATA

Next, we focus on two general approaches for extending generalized linear models to correlated data:

1. Marginal Models

2. Generalized Linear Mixed Models

The main emphasis will be on discrete response data, e.g., count data or binary responses.

Before discussing these approaches, we briefy review main features of generalized linear models for a univariate outcome.

Brief Review of Generalized Linear Models

Generalized linear models are a class of regression models; they include the standard linear regression model but also many other important models:

- Linear regression for continuous data
- Logistic regression for binary data
- Loglinear/Poisson regression models for count data

Generalized linear models extend the methods of regression analysis to settings where the outcome variable can be categorical.

Later (Sessions 5 and 6), we consider extensions of generalized linear models to longitudinal data.

Notation for Generalized Linear Models

Assume m independent observations of a single response variable, Y_i.

Associated with each response, Y_i, there is a $p \times 1$ vector of covariates, $X_{i1}, ..., X_{ip}$.

Goal: Primarily interested in relating the mean of Y_i, $\mu_i = E(Y_i|X_{i1}, ..., X_{ip})$, to the covariates.

In generalized linear models:

(i) the distribution of the response is assumed to belong to a family of distributions known as the exponential family, e.g., normal, Bernoulli, binomial, and Poisson distributions.

(ii) A transformation of the mean response, μ_i, is then linearly related to the covariates, via an appropriate link function:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where link function $g(\cdot)$ is a known function, e.g., $\log(\mu_i)$.

Mean and Variance of Exponential Family Distributions

Exponential family distributions share some common statistical properties.

The variance of Y_i can be expressed in terms of

$$\text{Var}\,(Y_i) = \phi\, v(\mu_i),$$

where the scale parameter $\phi > 0$.

The variance function, $v(\mu_i)$, describes how the variance of the response is functionally related to μ_i, the mean of Y_i.

Link Function

The link function applies a transformation to the mean and then links the covariates to the transformed mean,

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

where link function $g(\cdot)$ is known function, e.g., $\log(\mu_i)$.

This implies that it is the transformed mean response that changes linearly with changes in the values of the covariates.

Canonical link and variance functions for the normal, Bernoulli, and Poisson distributions.

Distribution	**Var. Function,** $v(\mu)$	**Canonical Link**
Normal	$v(\mu) = 1$	Identity: $\mu = \eta$
Bernoulli	$v(\mu) = \mu(1-\mu)$	Logit: $\log\left[\frac{\mu}{(1-\mu)}\right] = \eta$
Poisson	$v(\mu) = \mu$	Log: $\log(\mu) = \eta$

where $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$.

Common Examples

Normal distribution:

If we assume that $g(\cdot)$ is the identity function,

$$g(\mu) = \mu$$

then

$$\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

gives the standard linear regression model, with $\text{Var}(Y_i) = \phi$.

Note: Variance is unrelated to the mean.

Bernoulli distribution:

For the Bernoulli distribution, $0 < \mu_i < 1$, so we would prefer a link function that transforms the interval $[0, 1]$ on to the entire real line $(-\infty, \infty)$:

$$\text{logit} : \ln\left[\mu_i / \left(1 - \mu_i\right)\right]$$
$$\text{probit} : \Phi^{-1}\left(\mu_i\right)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

If we assume a logit link function then

$$\log\left[\frac{\mu_i}{(1-\mu_i)}\right] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

yields logistic regression model, with $\text{Var}\left(Y_i\right) = \mu_i(1 - \mu_i)$ (Bernoulli variance).

Poisson distribution:

For the Poisson distribution, $\mu_i > 0$, so we would prefer a link function that transforms the interval $(0, \infty)$ on to the entire real line $(-\infty, \infty)$.

If we assume a log link function then

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

yields Poisson or loglinear regression model, with $\text{Var}(Y_i) = \mu_i$ (Poisson variance).

Summary

In generalized linear models:

(i) response assumed to have exponential family distribution, e.g., normal, Bernoulli, binomial, and Poisson distributions.

(ii) transformed mean response is linearly related to the covariates, via an appropriate link function:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

MARGINAL MODELS FOR LONGITUDINAL DATA

The basic premise of marginal models is to make inferences about population averages.

The term 'marginal' is used here to emphasize that the mean response modelled is conditional only on covariates and not on other responses or random effects.

A feature of marginal models is that the models for the mean and the 'within-subject association' (e.g., covariance) are specified separately.

Notation

Let Y_{ij} denote response variable for i^{th} subject on j^{th} occasion.

