PSC 8185: Multilevel Modeling
Handout on Estimation for Nonlinear Models (e.g., binary, ordinal, etc)

For more details, see

- Skrondal and Rabe-Hesketh (S&R-H) (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC.
- Rabe-Hesketh, Skrondal, and Pickles (2002) (RH et al.) [available on Blackboard]
- Raudenbush and Byrk (2002), Ch. 14 from their book [available on Blackboard]

In a multilevel model, we start with the conditional distribution of the response (conditional on the random effects):

$$g^{(1)}(Y \mid X, \zeta; \theta)$$

where $z \sim \text{MVN}(0, S)$

Goal: Obtain the unconditional (marginal) distribution of the response for each cluster *j* by integrating out the random effects:

$$f^{(2)}(Y \mid X; \theta) = \int h(\zeta) \prod_{i=1}^{n_j} g^{(1)}(Y \mid X, \zeta; \theta) d\zeta$$

Full likelihood:

$$L = \prod_{j=1}^{J} f^{(2)}(Y \mid X; \theta)$$

Or:

$$L = \prod_{j=1}^{J} \int h(\zeta) \prod_{i=1}^{n_j} g^{(1)}(Y \mid X, \zeta; \theta) d\zeta$$

Computational Tasks:

- Full ML: Integrate out the random effects and maximize the likelihood
  - Numerical integration via Gauss-Hermite quadrature (GHQ): Replaces the continuous density of the random effects with a discrete distribution with *R* possible values based on quadrature weights and locations. As *R* increases, approximation becomes more accurate.
  - Adaptive quadrature (AQ): Adaptive quadrature allows the quadrature locations and weights to adapt to the cluster-specific nature of the data (see S&RH, pp. 165-169)
  - GLLAMM accommodates both GHQ and AQ.
  - In gllamm, one needs to specify the number of quadrature points; 12 usually does the trick for GHQ and AQ. One can also build up iteratively, from, e.g., 4 points to 8 to 12 to 16 to 20; the compare results to assess convergence.

- Alternatives:
  - Monte Carlo integration instead of numerical integration (S&R-H, p. 170)

– Bayesian approach; Markov chain Monte Carlo (MCMC) using Gibbs sampling (S&RH, p. 204-214).

Numerical Integration via Quadrature:

Gauss-Hermite quadrature (GHQ) is designed to evaluate integrals of the form:

$$\int \exp(-x^2) f(x) dx \approx \sum_{i=1}^{R} p_r^* f(a_r^*) \qquad \text{[see text, p. 165]}$$

- The exponential function, the weight function, is proportional to the normal density, so we can work with this. p=weight; a=location. Use *R* integration points.
- Recall, this is the integral we need to evaluate (for each cluster, *j*). Consider a random intercept model.

$$f^{(2)}(Y \mid X;\theta) = \int h(\zeta_j) \prod_{i=1}^{n_j} g^{(1)}(Y \mid X, \zeta_j;\theta) d\zeta_j$$

- We need to integrate out the normally distributed ζ_j.
- First, change z_j to a standard normal variable: $v_j = \zeta_j$ / sqrt(ψ).

$$f^{(2)}(Y \mid X;\theta) = \int \phi(v_j) \prod_{i=1}^{n_j} g^{(1)}(Y \mid X, \sqrt{\psi} v_j;\theta) dv_j$$

- ϕ is the standard normal density:

$$\phi(v_j) = \frac{1}{\sqrt{2\pi}} \exp(-v_j^2/2)$$

- Thus, applying G-H quadrature rule, the integral is approximated as:

$$\int \phi(v_j) \prod_{i=1}^{n_j} g^{(1)}(Y \mid X, \sqrt{\psi} v_j;\theta) dv_j = \sum p_r \prod g^{(1)}(y_{ij} \mid \sqrt{\psi} a_r$$

  See S&RH, p. 165 for more details.

- For multivariate integrals, e.g., if we have both a random intercept and a random slope, see p. 165. Just as we needed to work with the standard normal density in the random intercept case, in the multivariate case, we need to work with *independent* standard normal random effects, **v**, such that, for a given higher level, $\boldsymbol{\zeta} = \mathbf{Qv}$. **Q** is the Cholesky decomposition of the covariance matrix,$\boldsymbol{\Psi}$, of the random effects at a particular level. The Cholesky decomposition is analogous to taking the square root of $\boldsymbol{\Psi}$, such that $\boldsymbol{\Psi} = \mathbf{Q}'\mathbf{Q}$. Thus, the equation, $\boldsymbol{\zeta} = \mathbf{Qv}$, is analogous to $v_j = \zeta_j$ / sqrt(ψ) that we worked with in the random intercept case above.

