BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 16

Marginal Models and Generalized Estimating Equations

Marginal Models and Generalized Estimating Equations

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

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

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

$$\mathrm{Var}\left(Y_{ij}|X_{ij}\right) = \phi\, v\left(\mu_{ij}\right)$$

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

   Note: For continuous response, can allow $\mathrm{Var}(Y_{ij}|X_{ij}) = \phi_j v(\mu_{ij})$.

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:

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

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

In principle, can also specify higher-order moments.

Aside: Measures of Association for Binary Responses

With binary responses correlations are not the best choice for modelling the association because they are constrained by the marginal probabilities.

For example, if $E(Y_1) = Pr(Y_1 = 1) = 0.2$ and $E(Y_2) = Pr(Y_2 = 1) = 0.8$, then $\text{Corr}(Y_1, Y_2) < 0.25$.

The correlations must satisfy certain linear inequalities determined by the marginal probabilities.

These constraints are likely to cause difficulties for parametric modelling of the association.

With binary responses, the odds ratio is a natural measure of association between a pair of responses.

The odds ratio for any pair of binary responses, Y_j and Y_k, is defined as

$$OR(Y_j, Y_k) = \frac{Pr(Y_j = 1, Y_k = 1)Pr(Y_j = 0, Y_k = 0)}{Pr(Y_j = 1, Y_k = 0)Pr(Y_j = 0, Y_k = 1)}.$$

Note that the constraints on the odds ratio are far less restrictive than on the correlation.

$\Longrightarrow$ With binary response can model within-subject association in terms of odds ratios rather than correlations.

Examples of Marginal Models

Example 1. Continuous responses:

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

2. $\mathrm{Var}\,(Y_{ij}|X_{ij}) = \phi_j$
   (i.e., heterogeneous variance, but no dependence of variance on mean)

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

Example 2. Binary responses:

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

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

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

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

Example 3. Count data:

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

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

3. 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,

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

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,

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

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

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

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

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

$\longrightarrow$ β_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*

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

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 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

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

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

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

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

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

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

B and M can be estimated by replacing α, ϕ, and β by their estimates, and replacing $\text{Cov}\left(Y_i\right)$ 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 1: ***Clinical Trial of Antibiotics for Leprosy***

Placebo-controlled clinical trial of 30 patients with leprosy at the Eversley Childs Sanitorium in the Philippines.

Participants were randomized to either of two antibiotics (denoted treatment drug A and B) or to a placebo (denoted treatment drug C).

Baseline data on number of leprosy bacilli at 6 sites of body were recorded.

After several months of treatment, number of bacilli were recorded a second time.

Outcome: Total count of number of leprosy bacilli at 6 sites.

Table 1: Mean count of leprosy bacilli at six sites of the body (and variance) pre- and post-treatment.

Treatment Group	Baseline	Post-Treatment
Drug A (Antibiotic)	9.3	5.3
	(22.7)	(21.6)
Drug B (Antibiotic)	10.0	6.1
	(27.6)	(37.9)
Drug C (Placebo)	12.9	12.3
	(15.7)	(51.1)

Question: Does treatment with antibiotics (drugs A and B) reduce abundance of leprosy bacilli when compared to placebo (drug C)?

We consider the following model for changes in the average count

$$\log\, E(Y_{ij}) \;=\; \log\, \mu_{ij} = \beta_1 + \beta_2\, \texttt{time}_{\texttt{ij}} + \beta_3\, \texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{1i}} + \beta_4\, \texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{2i}},$$

where Y_{ij} is count of bacilli for i^{th} patient in j^{th} period ($j = 1, 2$).

$\texttt{trt}_1$ and $\texttt{trt}_2$ are indicator variables for drugs A and B respectively.

The binary variable, `time`, denotes the baseline and post-treatment follow-up periods, with $\texttt{time} = 0$ for the baseline period (period 1) and $\texttt{time} = 1$ for the post-treatment follow-up period (period 2).

To complete specification of the marginal model, we assume

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

where ϕ can be thought of as an overdispersion factor.

Finally, the within-subject association is accounted for by assuming a common correlation,

$$\text{Corr}(Y_{i1}, Y_{i2}) = \alpha.$$

The log-linear regression parameters, β, can be given interpretations in terms of (log) rate ratios.

Table 2: Parameters of the marginal log-linear regression model for the leprosy bacilli data.

Treatment Group	Period	$\log(\mu_{ij})$
Drug A (Antibiotic)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2 + \beta_3$
Drug B (Antibiotic)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2 + \beta_4$
Drug C (Placebo)	Baseline	β_1
	Follow-up	$\beta_1 + \beta_2$

Table 2 summarizes their interpretation in terms of the log expected counts in the three groups at baseline and during post-treatment follow-up.

For example, e^{β_2} is the rate ratio of leprosy bacilli, comparing the follow-up period to baseline, in the placebo group (drug C).

Similarly, $e^{\beta_2+\beta_3}$ is the corresponding rate ratio in the group randomized to drug A.