Y_{ij} can be continuous, binary, or a count.

We assume there are n_i repeated measurements on the i^{th} subject and each Y_{ij} is observed at time t_{ij}.

Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates, X_{ij}.

Covariates can be time-invariant (e.g., gender) or time-varying (e.g., time since baseline).

Features of Marginal Models:

The focus of marginal models is on inferences about population averages.

The marginal expectation, $\mu_{ij} = E\left(Y_{ij}|X_{ij}\right)$, of each response is modelled as a function of covariates.

Specifically, marginal models have the following three part specification:

1. The marginal expectation of the response, μ_{ij}, depends on covariates through a known link function

$$g(\mu_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_p X_{pij}.$$

2. The marginal variance of Y_{ij} depends on the marginal mean according to

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

where $v(\mu_{ij})$ is a known 'variance function' and ϕ is a scale parameter that may need to be estimated.

3. The 'within-subject association' among the responses is a function of the means and of additional parameters, say α, that may also need to be estimated.

For example, when α represents pairwise correlations among responses, the covariances among the responses depend on $\mu_{ij}(\beta)$, ϕ, and α:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \text{s.d.}(Y_{ij})\,\text{Corr}(Y_{ij}, Y_{ik})\,\text{s.d.}(Y_{ik}) \\ &= \sqrt{\phi\, v\,(\mu_{ij})}\,\text{Corr}(Y_{ij}, Y_{ik})\,\sqrt{\phi\, v\,(\mu_{ik})} \end{aligned}$$

where s.d.(Y_{ij}) is the standard deviation of Y_{ij}.

In principle, can also specify higher-order moments.

Examples of Marginal Models

Example 1. Continuous responses:

1. $\mu_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$.
   (i.e., linear regression)

2. $\text{Var}\,(Y_{ij}) = \phi$
   (i.e., homogeneous variance)

3. $\text{Corr}\,(Y_{ij}, Y_{ik}) = \alpha^{|k-j|}$ $(0 \leq \alpha \leq 1)$
   (i.e., autoregressive correlation)

Example 2. Binary responses:

1. $\text{Logit}\,(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$.
   (i.e., logistic regression)

2. $\text{Var}\,(Y_{ij}) = \mu_{ij}\,(1 - \mu_{ij})$
   (i.e., Bernoulli variance)

3. $\text{OR}\,(Y_{ij}, Y_{ik}) = \alpha_{jk}$
   (i.e., unstructured odds ratios)
   where

$$\text{OR}\,(Y_{ij}, Y_{ik}) = \frac{\Pr(Y_{ij} = 1, Y_{ik} = 1)\Pr(Y_{ij} = 0, Y_{ik} = 0)}{\Pr(Y_{ij} = 1, Y_{ik} = 0)\Pr(Y_{ij} = 0, Y_{ik} = 1)}.$$

Example 3. Count data:

1. $\text{Log}\,(\mu_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$.
   (i.e., Poisson regression)

2. $\text{Var}\,(Y_{ij}) = \phi\,\mu_{ij}$
   (i.e., extra-Poisson variance or "overdispersion" when $\phi > 1$)

3. $\text{Corr}\,(Y_{ij}, Y_{ik}) = \alpha$
   (i.e., compound symmetry correlation)

Interpretation of Marginal Model Parameters

The regression parameters, β, have 'population-averaged' interpretations (where 'averaging' is over all individuals within subgroups of the population):

- describe effect of covariates on the average responses

- contrast the means in sub-populations that share common covariate values

$\Longrightarrow$ Marginal models are most useful for population-level inferences.

The regression parameters are directly estimable from the data.

Of note, nature or magnitude of within-subject association (e.g., correlation) does not alter the interpretation of β.

For example, consider the following logistic model,

$$\text{logit}(\mu_{ij}]) = \text{logit}(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}.$$

Each element of β measures the change in the log odds of a 'positive' response per unit change in the respective covariate, for sub-populations defined by fixed and known covariate values.

The interpretation of any component of β, say β_k, is in terms of changes in the transformed mean (or "population-averaged") response for a unit change in the corresponding covariate, say X_{ijk}.

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

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

$$\beta_0 + \beta_1 X_{ij1} + \cdots + \beta_k x + \cdots + \beta_p X_{ijp}.$$

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

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

$$\beta_0 + \beta_1 X_{ij1} + \cdots + \beta_k (x+1) + \cdots + \beta_p X_{ijp}.$$

$\longrightarrow \beta_k$ is change in log odds for subgroups of the study population (defined by any fixed values of $X_{ij1}, ..., X_{ij(k-1)}, X_{ij(k+1)}, ..., X_{ijp}$).

