Multilevel Models

5. Multilevel Logit Models

Germán Rodríguez

Princeton University

April 9, 2018

Binary data

We now turn our attention to clustered and longitudinal *binary* data. Examples that we will consider include

- Data on the decision to deliver a birth in a hospital or elsewhere, with repeated observations on a sample of women.
- Contraceptive use by women in the Bangladesh DHS. The data are clustered by district, which may affect both levels and urban-rural differentials in contraceptive use.
- Immunization status for Guatemalan children, which are clustered by mother, which are in turn nested in communities.

For the first example and part of the second we can use fixed or random effects models.

For three or more levels, and more generally for random coefficient models, we need a multilevel approach.

We start with a quick reminder of fixed and random-effects models.

Fixed effects models

We consider a clustered binary outcome following the fixed-effects model

$$Y_{ij} \sim B(\pi_{ij}), \quad \text{with} \quad \text{logit}(\pi_{ij}) = \alpha_i + x_{ij}'\beta$$

where α_i is a separate parameter for each group.

The usual ML estimator, equivalent to adding a dummy variable for each group, is inconsistent not just for the group parameters α_i but for β as well, in contrast with linear models.

The solution is to condition on group totals, which happen to be minimal sufficient statistics for the group effects α_i.

The resulting likelihood involves only groups with variation in both the outcome and the predictors. Sometimes losing 90% of the data is disconcerting, but a necessary price to pay to control for group-level omitted variables.

Random effects models

An alternative model assumes that the group effects are random, so

$$Y_{ij} \sim B(\pi_{ij}), \quad \text{where} \quad \text{logit}(\pi_{ij}) = a_i + x'_{ij}\beta$$

where $a_i \sim N(0, \sigma_a^2)$, is independent of the covariates and of the implicit error term.

The model can be written in terms of a latent variable following a linear random-intercept model, where $Y_{ij} = 1$ if $Y_{ij}^* > 0$, and

$$Y_{ij}^* = a_i + x'_{ij}\beta + e_{ij}$$

where $a_i \sim N(0, \sigma_a^2)$ as before and e_{ij} has a standard logistic distribution with mean 0 and variance $\pi^2/3$ (or $N(0,1)$ for probit). Just as in logit models we fix the error variance to identify β.

Estimation by ML is implemented in Stata and R, but is not without some challenges that we now discuss.

Maximum likelihood estimation

In multilevel linear models the marginal likelihood is multivariate normal, so estimation is straightforward.

In multilevel logit models the likelihood is logistic-normal and, unfortunately, has no closed form. In the random intercept model the contribution from cluster i is

$$L_i = \int_{-\infty}^{+\infty} g(a) \prod_{j=1}^{n_i} \pi_{ij}(a)^{y_{ij}} [1 - \pi_{ij}(a)]^{1-y_{ij}} da$$

where $\pi_{ij}(a) = \text{logit}^{-1}(a + x_{ij}'\beta)$ and $g(a)$ is the $N(0, \sigma_a^2)$ density. This integral is intractable.

Not surprisingly, various researchers have proposed approximations. Regrettably, some of them don't work very well. I'll summarize the main approaches, see Rodríguez and Goldman (1995, 2001), henceforth RG1 and RG2, for more details.

MLQ: Marginal quasi-likelihood

The multilevel logit model can be written in general form as

$$\mathbf{y} = \boldsymbol{\pi} + \mathbf{e} \quad \text{where} \quad \boldsymbol{\pi} = \text{logit}^{-1}(\mathbf{X}\beta + \mathbf{Z}u)$$

MQL-1. Goldstein approximates the inverse logit using a first order Taylor series about $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\mathbf{u} = \mathbf{0}$ for a trial estimate $\boldsymbol{\beta}_0$. This leads to an approximating multilevel *linear* model, which is used to obtain an improved estimate. The procedure is iterated to convergence. Longford uses a quadratic approximation to the log-likelihood. RG1 show that it is equivalent to MQL-1.

MQL-2. A second-order approximation that uses second derivatives w.r.t. $\mathbf{u}$ only, ignoring second derivatives w.r.t. the fixed effects $\boldsymbol{\beta}$ as well as mixed derivatives. Convergence can be an issue.

Both procedures are implemented in MLwiN. Not surprisingly, they work well for very small $\mathbf{u}$.

PQL: Penalized quasi-likelihood

PQL-1. An obvious improvement is to approximate $\boldsymbol{\pi}$ using a Taylor series expansion about

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{and} \quad \mathbf{u} = \mathbf{u}_0$$

where $\boldsymbol{\beta}_0$ is the current ML estimate of $\boldsymbol{\beta}$ and $\mathbf{u}_0$ is the EB estimate of $\mathbf{u}$ evaluated at current parameter estimates.
This method has been derived by several authors using different perspectives. It was named PQL by Breslow and Clayton. It usually works better than MQL.

