3/25/2010 36-402/608 ADA-II H. Seltman

Handout #17: Mixed Models, part 2

1. Formal model (review)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$
$$\mathbf{b} \sim N(\mathbf{0}, \mathbf{G})$$
$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R})$$

   $\mathbf{y}$ is all outcomes, stacked, usually with multiple measurements from the same upper level object in adjacent rows. $\boldsymbol{\beta}$ is the usual vector of unknown, population parameters (of the fixed effects).

   $\mathbf{X}$ is all fixed covariates (including appropriately coded categorical variables), one row per measurement.

   $\mathbf{Z}$ is the matrix of random covariates, one row per measurement. $\mathbf{b}$ is the vector of random effects, one per column of $\mathbf{Z}$.

   $\mathbf{G}$ is the variance-covariance matrix for the $\mathbf{b}$s, and $\mathbf{R}$ is the variance-covariance matrix for the $\boldsymbol{\epsilon}$s.

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$
$$\text{Var}\,(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

2. Estimation

   (a) Maximum Likelihood Method (ML) estimation of $\boldsymbol{\beta}$, and the parameters of $\mathbf{G}$ and $\mathbf{Z}$ is iterative, not closed form. ML estimates of the variance parameters are biased (just like $s^2_{ML} = \frac{\sum(y_i - \bar{y})}{n}$ is biased low by $\frac{n-1}{n}$).

   (b) The Restricted Maximum Likelihood (REML) method is unbiased for the variance parameters.

   (c) Fixed effects parameters and components of $\mathbf{R}$ (σ^2, and possibly others such as an AR1 correlation parameter) are interpreted in the usual way, but represent effects for the average upper level unit. Standard errors represent uncertainty of the estimates around the true parameters, as usual.

   (d) Random effect parameters show up as variance or s.d. estimates (with or without a standard error and/or p-value). These represent estimates of the true (irreducible) group-to-group variation in the values of the fixed parameters.

3. Model selection

   (a) Each random effect p-value is a test of whether some variance is zero, say $\tau^2 = 0$. Because the null hypothesis is on the boundary of the parameter space, these are not very reliable.

(b) The most commonly suggested model selection procedure is to use AIC or BIC in this algorithm:

i. Set a rich fixed effects structure (to minimize bias, even at the expense of increased variance).

ii. Use AIC or BIC with REML to choose a random effects structure (i.e, the forms of Z, G and R).

iii. Use AIC or BIC with ML to reduce the fixed effects to only those needed.

iv. Report the final REML fit. Add /SOLUTION to the MODEL statement to get parameters with SEs.

4. Some common variants

(a) Random intercept (only) model for patients within hospitals or repeated measurements within subjects: use the "equicorrelation" (compound symmetry) form for G and no further correlation (diagonal form) for R.

(b) Random intercept (only) model for repeated measurements within subjects: use the "equicorrelation" (compound symmetry) form for G and within subject autoregression for R.

(c) Random intercept and random slopes, (critically) allowing correlation between the random intercept and random slope for each upper level object.

(d) Spatial correlation using a distance measure.

(e) Three or more levels of the hierarchy.

5. SAS

(a) `CLASS` sets explanatory variables as categorical as in "PROC GLM"

(b) `MODEL` sets the fixed effects as in "PROC GLM"

(c) Need `/SOLUTION` on the `MODEL` line to get fixed effects parameter estimates.

(d) The `RANDOM` statement controls the Z and G matrices. Use `INT` for a random intercept (equicorrelation structure) and covariate names for random slopes. Need `/SUBJECT=` to set the grouping variable. For more than one random effect, almost always start with `TYPE=UN` (unstructured) to allow correlated random effects.

(e) The `REPEATED` statement controls R and is not needed if you want R to be diagonal. E.g., `REPEATED time / SUBJECT=subjectID TYPE=AR(1)`, where `time` is a class variable in the right order. The equivalent form for unequal spacing is `TYPE=SP(POW)(time)` where time is quantitative.

6. Breakout and Discussion