Applied Multivariate and Longitudinal Data Analysis

Longitudinal Data Analysis: Generalized repeated measures:
A. Generalized Estimating Equations (GEE)
B. Generalized Linear Mixed Effects Models (GLMM)

Ana-Maria Staicu

SAS Hall 5220; 919-515-0644; astaicu@ncsu.edu

1 Introduction

Motivating application: Epileptic seizures and chemotherapy study

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

Seizure counts for 5 subjects assigned to placebo (0) and 5 subjects assigned to progabide (1).

	Period						
Subject	*1*	*2*	*3*	*4*	*Trt*	*Baseline*	*Age*
1	5	3	3	3	0	11	31
2	3	5	3	3	0	11	30
3	2	4	0	5	0	6	25
4	4	4	1	4	0	8	36
5	7	18	9	21	0	66	22
				⋮			
29	11	14	9	8	1	76	18
30	8	7	9	4	1	38	32
31	0	4	3	0	1	19	20
32	3	6	1	3	1	10	30
33	2	6	7	4	1	19	18

Questions: Is the mean trend of seizures different across the two treatment groups? The primary objective of the study was to determine whether progabide reduces the rate of seizures in subjects like those in the trial. Additionally does age affect the mean trend over time?
Overview: This chapter deals with analysis of repeated outcomes which are either counts or binary, or more generally non-normal. We discuss modeling approach, estimation, inference of the mean regression parameters.

General setting:

- Outcome of interest: $\{Y_{i1}, \ldots, Y_{im_i}\}$ for unit/subject 'i'
- Y_{ij} either binary, or counts, or rates, etc

- Times of the repeated measurements: $\{t_{i1}, \ldots, t_{im_i}\}$ unit/subject 'i'
- Covariates associates wit the jth measurement of the ith subject

$$X_{ij} = (X_{ij1}, \ldots, X_{ijK})^T;$$

for example $X_{ij1} = t_{ij}$, $X_{ij2} = t_{ij}^2$, $X_{ij3} = Treatment_i$ etc. The covariates describe by X_{ij} could change over time, or not change over time ('time stationary').

Let X_i be $m_i \times K$ matrix obtained by row-stacking X_{ij}.

Objective: Study the effect of the covariates on the mean response trend.

Challenge: The responses are correlated !

Approach: there are few approaches commonly used to model such data. The first approach is to extend the ideas of the GLM studied in last chapter to the setting. Intuition: use GLM to describe the distribution $Y_{ij}|X_{ij}$ - call it 'marginal distribution' , marginal in the sense used by general linear models methodology for normal responses. Then, model the association between the repeated measures in some way.

The models developed in this way are called population-average models or marginal models. These models are used primarily to make inferences about the population means. As a result marginal models for longitudinal data model separately the mean response and the dependence between the repeated measures. The parameter that captures the effect of the covariates on the mean response is the regression parameter and is the parameter of interest. The parameter that describes the data dependence is a nuisance parameter (however this parameter needs to be accounted for in order to make accurate inferences about changes in the population mean response).

2 Population-average models: specification of marginal models

The marginal models describe separately the conditional distribution for each individual response Y_{ij} and the correlation between the repeated measurements. More specifically, the marginal models for longitudinal data have the following 3 part specification:

1. marginal expectation - or the mean for each response Y_{ij}, conditional on the set of covariates X_{ij}, $E[Y_{ij}|X_{ij}] = \mu_{ij}$ depends on the explanatory variables X_{ij} through the link function

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

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

2. marginal variance - or the variance of Y_{ij} conditional on X_{ij} is assumed to depend on the mean function μ_{ij}

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

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

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

The three part specification of the marginal models makes the extension of generalized linear models to longitudinal data transparent.

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

The correlation is a natural measure of the linear dependence for continuous responses; however it is not the common measure to describe the association, otherwise. For example, for continuous responses the correlation can be any value between -1 and 1, and it is independent of the means; this is not the case for discrete cases. As a result with discrete responses, correlation is not the common way to describe association. Instead, the odds ratio (or log odds ratio) is a preferable metric to describe association among binary responses.

Marginal models for

- "Normal" repeated responses

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

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

- Count repeated responses

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

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

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

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

There is an implicit assumption made by the marginal models:

$$E[Y_{ij}|X_i] = E[Y_{ij}|X_{ij}], \text{ where } X_i = (X_{i1}, \ldots, X_{im_i}).$$

This assumption holds for time-invariant covariates (X_{ij} does not vary over j), or for time-varying covariates that are set a priori by study design in a manner completely unrelated to the longitudinal response . However when a time-varying covariate varies over time, this assumption may not hold. For example such assumption would be violated when the current value of Y_{ij} given the current covariate X_{ij} predicts the subsequent value of $X_{i(j+1)}$. For example this may arise in observational studies to assess the effect of physical exercise on reducing the blood glucose level.

The correlation model implied by this model specification is popularly referred to in the context of these models as *working correlation model*. This is because this correlation model carries still a lot of uncertainty. The model is considered only a 'working model' rather than necessarily representing what is probably a very complex truth.

