BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part X

Correlated Data: GLMM

Review

In last chapter, we learnt generalized estimating equations (GEE) for modeling marginal models. GEE methods are "semiparametric" because they do not rely on a fully specified probability model.

With GEE, the estimates are efficient if the working covariance assumptions are correct. If the working covariance assumptions are wrong, the estimated coefficients are still approximately unbiased, and SE's from the sandwich (empirical) method are reasonable if the sample is large. The philosophy of GEE is to treat the covariance structure as a nuisance.

This chapter presents an alternative model type that has a term in the model for each cluster. The cluster-specific term takes the same value for each observation in a cluster. This term is treated as varying randomly among clusters. It is called a random effect.

Overview

This chapter presents an alternative model type called Generalized linear mixed models (GLMMs).

- Generalized linear mixed models (GLMMs) are modern methods for handling correlated or clustered data
- They explicitly model cluster-specific or subject-specific effects
- This effect is treated as random effect across clusters, but takes the same value for each observation in a cluster.
- This approach contrasts with marginal models that average over the clusters or subjects

The model we considered here is also called random intercepts model. It is possible to have other types of random effects for instance random slopes.

- Revisit 2000 General Social Survey (R+SAS) and Longitudinal study on depression (R).
- Basketball Free Throw Success (R).
- R (lme4), SAS (GLIMMIX and NLMIXED)

References

- Li B, Lingsma HF, Steyerberg EW, and Lesaffre E. Logistic random effects regression models: a comparison of statistical packages for binary and ordinal outcomes. BMC Medical Research Methodology 2011, 11:77
- Pinheiro, José C., and Douglas M. Bates. Mixed-effects models in S and S-PLUS. Springer, 2000. APA

Matched-pair Example Revisit: 2000 General Social Survey

In the 2000 General Social Survey, 1144 subjects were asked whether, to help the environment, they would be willing to (1) pay higher taxes or (2) accept a cut in living standards.

	Cut Living Standards		
Pay Higher Taxes	Yes	No	Total
Yes	227	132	359
No	107	678	785
Total	334	810	1144

How can we compare the probabilities of a "yes" outcome for the two environmental questions?

Random Effects Model Approach

Let

$$\pi_{ij} = \Pr(Y_{ij} = \text{“Yes”}) \quad i = 1, 2, \ldots, 1144, \quad j = 1, 2$$

