Multilevel Models

1. Introduction. Variance Components

Germán Rodríguez

Princeton University

March 26, 2018

Website

The course has a website at **http://data.princeton.edu/pop510**, where you will find supporting materials including

- a course syllabus and bibliography, with useful links to other resources
- a collection of computing logs including
  - Stata and R logs fitting various linear and generalized linear multilevel models by maximum likelihood
  - Computing logs illustrating the use of Bayesian methods in multilevel analysis, including a random-effects logistic regression model fitted using WinBUGS and Stan
  - Some older runs using the classic package MLwiN

Some of my own research on multilevel models is housed at **http://data.princeton.edu/multilevel**. Resources include a list of publications, the simulated data used in a 1995 JRSS-A paper, and the actual data used in a 2001 JRSS-A paper (and earlier in a 1996 Demography paper)

Outline

1. Introduction. Variance-component models. Maximum likelihood and empirical Bayes estimates. Random intercepts.
2. Random slopes. Contextual predictors and cross-level interactions. Longitudinal data. Growth curve models. The general linear mixed model.
3. Models for binary data and the generalized linear mixed model. Likelihood approximations (MQL and PQL). Quadrature and adaptive quadrature.
4. Bayesian estimation in multilevel models. Markov chain Monte Carlo (MCMC). Gibbs sampling. The Metropolis algorithm. Hamiltonian Monte Carlo.
5. Models for categorical data. Ordered logits. Multinomial logits via SEM and Stan. Poisson models for count data. Small area estimation.
6. Multilevel survival models. Frailty models. Relationship with generalized linear mixed models. Interpreting results.

Software

The course will emphasize applications. Software used will be

- Stata, with `xtreg`, `xtlogit` and `xtpoisson` for random-intercept models and `mixed`, `melogit` and `mepoisson` for more general multilevel models,
- R's `lme4` package with the functions `lmer()` and `glmer()`,
- WinBUGS for Bayesian inference using the Gibbs sampler and Stan for Hamiltonian MCMC using NUTS.

Other specialized software in common use includes

- MLwinN, developed by Goldstein and collaborators at the Centre for Multilevel Modelling at Bristol University, and
- HLM, developed by Raudenbush and collaborators at the University of Michigan and distributed by Scientific Software International (SSI). A free student edition is available

Bibliography

The closest thing to a course textbook is

- Rabe-Hesketh, S., and A. Skrondal. (2012). *Multilevel and Longitudinal Modeling using Stata*, 3rd edition. Volume I: Continuous Responses and Volume II: Categorical Responses, Counts, and Survival. Stata Press.

The online syllabus cites other sources. Of particular note are

- Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd edition. London: Edward Arnold. 4th edition is print on demand. 2nd edition is available free in electronic form
- Gelman, A., and J. Hill. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, an excellent book on statistical modeling including multilevel models

Bates (2010) has posted chapters from a new book on mixed models with R.

Variance components

We start by considering a simple situation: 2-level clustered or longitudinal data with no covariates. The online example concerns children nested in schools.

Let Y_{ij} denote the outcome for the jth member of the ith group. (Notation is not consistent, MALMUS uses the subscripts in the reverse order. Levels may be counted up or down, and some count only grouping levels!)

The model is

$$Y_{ij} = \mu + a_i + e_{ij}$$

where μ is an overall mean, $a_i \sim N(0, \sigma_a^2)$ is a random effect for group i and $e_{ij} \sim N(0, \sigma_e^2)$ is the usual individual error term, independent of a_i.

Expectation and variance

The expected outcome in this model is just

$$E(Y_{ij}) = \mu$$

The variance of the outcome has two components

$$\text{var}(Y_{ij}) = \sigma_a^2 + \sigma_e^2$$

The covariance between two outcomes in the same group is

$$\text{cov}(Y_{ij}, Y_{ik}) = \sigma_a^2, \quad j \neq k$$

The correlation between two observations in the same group is

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2},$$

and is called the *intraclass* correlation.

Estimation of the mean

The OLS estimator of μ is the grand mean, which can be written as a weighted average of the group means

$$\bar{Y} = \sum_i n_i \bar{Y}_i / n$$

with weights proportional to sample sizes. This estimator is consistent but not fully efficient with unbalanced data.

The ML estimator of μ given the variances σ_a^2 and σ_e^2 is also a weighted average of the group means

$$\hat{\mu} = \sum_i w_i \bar{Y}_i / \sum w_i$$

but with weights w_i inversely proportional to their variances $\text{var}(\bar{Y}_i) = \sigma_a^2 + \sigma_e^2/n_i$, for which we plugin estimates.