Statistical Inference for Marginal Models

Maximum Likelihood (ML):

Unfortunately, with discrete response data there is no simple analogue of the multivariate normal distribution.

In the absence of a "convenient" likelihood function for discrete data, there is no unified likelihood-based approach for marginal models.

Alternative approach to estimation - *Generalized Estimating Equations* (GEE).

GENERALIZED ESTIMATING EQUATIONS

Avoid making distributional assumptions about Y_i altogether.

Potential Advantages:

Empirical researcher does not have to be concerned that the distribution of Y_i closely approximates some multivariate distribution.

It circumvents the need to specify models for the three-way, four-way and higher-way associations (higher-order moments) among the responses.

It leads to a method of estimation, known as generalized estimating equations (GEE), that is straightforward to implement.

The GEE approach has become an extremely popular method for analyzing discrete longitudinal data.

It provides a flexible approach for modelling the mean and the pairwise within-subject association structure.

It can handle inherently unbalanced designs and missing data with ease (albeit making strong assumptions about missingness).

GEE approach is computationally straightforward and has been implemented in existing, widely-available statistical software.

The GEE estimator of β solves the following *generalized estimating equations*

$$\sum_{i=1}^{m} D_i' V_i^{-1} (y_i - \mu_i) = 0,$$

where V_i is the so-called "working" covariance matrix.

By "working" covariance matrix we mean that V_i approximates the true underlying covariance matrix for Y_i.

That is, $V_i \approx \text{Cov}(Y_i)$, recognizing that $V_i \neq \text{Cov}(Y_i)$ unless the models for the variances and the within-subject associations are correct.

$D_i = \partial \mu_i / \partial \beta$ is the "derivative" matrix (of μ_i with respect to the components of β); $D_i(\beta)$ transforms from the original units of Y_{ij} (and μ_{ij}) to the units of $g(\mu_{ij})$.

Therefore the generalized estimating equations depend on both β and α.

Because the generalized estimating equations depend on both β and α, an iterative two-stage estimation procedure is required:

1. Given current estimates of α and ϕ, an estimate of β is obtained as the solution to the 'generalized estimating equations'

2. Given current estimate of β, estimates of α and ϕ are obtained based on the standardized residuals,

$$r_{ij} = \left(Y_{ij} - \widehat{\mu}_{ij}\right) / v\left(\widehat{\mu}_{ij}\right)^{1/2}$$

For example, ϕ can be estimated by

$$1/\left(mn-p\right)\sum_{i=1}^{m}\sum_{j=1}^{n}r_{ij}^{2}$$

The correlation parameters, α, can be estimated in a similar way.
For example, unstructured correlations, $\alpha_{jk} = \mathrm{Corr}\left(Y_{ij}, Y_{ik}\right)$, can be estimated by

$$\widehat{\alpha}_{jk} = \left(1/(m-p)\right)\widehat{\phi}^{-1}\sum_{i=1}^{m}r_{ij}r_{ik}$$

Finally, in the two-stage estimation procedure we iterate between steps 1) and 2) until convergence has been achieved.

Properties of GEE estimators

$\widehat{\beta}$, the solution to the generalized estimating equations, has the following properties:

1. $\widehat{\beta}$ is consistent estimator of β

2. In large samples, $\widehat{\beta}$ has a multivariate normal distribution

3. $\text{Cov}(\widehat{\beta}) = B^{-1}MB^{-1}$
   where

$$B = \sum_{i=1}^{m} D_i' V_i^{-1} D_i$$

$$M = \sum_{i=1}^{m} D_i' V_i^{-1} \mathrm{Cov}\left(Y_i\right) V_i^{-1} D_i$$

B and M can be estimated by replacing α, ϕ, and β by their estimates, and replacing $\mathrm{Cov}\left(Y_i\right)$ by $\left(Y_i - \widehat{\mu}_i\right)\left(Y_i - \widehat{\mu}_i\right)'$.

Note: We can use this empirical or so-called 'sandwich' variance estimator even when the covariance has been misspecified.