and

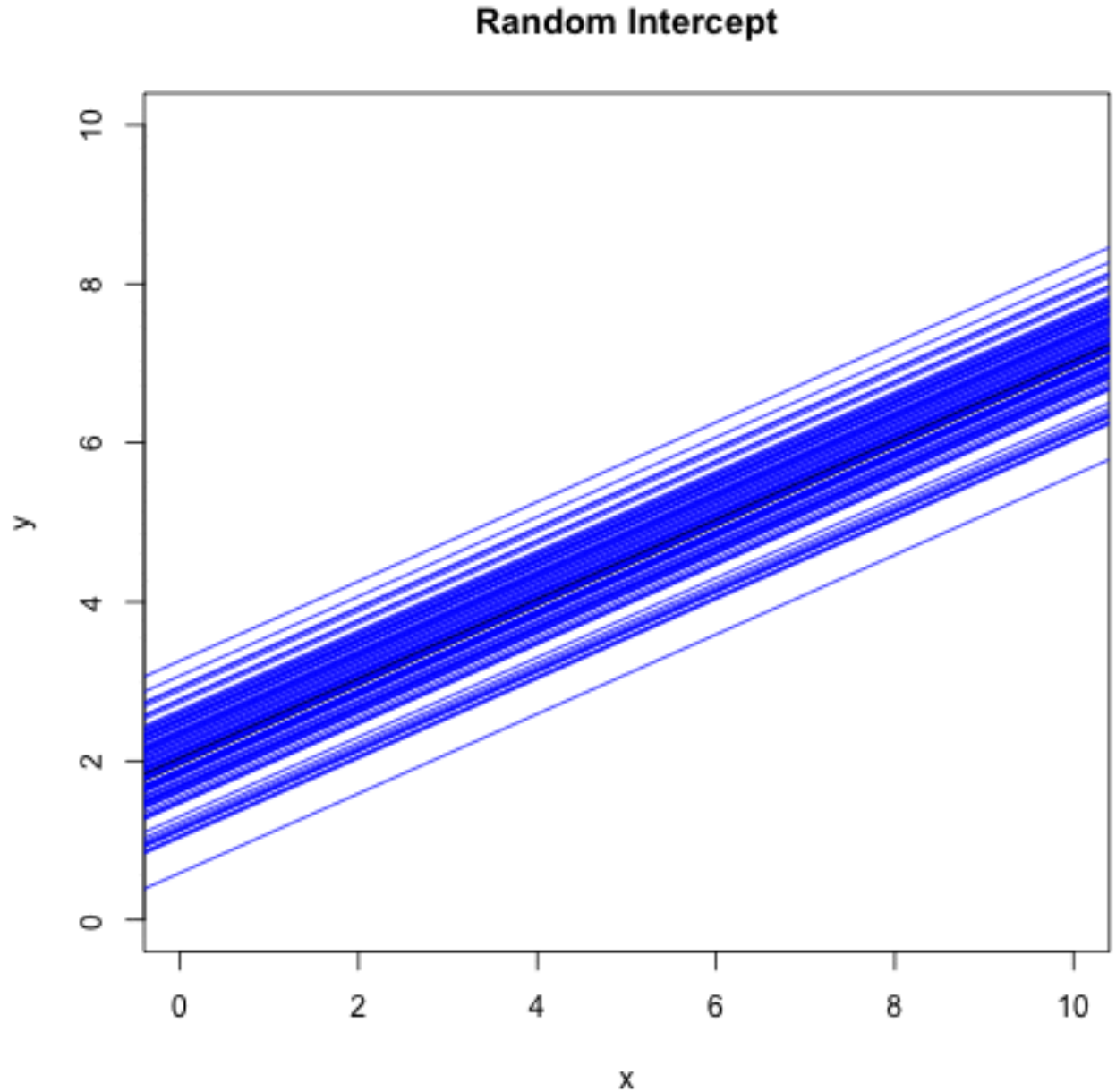
$$x_{i1} = 1 \text{ (Question 1)}, \quad x_{i2} = 0 \text{ (Question 2)}$$

