36-720: Generalized Linear Mixed Models

Brian Junker

October 10, 2007

- Review: Generalized Linear Models (GLM's)
- Generalized Linear Mixed Models (GLMM's)
- Computational Notes
- Facilities in R
- Examples

[Related to Christensen Chapter 13; Borrows from: Peter Dalgaard's short course at `http://staff.pubhealth.ku.dk/~pd/mixed-jan.2006/`; Brian Ripley's `lme4` notes at `http://www.stats.ox.ac.uk/~ripley/`]

Review: Generalized Linear Models (GLM's)

The Basic Idea

Suppose y_i follows an exponential family distribution of the form

$$f(y_i;\eta_i,\tau) = g(y;\tau)\exp\left(\frac{b(\eta_i)+k(y)\gamma(\eta_i)}{h(\tau)}\right) \equiv g(y_i;\tau)\exp\left(\frac{b(\eta_i)+y_i\eta_i}{h(\tau)}\right)$$

with τ known, so that η_i is the *natural parameter* and y_i is the *sufficient statistic*. You showed in homework, by differentiating $1 = \int f(y;\eta,\tau)dy$, that

$$\mu_i \;=\; E[y_i] \;=\; -b'(\eta_i) \;\equiv\; \ell^{-1}(\eta_i) \qquad (*)$$

and, by similar methods, you can show

$$\mathrm{Var}\,(y_i) \;=\; -b''(\eta_i)h(\tau) \;\equiv\; h(\tau)/\ell'(\mu_i) \qquad (**)$$

Rewriting the natural parameter η_i as a linear function of covariates X_i, we get

$$\ell(\mu_i) \;=\; \ell(E[y_i]) \;=\; \eta_i \;=\; X_i\beta$$

Plugging this back into the density, and considering the likelihood of n independent y_i's, we get

$$f(y_1,\ldots,y_n|\beta,\tau) = \prod_{i=1}^{n} g(y_i;\tau)\exp\left(\frac{\sum_i b(X_i\beta) + \sum_i y_i X_i\beta}{h(\tau)}\right)$$

Setting $\partial \log f(\cdots)/\partial\beta_j = 0$ we obtain the normal equations

$$0 = \left(\sum_i b'(X_i\beta)X_{ij} + \sum_i y_i X_{ij}\right)\Big/ h(\tau) = \sum_i \frac{(y_i - \mu_i)X_{ij}}{h(\tau)}$$

or, cancelling $h(\tau)$ and collecting terms,

$$(y-\mu)^T X = 0 \qquad (***)$$

These are "exactly" the same normal equations that we get from setting $\frac{\partial}{\partial\beta_j}\sum(y_i - X_i\beta)^2 = 0$ in OLS, except that here $\mu_i = \ell^{-1}(X_i\beta)$.

- From $(*)$, $\ell(E[y_i]) = X_i\beta$; $\quad \ell(\cdot)$ is called the *natural link function*;
- From $(**)$, $\mathrm{Var}\,(y_i) = h(\tau)/\ell'(X_i\beta)$; τ is called a *dispersion parameter*.

This is the basis of *generlized linear models (GLM's)*.

Examples:

- Normal linear regression. The Normal density is,

$$f(y_i;\mu_i,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma}\right)^2} = \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2}\right) \exp\left(\frac{-\frac{1}{2}\mu_i^2 + y\mu_i}{\sigma^2}\right)$$

  so μ_i is the natural parameter, $\ell(\mu_i) = \mu_i$ is the natural link function, and σ^2 is the dispersion parameter. So the GLM is $\mu_i = X_i\beta$.

- Poisson regression. The Poisson density is $f(y;\mu) = \mu^y e^{-\mu}/y! = (1/y!)\exp(-\mu + y\log\mu)$; the *over-dispersed Poisson family* has the form

$$f(y_i;\mu_i,\tau) = g(y_i;\tau)\exp\left(\frac{-\mu_i + y_i\log\mu_i}{\tau}\right)$$

  In this family, $\log\mu_i$ is the natural parameter, τ is the dispersion parameter, and we build GLM's of the form

$$\log\mu_i = X_i\beta$$

  We have also encountered this as the *log-linear model for Poisson sampling*.