Estimation of the variances

There are three approaches to estimating the variance components

- A method of moments or anova estimator equates the between-groups and within-groups sums of squares to their expected values.
- The maximum likelihood (ML) estimator maximizes the multivariate normal likelihood of the data with respect to all parameters, implemented in Stata and R.
- The restricted maximum likelihood (REML) estimator uses the likelihood of error contrasts, which allow for the estimation of the fixed parameters (analogous to using $n - p$ instead of n in OLS), also implemented in Stata and R.

Each of these approaches leads to a (usually slightly) different estimator of the mean. I generally prefer ML because it allows likelihood ratio tests for nested models.

Scores in language tests

For the language data in the computing logs the MLE of the mean is

$$\hat{\mu} = 40.364$$

compared to a grand mean of 40.935.
The estimated variance components are

$$\hat{\sigma}_a^2 = 4.408^2 \quad \text{and} \quad \hat{\sigma}_e^2 = 8.035^2$$

The intraclass correlation, or correlation between the scores of two students in the same school, is

$$\hat{\rho} = \frac{4.408^2}{4.408^2 + 8.035^2} = 0.231$$

so 23% of the variation in language scores can be attributed to the schools.

Tests of hypotheses

To test hypotheses about μ we used the fact that the ML (or REML) estimate is asymptotically normal, so we can use a Wald test. In particular, we can compute a 95% confidence interval

$$\hat{\mu} \pm 1.96\, \hat{\text{se}}(\hat{\mu})$$

For the school data the 95% interval for the mean is $(39.52, 41.20)$

To test hypotheses about σ^2_a we can use a likelihood ratio test, fitting models with and without school effects. Because the hypothesis is on a boundary of the parameter space, however, the test criterion does not have the usual χ^2_1 distribution, but is best treated as a 50:50 mixture of 0 and χ^2_1 by halving the p-value.

For the school data the test criterion is $\bar{\chi}^2_{01} = 287.98$, so we have highly significant school effects on language scores.

Predicting the random effects - ML

Consider "estimating" the school means

$$E(Y_{ij}|a_i) = \mu + a_i$$

Because these involve the random variables a_i we prefer to use the term "prediction". MALMUS uses "assigning numbers".

One approach is to treat the a_i as if they were fixed parameters and everything else as known, and then use ML. The resulting estimate is the difference between the group mean and the MLE of μ

$$\hat{a}_i^{ML} = \bar{y}_i - \hat{\mu}$$

Thus the ML estimate of the school mean is the sample mean $\bar{y}_i$.

Predicting the random effects - BLUP

An alternative approach is to minimize the prediction variance using a best linear unbiased predictor (BLUP), which has the form

$$\hat{a}_i^{BLUP} = \hat{a}_i^{ML} \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2/n_i}$$

The fraction on the right "shrinks" the ML towards zero by an amount that depends on the reliability of $\bar{y}_i$.

The corresponding estimator of the school mean is a compromise between the sample school mean and the overall mean

$$\hat{\mu}_i = \bar{y}_i w_i + \hat{\mu}(1 - w_i)$$

where w_i is the fraction in the top equation. These estimators have an empirical Bayes interpretation which we discuss next.

Bayes theorem

Bayes theorem gives us the conditional probability of $P|D$ as a function of the conditional probability of $D|P$

$$\Pr(P|D) = \frac{\Pr(PD)}{\Pr(D)} = \frac{\Pr(D|P)\Pr(P)}{\Pr(D)}$$

Suppose P are the parameters (which Bayesians view as random) and D are the data, so $\Pr(D|P)$ is the usual likelihood, $\Pr(P)$ is called the prior, and $\Pr(P|D)$ is the posterior distribution, so we can write

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Here we ignored $\Pr(D)$, which is just a normalizing constant.

Classical statisticians base their inferences on the likelihood, while Bayesians use the posterior. If the prior is vague or uninformative the two approaches give similar results.

Empirical Bayes

In the variance components model we treat a_i as a parameter with prior $N(0, \sigma_a^2)$.

The likelihood is the distribution of the $Y_{ij}|a_i$ for $j = 1, \ldots, n_i$ which are independent $N(\mu + a_i, \sigma_e^2)$. Maximizing yields $\hat{a}_i^{ML}$.

The posterior or distribution of $a_i|y_{ij}$ is proportional to the product of the prior and likelihood and can be shown to be a normal distribution, so the mean and mode coincide and give $\hat{a}_i^{EB}$.

This is not a full Bayesian approach, because instead of assuming priors for μ, σ_a^2 and σ_e^2 we simply plugged estimates. Hence the name *empirical* Bayes (EB).

This approach can be used to produce demographic estimates for small areas, where sample means may be unreliable

Shrinking towards zero

In the computing logs we show how to fit the variance-components model and predict the school random effects using ML and EB. The following graph shows substantial "shrinkage" for three small schools.