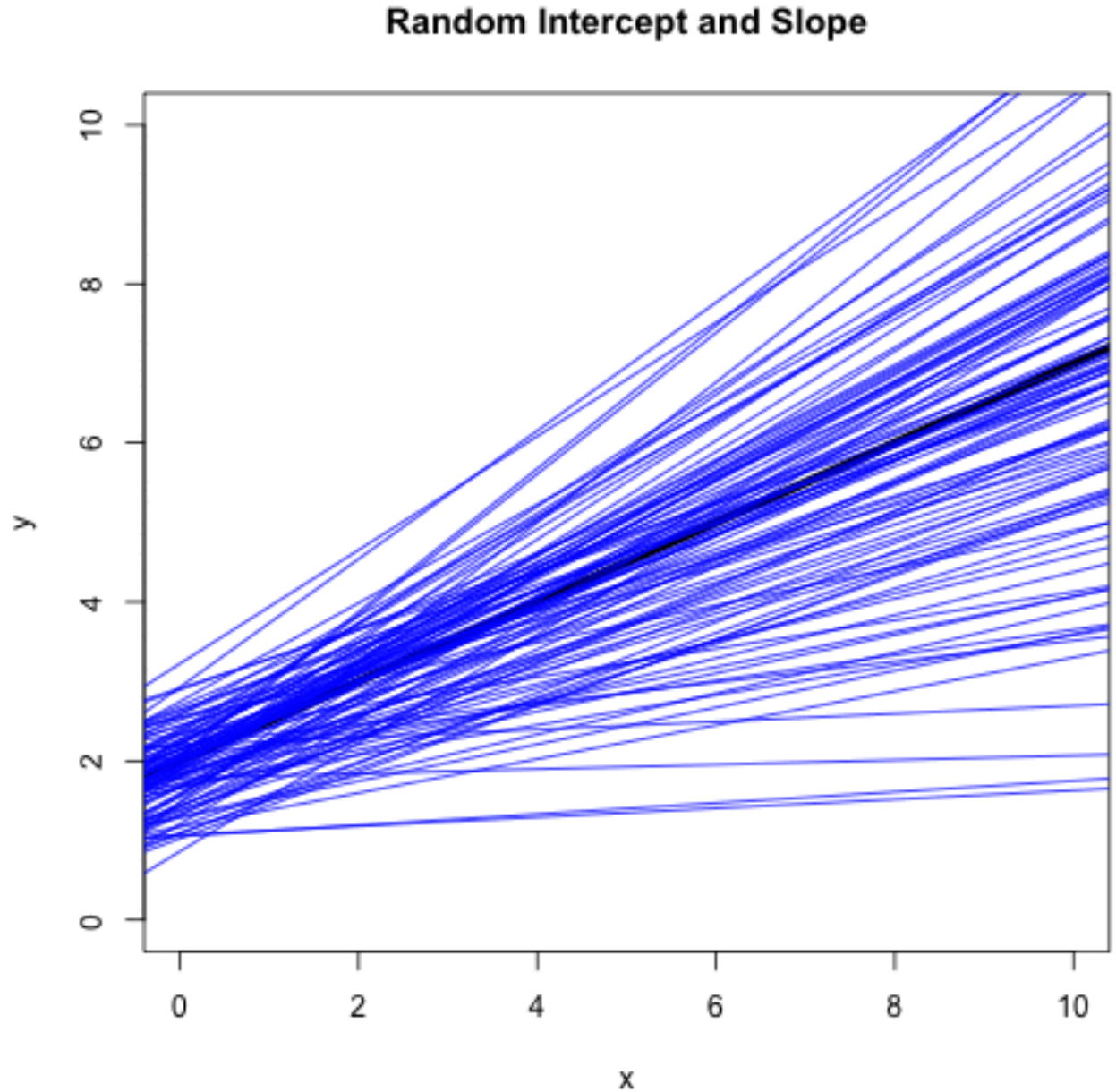
The GLMM model is

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \alpha_i + \beta x_{ij}, \\ \alpha_i &\sim N(\alpha, \sigma^2) \end{aligned} \tag{1}$$

α_i is the random effect of individual i

- Model (1) is subject-specific. Each subject has his/her own intercept
- This model is also called logistic-normal
- It is possible to use other distribution for the random effect. But normal is almost the standard choice
- The random effect is assumed for the intercept


Random Intercept
y
x
0
2
4
6
8
10


Random Intercept and Slope
y
x
0
2
4
6
8
10

An equivalent form of the GLMM in (1) is

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \alpha + u_i + \beta x_{ij}, \quad i = 1, 2, \ldots, 1144, \quad j = 1, 2 \\ u_i &\sim N(0, \sigma^2) \end{aligned}$$

The likelihood function is

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \Pr(Y_{i1}, Y_{i2} | x_{i1}, x_{i2}) \\ &= \prod_{i=1}^{n} \int \prod_{j=1}^{2} \Pr(Y_{ij} | u_i; x_{ij}, \alpha, \beta) \Pr(u_i; \sigma^2) du_i \end{aligned}$$

since Y_{i1} and Y_{i2} are conditionally independent given u_i.