- Logistic regression. The binomial density is $f(y;p) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} e^{n \log(1-p) + y \log \frac{p}{1-p}}$; the *overdispersed binomial* has the form

$$f(y_i; p_i, \tau) = g(y_i; \tau) \exp\left(\frac{n \log(1-p_i) + y_i \log \frac{p_i}{1-p_i}}{\tau} \right)$$

  The natural parameter is $\log \frac{p_i}{1-p_i}$, τ is the dispersion parameter, and we build GLM's of the form

$$\log \frac{p_i}{1-p_i} = X_i \beta$$

  This is of course the form of the *logistic regression model*.

Further examples are possible…

- Other location-scale members of the exponential family;
- Other link functions (the basic challenge of numerically solving the normal equations $(***)$ is about the same);

…but Normal regression, Poisson regression, and logistic regression are quite common.

Generalized Linear Mixed Models (GLMM's)

The basic idea is to take a GLM

$$\ell(\mu_i) = X_i\beta$$

and add random effects by analogy with the LMM:

$$\ell(\mu_i) = X_i\beta + Z_i u$$

Once again we generally take $u \sim N(0, \Psi)$, but because the error structure in a GLM is usually non-additive (it is handled by the underlying exponential family model), we do not add ε_i.

The prototypical case is logistic regression (Stiratelli, Laird & Ware, 1984):

$$\log \frac{p_i}{1 - p_i} = X_i\beta + Z_i u$$

Examples

- *Growth curves for binary outcomes.* With practice, people generally get better at problem solving. Let y_{ij} be the outcome (0 = incorrect, 1 = correct) for performing a task by person i on the t^{th} attempt. One version of the *power law of learning* says that the odds of performing the task correctly should increase like a power of t,

$$\frac{p}{1-p} = a \cdot t^b, \quad b < 0$$

This leads to a logistic regression model (GLM) of the form

$$\log \frac{p_{it}}{1-p_{it}} = \beta_0 + \beta_1 \log t$$

but, since there may be small individual differences in the rate of learning, we may wish to build a GLMM instead:

$$\log \frac{p_{it}}{1-p_{it}} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i}) \log t = (\beta_0 + \beta_1 \log t) + (u_{0i} + u_{1i} \log t)$$

- *Discrete outcomes in clustered survey sampling.* After taking a survey we will analyze a cross-classified table of counts for the discrete variables Sex ($i = 1, 2$), Income ($j = 1, 2, 3$), and Education ($k = 1, 2, 3$). Normally this would lead to a log-linear model

$$\log m_{ijk} = \beta_0 + \beta_i + \beta_j + \beta_k + \beta_{ij} + \cdots$$

However, in many national surveys, sampling is done in stages: first we sample a census block (say), and then we sample individuals within the census block. Because they live close together, people in the same census block will be more alike than people in different census blocks. A standard way to model this cluster-level dependence is by adding random effects to the model:

$$\log m_{ijkc} = (\beta_0 + u_c^{(0)}) + (\beta_i + u_{ic}^{(1)}) + (\beta_j + u_{jc}^{(2)}) + (\beta_k + u_{kc}^{(3)}) + (\beta_{ij} + u_{ijc}^{(4)}) + \cdots$$

(note that the β's play the role of u-terms in our earlier log-linear work, and the u's are the random effects here).

Computational Notes

For LMM's we could handle the random effects by computing a general error variance $V(\omega) = \text{Var}\,(\varepsilon) + Z\Psi(\omega)Z^T$, sidestepping an ugly integral.

For GLMM's, the random effects introduce an integral into the likelihood, of the form

$$\int f_{ij}(y_{ij}|u_i, \beta, \Psi)\, f(u_i|Data) du_i$$

(Molenberghs & Verbeke, 2005, Springer). There is no REML shortcut, and the full MLE's are usually computed using one of three approximations:

- Approximating the data: penalized quasi-likelihood (PQL);
- Approximating the integrand: Laplace's method;
- Approximating the integral: adaptive Gaussian quadrature (AGQ).

(EM is used by `nlme`; but apparently it is not very fast...)