Working correlation matrix. The model for the pairwise correlation is attempting to represent all sources of variation that could lead to associations among the observations:

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

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

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

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

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

Remarks

- In the case where Y_{ij} are normal responses, and furthermore the distribution of Y_i is multivariate normal, then the specification of the conditional mean and variance, (conditional on X_{ij}) for each repeated observation, along with the specification of the correlation $\Gamma_i(\omega)$ identifies the distribution of the vector of responses Y_i (again conditional on X_i).
- This is NOT the case for the vector of generalized responses. To specify the likelihood function for multivariate discrete data, even if the marginal distribution is given additional knowledge about higher order moments are required.

2.1 Statistical inference for marginal models

With discrete response data there is no analogue of the multivariate, normal distribution. Thus there is no 'convenient' likelihood function. Furthermore there is no unified likelihood based framework for marginal models. The estimation is based on an alternative approach called *Generalized Estimating Equations*.

Generalized Estimating Equations (GEE)

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

Connection to previous estimation methods. Recall the estimating equations for the GLM when the response per subject is vector and normally distributed $Y_i = X_i\beta + \epsilon_i$.

$$\sum_{i=1}^{n} X_i^T \Sigma_i^{-1}(Y_i - X_i\beta) = 0$$

The estimating equations for scalar generalized response, $Y_i \sim EF(\eta_i, \phi)$ and $E[Y_i] = \mu_i$ and $g(\mu_i) = \eta_i = X_i\beta$, are:

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

this is equivalent to $\Delta^T V^{-1}(Y - \mu) = 0$. Recall Δ is $n \times K$ matrix of $\partial\mu_i/\partial\beta$ and V is $n \times n$ diagonal matrix with the ith diagonal element equal to $v(\mu_i)$.

Return now to our setting: generalized linear models for repeated measures. Let $Y_i = (Y_{i1}, \ldots, Y_{im_i})$ be the vector of repeated measured for subject/unit i.

- Conditional on X_{ij} the distribution of $Y_{ij} \sim EF(\eta_{ij}, \phi)$

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

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

  where $v(\cdot)$ is the variance (known) function and ϕ is the unknown dispersion parameter

- the correlation of Y_{ij} and $Y_{ij'}$ is a function of additional parameters ω. Let Λ_i be the working covariance matrix $\Lambda_i = \phi\{V(\mu_i)\}^{1/2}\Gamma_i(\omega)\{V(\mu_i)\}^{1/2}$.

The **generalized estimating equations** (GEE) for β are

$$\sum_{i=1}^{n} \Delta_i^T \Lambda_i^{-1}(Y_i - \mu_i) = 0,$$

where

- μ_i is the vector of μ_{ij},
- $\Delta_i = \partial\mu_i/\partial\beta$ is the $m_i \times k$ matrix with the jth row equal to $\partial\mu_{ij}/\partial\beta$
- Λ_i is the working covariance matrix; $\Lambda_i = \phi\{V(\mu_i)\}^{1/2}\Gamma_i(\omega)\{V(\mu_i)\}^{1/2}$.

Remark 1: GEE may be viewed as the minimizer of the objective function $\sum_{i=1}^{n}(Y_i - \mu_i(\beta))^T\Lambda_i^{-1}(Y_i - \mu_i(\beta))$.

Remark 2 that the GEE depends on both β and ω; a one step optimization is challenging. Typically, a two-stage estimation is needed

- Obtain an initial estimator of β by assuming all observations across all individuals are independent;
- Given the current estimate of β, say $\widehat{\beta}$, estimates of ω and ϕ are obtained based on the standardized residuals, and the assumed "working correlation" matrix

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

$$\widehat{\phi} = \frac{1}{N-K}\sum_{i=1}^{n}\sum_{j=1}^{m_i} r_{ij}^2, \qquad \widehat{\omega}_{jj'} = \sum_{i=1}^{n}\frac{r_{ij}r_{ij'}}{\widehat{\phi}n}$$

  where $\widehat{\mu}_{ij} = g^{-1}(X_{ij}\widehat{\beta})$ and N is the total number of observations.
- Given the current estimate of ω and ϕ construct an estimate for Λ_i, $\widehat{\Lambda}_i$. An estimate of β is obtained by a numerical technique which may be viewed as an extensions of the IRWLS method used in the ordinary generalized linear model.

2.2 GEE estimator: Sampling properties

It is important to note that GEE estimator is NOT an ML estimator, as it does not rely on the distributional assumption of Y_i; instead it was derived from an ad-hoc procedure. Nevertheless we can establish theoretical properties of this estimator.

Assuming that the estimators of ω and ϕ are consistent, then $\widehat{\beta}$, the solution of the GEE has the following properties:

- $\widehat{\beta}$ is a consistent estimator of β

- for large samples (large n), $\widehat{\beta}$ has approximately multivariate normal distribution,

$$\widehat{\beta} \sim N\left\{\beta, \phi\left(\sum_{i=1}^{n} \Delta_i^T \Lambda^{-1} \Delta_i\right)^{-1}\right\}.$$

  In practice we estimate Δ_i and Λ_i. Thus $\widehat{V}_{\widehat{\beta}} = \widehat{\phi}\left(\sum_{i=1}^{n} \widehat{\Delta}_i^T \widehat{\Lambda}^{-1} \widehat{\Delta}_i\right)^{-1}$.