Maximizing $L(\alpha, \beta, \sigma^2)$ is not trivial. The difficulty is in the evaluation of the integration. Approximation is used:

- Laplace approximation
- Gauss-Hermite quadrature (replace integration by bins, whose number needs to be specified)

Estimate from SAS/R

Software	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2_\alpha$	-2 log-likelihood
PROC GLIMMIX	$-1.417(0.236)$	$-0.209(0.130)$	8.112 (1.203)	2520.5
PROC NLMIXED	$-1.427(0.238)$	$-0.210(0.130)$	8.274 (1.275)	2520.3
R glmer	$-1.403(0.235)$	$-0.210(0.130)$	8.035 (NA)	2521.4

- Estimates of β are similar in the three methods, but not for α and σ^2
- PROC NLMIXED has smaller value on -2 log-likelihoods
- MH estimates of $\hat{\beta}$ is equal to

$$\log(132/107) = 0.21$$

with

$$SE = \sqrt{1/107 + 1/132} = 0.130$$

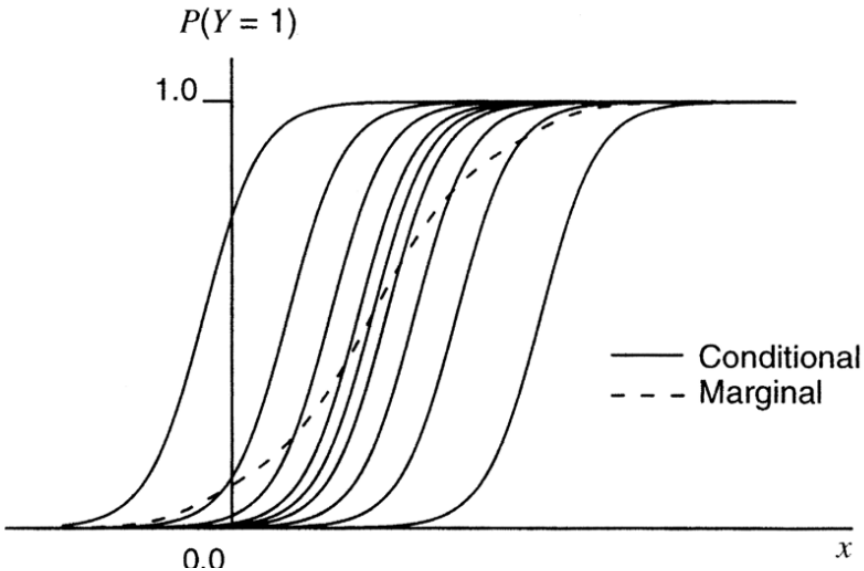
- "Whenever the sample log odds ratio in such a table is nonnegative, as it usually is, the ML estimate of β with this random effects approach is identical to the conditional ML estimate"

How does random effect model differ from the marginal model

Recall that the marginal model ignores the variation in α_i:

$$\text{logit}[\Pr(Y_i = \text{“Yes”})] = \alpha + \beta x_i, \quad i = 1, 2, \ldots, 2288$$

The estimates for β is $\log((359 \times 810)/(785 \times 334) = 0.104 < 0.21$



“When the link function is nonlinear, population-averaged effects of marginal models are typically smaller in magnitude than the cluster-specific effects of GLMMs”

Interpretation for estimated effects

For the conditional model: for any given subject, the estimated odds of a "yes" response on higher taxes are $\exp(0.210) = 1.23$ times the estimated odds of a "yes" response on lower standard of living.

For the marginal model: the estimated odds of a "yes" response on higher taxes for a randomly selected subject are $\exp(0.104) = 1.11$ times the estimated odds of a ?yes? response on lower standard of living for a different randomly selected subject.

Another example for understanding the interpretation from a study investigating smoking behavior on respirator function (Zeger et al., 1988).

- The population-averaged effects for marginal model: β estimates difference in infection rate between the population of smokers and non-smokers.
- The subject-specific effects from conditional model: β estimates difference in an individual's probability of infection given a change in his smoking status.

Zegar, S.L., Liang, K.-Y., and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. Biometrics 44, 1049-1060.

Small-area estimation of binomial probabilities

Small-area estimation refers to estimation of parameters for many geographical areas when each may have relatively few observations.

For example, a study might find county-specific estimates of characteristics such as the unemployment rate. With a national or statewide survey, counties with small populations may have few observations.

