Mixed versus GEE models

Oksana Pugach, PhD

Institute for Health Research and Policy
University of Illinois at Chicago
May, 2013

Repeated/Clustered Measurements

Two approaches have different targets for inferences and address subtly different questions about longitudinal change

- Marginal (population-average)

  Merely acknowledge the correlation among repeated measurements by robust variance estimation

- Conditional (subject-specific)

  Provide an explanation for the source of correlation at different levels

Linear Model: special case

Model for the mean response vector

$$E\left(Y_i\right) = X_i\beta$$

To account for correlated data:

- Covariance pattern model (autoregressive, Toeplitz)
- Introduce random effect in the model for the mean response

$$E\left(Y_i \mid b_i\right) = X_i\beta + Z_i b_i$$

b_i - random vector with distribution. Different between subjects. Induces within-subject association in outcome.

Linear Model: special case

Interpretation for fixed effect parameters

β has the same interpretation in both models

It describes how the mean response in study population changes with time and how these changes are related to the covariates.

Can be interpreted as either conditional effect or as marginal effect

Why: the marginal effect is derived from the conditional effect by averaging over linear change in individuals

$$E\left(E\left(Y_i \mid b_i\right)\right) = E\left(X_i\beta + Z_i b_i\right) = X_i\beta + Z_i E\left(b_i\right) = X_i\beta = E\left(Y_i\right)$$

Generalized Linear Models

One extra component: link function

Link function relates the mean of Y_i to the linear predictor

In linear models, link function is identity link function

Marginal model for the mean response:

$$g\left(\mu_i\right) = g\left(E\left(Y_i\right)\right) = X_i\beta$$

β is a change in the transformed mean response in the study population (For logistic regression, it is a change in log odds of success in study outcome)

For any known link function there is an inverse link function. For example for logistic model:

$$\mu_i = E\left(Y_i \mid X_i\right) = \frac{\exp\left(X_i\beta\right)}{1+\exp\left(X_i\beta\right)}$$

Note that marginal mean depends on the index *i* only via fixed and known covariate values

Generalized Estimating Equation (GEE) Models

- GEE is extension of GLM for correlated mesures
- The joint distribution of outcome Y_i is not specified
- Instead, only marginal distribution of Y_{ij} at each time point is specified
- Link function is specified (identity, logit, log link)
- Variance is specified as a function of a mean
- "Working" correlation structure is specified (independence, exchangeable, AR(1), m-dependent, unstructured). Matrix is of size $n \times n$, because it is assumed that there are a fixed number of timepoints n that subjects are measured at
- Upon convergence, SE are obtained based on naïve (model-based) or empirical (robust) correlation structures

Generalized Linear Mixed Model

$$g\left(E\left(Y_i \mid X_i, b_i\right)\right) = X_i\beta^* + Z_i b_i$$

models conditional mean of Y_i, given a vector of random effects b_i

β^* have subject-specific interpretation in terms of change in the transformed mean response for any individual. In other words, it's a unit change in the corresponding covariate holding b_i fixed.

The most natural way to hold b_i fixed is to focus on any given individual – most natural for time-varying covariate

We can compare two individuals who have the same values of b_i but who differ by a single unit in the corresponding covariate – most natural for time-invariant covariate.

Marginal vs Conditional Fixed Effect

Special case of logistic regression with random intercept

$$\beta \approx \frac{\beta^*}{\sqrt{1+k^2\sigma_b^2}}$$

coefficient in the marginal model is attenuated relative to corresponding fixed effect in the mixed effect model

$$k^2 = \frac{16\sqrt{3}}{15\pi} = 0.346$$

Marginal β is smaller in absolute value than conditional β^*

For more general model, with a vector of random effects, this relationship holds: marginal β are always attenuated toward zero when compared to β^*

Marginal vs Conditional Fixed Effect

Note:

It is possible to obtain marginal estimates from a generalized mixed effect model. But the assumed form of the distribution of the outcome (logistic, log-linear, etc) for the conditional model no longer holds for the resulting marginal means when averaged over the distribution of random effects.