Finally, $e^{\beta_2+\beta_4}$ is the corresponding rate ratio in the group randomized to drug B.

Thus, β_3 and β_4 represents the difference between the changes in the log expected rates, comparing drug A and B to the placebo (drug C).

Estimated regression coefficients are displayed in Table 3 (with SEs based on "sandwich" estimator).

A test of H$_0$: $\beta_3 = \beta_4 = 0$, produces a (multivariate) Wald statistic, $W^2 = 6.99$, with 2 degrees of freedom ($p < 0.05$).

Note: Magnitudes of effects are similar and indicate that treatment with antibiotics reduces leprosy bacilli.

A test of H$_0$: $\beta_3 = \beta_4$, produces a Wald statistic, $W^2 = 0.08$, with 1 degree of freedom ($p > 0.7$).

Table 3: Parameter estimates and standard errors from marginal log-linear regression model for the leprosy bacilli data.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0801	29.62
$\texttt{time}_{\texttt{ij}}$	−0.0138	0.1573	−0.09
$\texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{1i}}$	−0.5406	0.2186	−2.47
$\texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{2i}}$	−0.4791	0.2279	−2.10

Estimated scale or dispersion parameter: $\widehat{\phi} = 3.45$.
Estimated working correlation: $\widehat{\alpha} = 0.797$.

To obtain a common estimate of the log rate ratio, comparing both antibiotics (drugs A and B) to placebo, we can fit the reduced model

$$\log E(Y_{ij}) \;=\; \log \mu_{ij} = \beta_1 + \beta_2\, \texttt{time}_{\texttt{ij}} + \beta_3\, \texttt{time}_{\texttt{ij}} \times \texttt{trt}_{\texttt{i}},$$

where the variable trt is an indicator variable for antibiotics, with trt $= 1$ if a patient was randomized to either drug A or B and trt $= 0$ otherwise.

We retain the same assumptions about the variance and correlation as before.

The estimated regression coefficients are displayed in Table 4

Table 4: Parameter estimates and standard errors from marginal log-linear regression model for the leprosy bacilli data.

Variable	Estimate	SE	Z
Intercept	2.3734	0.0801	29.62
$\mathtt{time}_{\mathtt{ij}}$	-0.0108	0.1572	-0.07
$\mathtt{time}_{\mathtt{ij}} \times \mathtt{trt}_{\mathtt{i}}$	-0.5141	0.1966	-2.62

Estimated scale or dispersion parameter: $\widehat{\phi} = 3.41$.
Estimated working correlation: $\widehat{\alpha} = 0.780$.

The common estimate of the log rate ratio is -0.5141.

Rate ratio is 0.60 (or $e^{-0.5141}$), with 95% confidence interval, 0.41 to 0.88, indicating that treatment with antibiotics significantly reduces the average number of bacilli when compared to placebo.

For placebo group, there is a non-significant reduction in the average number of bacilli of approximately 1% (or [1 - $e^{-0.0108}$] × 100%).

In the antibiotics group there is a significant reduction of approximately 40% (or [1 - $e^{-0.0108-0.5141}$] × 100%).

Estimated pairwise correlation of 0.8 is relatively large.

Estimated scale parameter of 3.4 indicates substantial overdispersion.

Case Study 2: *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,

$$\text{logit}E(Y_{ij}) = \beta_1 + \beta_2\text{Month}_{ij} + \beta_3\text{Trt}_i * \text{Month}_{ij}$$

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 5: 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 × 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").

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.

GEE and Overdisperson using PROC GENMOD

There are multiple ways to handle overdisperson in a GEE analysis in SAS PROC GENMOD.

1. The default method: PROC GENMOD uses the robust, empirical standard errors, which adjusts for overdisperson automatically because the standard errors are valid even if the variance assumption is misspecified.

2. Use the model standard errors but explicitly estimate the overdisperson parameter.

3. No adjustment: One has the ability in PROC GENMOD to set $\phi = 1$ (no overdisperson). This is not advised because if overdisperson is present the estimated standard errors will be too small.

Table 6: 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 7: 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 8: Illustrative commands for a marginal log-linear regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS. Here overdispersion is accommodated by empirical standard errors (the default)

```
PROC GENMOD;

  CLASS id group;

  MODEL y=group time group*time / DIST=POISSON LINK=LOG;

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

Table 9: Illustrative commands for a marginal log-linear regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS. Here model standard errors are used and the overdisperson parameter is estimated (SCALE=).

```
PROC GENMOD;

   CLASS  id  group;

   MODEL  y=group  time  group*time / DIST=POISSON  LINK=LOG SCALE=PEARSON

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

Table 10: Illustrative commands for a marginal log-linear regression, with within-subject associations specified in terms of correlations, using PROC GENMOD in SAS. Here the assumption is no overdisperson, with ϕ set to be 1.

```
PROC GENMOD;

   CLASS  id  group;

   MODEL  y=group  time  group*time / DIST=POISSON  LINK=LOG SCALE=1;

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