To see this, we use the basketball free throw success example to illustrate. Table 10.2 shows results of free throws (a standardized shot taken from a distance of 15 feet from the basket) for the 15 top-scoring centers in the National Basketball Association after one week of the 2005-2006 season.

$$\begin{aligned} \text{logit}(\Pr(\text{“success”})) &= \alpha_i, \quad i = 1, \ldots, n \\ &= \alpha + u_i, \quad i = 1, \ldots, n \end{aligned}$$

where

$$u_i \sim N(0, \sigma^2)$$

Based on the output,

- For a player with random effect $u_i = 0$, the predicted logit is $\hat{\alpha} = 0.9076$. The variance of this prediction is $\hat{\sigma}^2 = 0.1779$. 95% of the predicted logits fall within

$$0.908 \pm 1.96\sqrt{0.1779} = (0.08, 1.73)$$

  The predicted probability of making a free throw is

$$\frac{\exp(0.9076)}{1 + \exp(0.9076)} = 0.71$$

  The 95% confidence limits are

$$\exp\{(0.08, 1.73)\} = (0.52, 0.85)$$

- For player Yao, who has random effect $u_i = 0.0896$, the predicted logit is $\hat{\alpha} + u_i = 0.9972$. The variance of this prediction remains $\hat{\sigma}^2 = 0.1779$. Other calculations are parallel to above

player-specific sample proportions vs its estimated proportion

- p_i: proportion of successful ones
- π_i: Pr("success")

Table 10.2. Estimates of Probability of Centers Making a Free Throw, Based on Data from First Week of 2005–2006 NBA Season

Player	n_i	p_i	$\hat{\pi}_i$	Player	n_i	p_i	$\hat{\pi}_i$
Yao	13	0.769	0.730	Curry	11	0.545	0.663
Frye	10	0.900	0.761	Miller	10	0.900	0.761
Camby	15	0.667	0.696	Haywood	8	0.500	0.663
Okur	14	0.643	0.689	Olowokandi	9	0.889	0.754
Blount	6	0.667	0.704	Mourning	9	0.778	0.728
Mihm	10	0.900	0.761	Wallace	8	0.625	0.692
Ilgauskas	10	0.600	0.682	Ostertag	6	0.167	0.608
Brown	4	1.000	0.748				

Note: p_i = sample, $\hat{\pi}_i$ = estimate using random effects model.
Source: nba.com.

- Estimated proportions are less variation in them that the player-specific sample proportions
- Shrinkage towards the overall proportion $101/143 = 0.706$

Inference on Variance Components

In the Basketball Free Throw Success example, suppose we are interested in testing $\alpha_1 = \alpha_2 = \ldots = \alpha_{15}$ to see whether the success rates of the players are different. In random effects model, this is equivalent to testing

$$H_0 : \sigma^2 = 0 \text{ vs } H_1 : \sigma^2 > 0$$

The alternative H_1 is one-sided because σ^2 can not be negative.

- The likelihood ratio statistic no longer has a limiting chi-square distribution: with 50% chance a chi-square distribution with 1df; other 50% chance being equal to 0.
- The threshold for a level 5% test is 2.706 (i.e., not 3.84 anymore). (R code: *qchisq(0.9,1)*)
- The p-value is half of the right-tail of the corresponding χ^2. (R code: *0.5*(1-pchisq(LRTvalue,1))*)
- We can use the LRT to compare two nested covariance matrices; otherwise, choose with AIC/BIC

Choosing Marginal or Conditional Models

Marginal model approach

- GEE is computationally simpler and more amenable to standard software
- Likelihood-based inferences are not possible with GEE
- GEE does not provide estimates of subject-specific effects

Conditional model approach

- Useful if one wants to model the subject-specific effects
- Provides estimates of subject-specific effects (α_is) and their variation (σ^2) or other subject-specific characteristics
- Computationally demanding
- Risk of model mis-specification

Conditional Models: Random Effects versus Conditional ML

The conditional ML approach

- removes subject-specific terms in the model
- does not need an assumed distribution for subject-specific terms. This is good
- provides no information on subject-specific terms. Can not do between subject comparison. This is not good
- Computation intensive with large sample size
- can be less efficient than random effects model