When one or more covariates are quantitative and/or some confounding is present, no simple summaries of the effects of covariates on μ_i are readily available from generalized linear mixed models (Fitzmaurice, Laird, and Ware, *Applied Longitudinal Analysis*, p. 478)

Example

Two-period crossover study, N=67

Outcome – potential side effect (Y=1 abnornal ECG, Y=0 normal ECG)

Sequence of treatments: P->D; D->P

GEE model:

$$\text{logit}\left(\mu_{ij}\right) = \beta_1 + \beta_2 Treat_{ij} + \beta_3 Period_{ij}$$

Mixed model:

$$\text{logit}\left(E\left(Y_{ij} \mid b_i\right)\right) = \beta_1^* + b_i + \beta_2^* Treat_{ij} + \beta_3^* Period_{ij}$$

b_i - patient's underlying propensity for an abnormal ECG

Example

	GEE			Mixed		
Variable	Estimate	SE	OR	Estimate	SE	OR
Intercept	-1.24	0.29		-4.08	1.67	
Treatment	0.57	0.23	1.8	1.86	0.93	6.4
Period	0.29	0.23		1.04	0.82	
$Var(b_i)$				24.43	18.85	

The greater underlying heterogeneity among patients, the greater is the discrepancy between estimates.

Example

Distinction between two analytic approaches:

Marginal: Average rates (in odds) of abnormal ECG in the study population if patients were treated with the drug

Mixed: Increase in odds of abnormal ECG for any patient treated with the drug.

Question: How harmful is the drug

Answer: Depends on interest in study population (public health, health insurance agent) or on an individual drawn at random from that population (physician). When the answers to both questions is of interest, there is no contradiction in reporting both estimates

What if we ignore clustering or time dependency?

- Regression coefficients will likely remain unbiased
- Regression models that ignore correlated measures tend to:
  - overestimate SE of time-varying covariates;
  - underestimate SE of time-invariant covariates

- How bad is it?
  - Depends on design effect (ICC and on cluster size)
  - Large clusters with small ICC can have similar design effect as small clusters with large ICC

Missing Data Assumptions

- Mixed models use maximum likelihood estimation and provide valid inferences in the presence of ignorable non-responses (MAR)

  MAR – probability of missingness depends on measured covariates and/or previously observed values of the outcome

- Marginal models require stronger missing data assumption: MCAR

  MCAR – missingness does not depend on any individual characteristics. Special case of MCAR is when missingness can depend on measured covariates

Summary of approaches for mixed models and GEE

- Hubbard et al., "To GEE or Not to GEE."

	Mixed Model	GEE
Focus of interest	Variance component and regression coefficients	Regression coefficients
Parameter interpretation	Individual (subject, school, neighborhood) specific	Population average
Linear Model (estimates equivalent)	Change in the mean outcome for a unit change in subject exposure, keeping the random effect (subject) fixed	Change in the mean outcome for a unit change in subject exposure across all of the subjects observed
Binary Model (estimates NOT equivalent)	The log(OR) of an outcome for a unit change in subject exposure, keeping the subject fixed	The log(OR) of an outcome for a unit change in subject exposure across all of the subjects observed
Assumptions	Correctly specified error distribution. Sensitive to different assumptions about variance and covariance structure, which are usually difficult to validate	Number of subjects should be sufficiently large for robust estimation of standard errors
Software		Can only accommodate two levels of hierarchy in many packages

More notes on GEE

- GEE uses 'sandwich' estimator, which produce consistent estimates even if correlation structure is specified incorrectly
- GEE preferred when the number of clusters is large (if five predictors, then 25 clusters is good)
- One additional correlation matrix is estimated: working (model-based) correlation. It is based entirely on hypothesized correlational structure
- If there is small number of clusters, an incorrectly specified model-based SE is better than a correctly specified robust SE
- No likelihood based measures for model selection (logL, AIC, BIC) is available (can do multi-parameter Wald test)

Thank you!