What happens with these properties of the GEE estimator of β when the assumption on correlation is incorrect? One solution is to re-evaluate the covariance of the GEE estimator. The robust estimator of the covariance $Var(\widehat{\beta})$ is

- $V_{\widehat{\beta}} = cov(\widehat{\beta}) = F^{-1}GF^{-1}$, where

$$F = \sum_{i=1}^{n} \Delta_i^T \Lambda_i^{-1} \Delta_i$$

$$G = \sum_{i=1}^{n} \Delta_i^T \Lambda_i^{-1} cov(Y_i) \Lambda_i^{-1} \Delta_i$$

where

- F, G depend on the true parameter values ω, ϕ and β ...

- Cov(Y_i) can be estimated by the sample covariance of the residuals as $\widehat{\text{Cov}}(Y_i) = (Y_i - \widehat{\mu}_i)(Y_i - \widehat{\mu}_i)^T$.

Remark: If Cov(Y_i) is estimated using the model based covariance then, the above expression gives the covariance Cov$(\widehat{\beta})$ discussed earlier.

Final Remarks In summary the GEE estimators have the following properties

- The GEE estimator $\widehat{\beta}$ is consistent even when the covariance of Y_i is misspecified. Question: Then why bother accounting for the dependence?

- Standard errors of $\widehat{\beta}$ can be obtained using the empirical or so-called *sandwich* covariance estimator. Remark: The robustness of the empirical sandwich estimator is a large sample property. The empirical sandwich estimator is not appealing in the following situations:
  - the number of subjects is small comparative to the number of repeated observations per subject/unit
  - the sampling design is unbalanced;
  - subj/units cannot be grouped on the basis of having identical covariate design matrices

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

- Hypotheses tests. Reformulate the hypothesis testing as $H_0 : L\beta = h$. Use Wald testing procedures as in hypothesis testing for the ordinary generalized linear models.

  Specifically use the test statistic $\chi^2 = (L\widehat{\beta} - h)^T (L\widehat{V}_{\widehat{\beta}} L^T)^{-1} (L\widehat{\beta} - h)$. Under the null hypothesis, this test statistic has $\chi^2_{\text{number of rows in L}}$

2.3 Model selection for GEE

Quasi-likelihood information criterion (QIC) was developed by Pan(2001) as a modification of the AIC to apply to models fit by GEE.

Let R be a working correlation model (the model on which the working covariance Γ_i is based). Define

$$Q\{\widehat{\beta}(R), \phi\} = \sum_i \sum_j Q\{\widehat{\beta}(R), \phi, Y_{ij}, X_{ij}\}$$

where $Q\{\widehat{\beta}(R), \phi, Y_{ij}, X_{ij}\} = Q_{ij}/\phi$ is the contribution of observation j to the cluster determined by i. Here Q_{ij} are quasi-likelihood functions which differ according to the model used.

Here are some

- Normal data: $Q_{ij} = -[1/(2w_{ij})](Y_{ij} - \mu_{ij})^2$
- Poisson data: $Q_{ij} = w_{ij}(Y_{ij} \log(\mu_{ij}) - \mu_{ij})$
- Binomial data: $Q_{ij} = w_{ij}\{r_{ij} \log(\mu_{ij}) + (n_{ij} - r_{ij}) \log(1 - \mu_{ij})$

where w_{ij} is an a priori specified weight (default $w_{ij} = 1$).

$$QIC(R) = -2Q\{\widehat{\beta}(R), \phi\} + 2Trace\{\widehat{\Omega}_I \widehat{V}_R\}$$

- $\widehat{\Omega}_I$ in the inverse of the model based covariance estimate under the independent working correlation model assumption.
- $\widehat{V}_R$ is the robust estimate of the covariance.

$$QIC_u(R) = -2Q\{\widehat{\beta}(R), \phi\} + 2K\}$$

where K is the number of regression parameters.

Remark: QIC can be used for (dependence/correlation) model selection when the responses are non-normal QIC_u may used for selection of the regression parameters.

2.4 Fitting GEE in R

Recall: population-average or marginal model, provides a regression approach for generalized linear models when the responses are not independent (correlated/clustered data). Goal is to make inferences about the population, accounting for the within-subject correlation.

There are two ways to fit marginal models: (1) the function gee in the `R` package gee (Carey et al., 2012) ; and (2) the function `geeglm` in the `R` package `geepack`

```
gee(Y~ systematic_mean, id=id, data=data, family=binomial, corstr="exchangeable") )
```

```
geeglm(formula, family=gaussian, data, id, zcor=NULL, constr, std.err="san.se")
```

The major difference between gee and `geepack` is that `geepack` contains an ANOVA method that allows us to compare models and perform Wald tests.

Not covered: Generalized linear mixed models (GLMM) which extend the LMM aproach to handle generalized responses.