- Then you write out the integrals over the *M* random effects terms at a particular level, as in equation 6.10 (p. 165). The integral is evaluated using a rectangular grid of points, as in figure 6.1, p. 166. Note that this is a case where there are two random effects (e.g., for a model with a random intercept and one random slope). R=8 in this case, but since there are two random effects, there are 8*8=64 total evaluations. Note that this a *rectangular* grid (unlike the adaptive quadrature grid) because in GHQ, we assume the two random effects are *independent*, and thus orthogonal (as visualized in the top-left panel of Fig. 6.1).

Issues with GHQ: Need a large number of quadrature points (*R*) for precise approximation. 12 or more for sure. GHQ can become less precise as cluster size increases and variance of the random effects increases.

Adaptive Quadrature:

- Adaptive quadrature (AQ) addresses the problems inherent in GHQ using empirical Bayes intuition.
  - In GHQ, we used the *prior* density (with expected means=0 and expected covariances=0; recall independence assumption) as the weight function in the integral; thus, the integral was a product of a *prior* density (prior) and the joint distribution of responses (likelihood for level-1 units in a particular cluster *j*).
  - Thus, the integrand is proportional to the *posterior* density of the responses, which can be estimated with normal density with a cluster-specific mean and variance.
  - The weight function is now the *posterior* density instead of the prior density. And this is why, in the upper-right plot of Fig. 6.1, we see that the two latent variables are not treated as orthogonal; they are allowed to be related since we're using the posterior density.
  - Adaptive quadrature allows the quadrature locations and weights to adapt to the cluster-specific nature of the data, as seen in the lower figure on p. 169.
  - See pp. 167-168 for more details; see also, RH et al., pp. 5-7
- Estimation done iteratively; update mean and var of the posterior iteratively until convergence (see S&RH, pp. 168-170).
- AQ will be more accurate than GHQ for large cluster sizes and large b/w cluster variances, since it's using the posterior density as the weight function and therefore adapting to the cluster specific nature of the random effects.
- Tough stuff! For more info, see text, 165-70; RH et al.

Marginal Quasi-Likelihood (MQL) and Penalized Quasi-Likelihood (PQL)

The intuition behind these estimation procedures is explained well on page 3 of the RH et al. Recall in GLS for linear models, the idea is to go back and forth between (1) estimating the parameter estimates (the β's), which are conditional on the var-cov matrix of the random effects, and (2) using the residuals to estimate the var-cov matrix of the random effects. This is iterative GLS (IGLS) or feasible GLS (FGLS). MQL and PQL seek to apply this IGLS procedure used in linear models to generalize linear models (i.e., binary, ordinal, count, and other DVs).

Think back to generalized linear modeling:

1. Specify the sampling model for the DV (normal for continuous, Bernoulli for binary, etc).
2. Write out the conditional expectation of the response, which is equal to the inverse link function.
3. Structural model: Write the link as a linear function of the level-1 independent variables and parameters.

Thus, MQL and PQL seek to "linearize" non-linear models by writing the response as a linear function of the linear predictor (x'b) and a heteroscedastic error component. Then, you use IGLS to iterate between estimating the fixed effects and the var-cov elements for the random effects.

- In MQL, the random effects terms are initially set to zero (the "prior" expectation).
- In PQL, estimation is improved by setting the random effects terms to their posterior modes.
- MQL and PQL can be improved by using second order Taylor series expansion for the random effects terms. These improved procedures are known as MQL-2 and PQL-2.
- See S&RH, pp. 195-197 for more details on this. See also the RH et al. article and the Rodriguez and Goldman article.

Issues:

- MQL and PQL, as well as MQL-2 and PQL-2, often underestimate the variances of the random effects, particularly for small cluster sizes (see especially Rodriguez and Goldman 2001).
- Rodriguez and Goldman found that even when using PQL-2, both the fixed and random effects estimates were attenuated (in the case of binary responses) for cases involving fairly large random effects variances.
- Bottom line: Full ML (via either GHQ or AQ) and MCMC are the "standards" for estimating models with non-continuous responses.