Summary

The GEE estimators have the following attractive properties:

1. In many cases $\widehat{\beta}$ is almost efficient when compared to MLE.
   For example, GEE has same form as likelihood equations for multivariate normal models and also certain models for discrete data

2. $\widehat{\beta}$ is consistent even if the covariance of Y_i has been misspecified

3. Standard errors for $\widehat{\beta}$ can be obtained using the empirical or so-called 'sandwich' estimator

Case Study: *Oral Treatment of Toenail Infection*

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

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

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

Interested in the rate of decline of the proportion of patients with onycholysis over time and the effects of treatment on that rate.

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

$$\text{logit}\left\{E(Y_{ij})\right\} = \beta_0 + \beta_1 \text{Month}_{ij} + \beta_2 \text{Trt}_i * \text{Month}_{ij}$$

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

Here, we assume that $\text{Var}(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$.

We also assume an unstructured correlation for the within-subject association (i.e., estimate all possible pairwise correlations).

Table 1: GEE estimates and standard errors (empirical) from marginal logistic regression model for onycholysis data.

PARAMETER	ESTIMATE	SE	Z
INTERCEPT	-0.698	0.122	-5.74
Month	-0.140	0.026	-5.36
Trt $\times$ Month	-0.081	0.042	-1.94

Results

From the output above, we would conclude that:

1. There is a suggestion of a difference in the rate of decline in the two treatment groups (P = 0.052).

2. Over 12 months, the odds of infection has decreases by a factor of 0.19 [exp(-0.14*12)] in treatment group A.

3. Over 12 months, the odds of infection has decreases by a factor of 0.07 [exp(-0.221*12)] in treatment group B.

4. Odds ratio comparing 12 month decreases in risk of infection between treatments A and B is approx 2.6 (or $e^{12*0.081}$).

5. Overall, there is a significant decline over time in the prevalence of onycholysis for all randomized patients.

Summary of Key Points

The focus of marginal models is on inferences about population averages.

The regression parameters, β, have 'population-averaged' interpretations (where 'averaging' is over all individuals within subgroups of the population):

- describe effect of covariates on marginal expectations or average responses

- contrast means in sub-populations that share common covariate values

$\Longrightarrow$ Marginal models are most useful for population-level inferences.

Marginal models should not be used to make inferences about individuals ("ecological fallacy").

STATISTICAL SOFTWARE: GENERALIZED ESTIMATING EQUATIONS

SAS and *Stata*, which are widely available, can perform all analyses presented in these lectures.

Alternative software packages, e.g. *SPSS* and *S-PLUS*, can also be used.

Caveat: Statistical software is constantly evolving.

GEE using PROC GENMOD in SAS

PROC GENMOD in SAS is primarily a procedure for fitting generalized linear models to a single response.

However, PROC GENMOD has incorporated an option for implementing GEE approach using a REPEATED statement (similar to PROC MIXED).

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

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

Table 2: Illustrative commands for a marginal logistic regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;

   CLASS id group;

   MODEL y=group time group*time / DIST=BINOMIAL LINK=LOGIT;

   REPEATED SUBJECT=id / WITHINSUBJECT=time TYPE=UN;
```

Table 3: Illustrative commands for a marginal logistic regression, with within-subject associations specified in terms of log odds ratios, using PROC GENMOD in SAS.

```
PROC GENMOD DESCENDING;
   CLASS id group;
   MODEL y=group time group*time / DIST=BINOMIAL LINK=LOGIT;
   REPEATED SUBJECT=id / WITHINSUBJECT=time LOGOR=FULLCLUST;
```

Table 4: Illustrative commands for a marginal log-linear regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS.

```
PROC GENMOD;
   CLASS  id  group;
   MODEL  y=group  time  group*time / DIST=POISSON  LINK=LOG;
   REPEATED  SUBJECT=id / WITHINSUBJECT=time  TYPE=UN;
```

GEE using xtgee in Stata

Table 5: Illustrative commands for a marginal logistic regression, with within-subject associations specified in terms of correlations, using xtgee in Stata.

```
. generate grp_time=group*time
. tsset id time
. xtgee y group time grp_time,family(binomial) link(logit) corr(unstr) robust
```

Table 6: Illustrative commands for a marginal log-linear regression, with within-subject associations specified in terms of correlations, using xtgee in Stata.

```
. generate grp_time=group*time
. tsset id time
. xtgee y group time grp_time,family(poisson) link(log) corr(unstr) robust
```

FURTHER READING

Diggle, P.J., Heagerty, P., Liang, K-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford University Press. (See Chapters 7 and 8).

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*. Wiley. (See Chapters 10 and 11).