Penalized Quasi-Likelihood (PQL)

Replace the GLM

$$\ell(\mu_{ij}) = x'_{ij}\beta + z'_{ij}u$$

with the nonlinear least-squares model

$$y_{ij} = h(x'_{ij}\beta + z'_{ij}u) + \varepsilon_{ij}$$

Taylor expansion of $h(\cdot)$ yields (in vector notation)

$$\hat{V}_i^{-1}(Y_i - \hat{\mu}_i) + X_i\hat{\beta} + Z_i\hat{u}_i \approx X_i\beta + Z_i u_i + \varepsilon_i^*$$

which gives a straightforward updating scheme, considering the LHS as pseudo data.

This is known as *penalized quasi-likelihood* because it obtains from optimizing a quasi-likelihood (involving only 1st and 2nd derivatives) with a penalty term on the random effects.

Laplace's Method

The integral amounts to a posterior mean, which can be approximated by careful Taylor expansion of the log-integrand. The usual approach is:

$$E_{u_i}[f_{ij}(y_{ij}|u_i,\beta,\Psi)|Data] = \frac{\int f_{ij}(y_{ij}|u_i,\beta,\Psi)f(Data|u_i)f(u_i)du_i}{\int f(Data|u_i)f(u_i)du_i}$$

$$= \frac{\int e^{-nh^*(u_i)}du_i}{\int e^{-nh(u_i)}du_i} = \frac{\sqrt{1/h^{*\prime\prime}}(u_i^*)e^{-nh^*(u_i^*)}}{\sqrt{1/h''(\hat{u}_i)}e^{-nh(\hat{u}_i)}}[1+O(n^{-2})]$$

(Tierney & Kadane, 1986, *JASA*).

The Hessians $h''(u)$, $h^{*\prime\prime}(u)$ and maximizers $\hat{u}$, u^* come automatically from optimization routines; and n is an appropriate sample size.

Thus integration is replaced with differentiation, which is faster and more stable (as long as the Taylor expansion holds).

Adaptive Gaussian Quadrature (AGQ)

Quadrature is another name for numerical integration using a weighted sum of the form

$$\int f(x)dx = \sum_i w_i f(x_i)$$

When $f(x)$ is a normal density, or has a log-quadratic factor, it can be expanded in Hermite polynomials, and then the w_i can be *Gaussian quadrature weights*, which are efficient for the problem.

Adaptive quadrature optimizes over the placement and number of the x_i, and the choice of w_i.

PROC MIXED / PROC NLMIXED in SAS use AGQ; and AGQ is the method of choice in Rabe-Hesketh, Skrondal & Pickles (2004, *Psychometrika*) `GGLAMM` package for `Stata`.

Facilities in R

R (and Splus) provide one "default" package for GLMM analysis, and other packages exist:

- `library(lme4)` is a rewrite of `nlme` that provides `lmer()`, for both *LMM's* **and** *GLMM's*.

  `lmer()` has a `method`= argument which can take the value `"Laplace"`, `"PQL"`, or `"AGQ"`. PQL is the default, and AGQ is not yet implemented.

- `library(MASS)` (for R or Splus) provides `glmmPQL()` (uses PQL).

- Other installable packages from `http://cran.r-project.org/` (R only) include `glmmML` (uses Laplace or AGQ) and `glmmAK` (extends to multinomial logit models; see `VGAM` for the fixed-effects-only case).

As with all things, R is great for "breadboarding", and for analyses of problems of moderate size.

For larger problems, SAS provides PROC MIXED and PROC NLMIXED, that provide approximately the same functionality as `lme, nlme` and `lmer`.

Examples

- A meta analysis of published studies of the use of beta-blockers to prevent death after heart attack; random effects are used since the size of the effect of beta-blockers varies moderately from study to study.
- Analysis of five cognitive items in a section of the LSAT exam. There are 1000 examinees and 5 items, so an additive fixed-effects model would require at least 1005 parameter estimates, very slow (after 4 hours I stopped trying!), and inconsistent (Neyman-Scott problem; see also Andersen and Haberman on the Rasch model). If we treat the student effects as random, then we have only 6 parameters (5 fixed effects for the items, plus a variance component for student effects), and estimates are consistent as number of students grows.

[See R notes in class]

Once again, these examples only scratch the surface. For more details, see Dalgaard's or Ripley's notes. Also, there are many similar examples in the "Examples" manuals of WinBUGS (`http://www.mrc-bsu.cam.ac.uk/bugs/`) software for doing applied Bayes via MCMC.