PQL-2. Goldstein and Rasbash proposed a second-order PQL approximation using second derivatives w.r.t. $\mathbf{u}$ only, ignoring second derivatives w.r.t. the fixed effects $\boldsymbol{\beta}$ as well as mixed derivatives, just as before.

MLwiN implements both forms of PQL.

A simulation study

RG1 conducted a simulation study using several scenarios, involving small and large random effects and designs with small and large clusters, as found in education and demographic research.

Of particular interest is a set of simulations using the same structure as a real dataset from Guatemala, which concerned prenatal care for 2449 births among 1558 women nested in 161 communities. In fact, it was doubts about conventional estimates obtained with the actual data that motivated the simulation study.

We simulated data using known values of the fixed coefficients and of the variances of the random effects, and then fitted a three-level random intercept model using MQL and PQL.

The data were made available through JRSS-A and on my website, and have been used by several authors, including Nelder, Goldstein and Rasbash, and Browne and Draper.

Comparison of estimates

Here are some results from Table 9.1 in Rodríguez (2008), which has the most complete set of estimates for a simulation using large random effects.

Estimation	Fixed Part (β)			Random Part (σ)	
method	Individual	Family	Community	Family	Community
True value	1.000	1.000	1.000	1.000	1.000
MQL-1	0.738	0.744	0.771	0.100	0.732
MQL-2	0.853	0.859	0.909	0.273	0.763
PQL-1	0.808	0.806	0.831	0.432	0.781
PQL-2	0.933	0.940	0.993	0.732	0.924

MQL-1 underestimates βs by 23-26% and σs by 27 and 90%! MQL-2 is more accurate but doesn't always converge. PQL-1 is better than MQL-1, competitive with MQL-2, and more likely to converge. PQL-2 is best in the series, with 1-7% bias for the βs, but still underestimates σs by 8 and 27% and may not converge.

Gaussian quadrature

In light of these results we turned to ML via numerical integration of the likelihood function using Gaussian quadrature.

Quadrature rules approximate an integral as a weighted sum over a grid of points. Gaussian quadrature chooses both the weights and the evaluation points to minimize error for different integrands.

Gauss-Hermite quadrature can be used with integrals of the form

$$\int f(x) e^{-x^2} dx = \sum_{k=1}^{q} w_k f(x_k)$$

The evaluation points are zeroes of the Hermite polynomials and, together with the weights, can be obtained from tables or code.

This method can be applied to the integral in slide 5 through a simple change of variables.

Adaptive quadrature

An alternative procedure that achieves remarkable accuracy with fewer points moves the evaluation points to cover the posterior rather than the prior distribution of the random effects.

Liu and Pierce approximate the posterior using a normal distribution with the same mode and curvature at the mode. This has the effect of sampling the integrand in a more relevant range. The method with just one point is equivalent to a Laplace approximation or PQL-1.

Rabe-Hesketh and collaborators, building on work by Naylor and Smith, use the posterior mean and variance of the random effects instead of the mode and curvature. This leads to somewhat simpler calculations and was first implemented in their `gllamm` command.

Pinheiro and Bates see adaptive quadrature as a deterministic version of importance sampling and use it in non-linear models.

Validating quadrature methods

Does it work? We validated ML via quadrature using the simulated data before using it on actual data, with the following results

Estimation	Fixed Part (β)			Random Part (σ)	
method	Individual	Family	Community	Family	Community
True value	1.000	1.000	1.000	1.000	1.000
ML-5	0.983	0.988	1.037	0.962	0.981
ML-20	0.983	0.990	1.039	0.973	0.979

Obviously numerical integration works very well indeed, even with as few as 5 points.

Our analysis of the Guatemalan data, published in *Demography* and used as a case study in RG2, used 20 quadrature points at each level. I later was able to reproduce the results exactly using 12-point adaptive quadrature. The page `maxlik.html` has some vintage runs and comparisons, but these days we use Stata or R.

Software notes

Stata implements quadrature procedures in two commands:

`xtlogit` fits random-intercept models. The option `intmethod()` can be `ghermite` for classic Gauss-Hermite, `aghermite` for adaptive G-H using mode and curvature, or `mvaghermite` for adaptive G-H using the mean and variance. The default is `mv`. The number of points is specified with the `intpoints()` option and defaults to 12.

`melogit` fits random-coefficient models using adaptive Gauss-Hermite with 7 points per effect as the default. In addition to the `intmethod()` and `intpoints()` options, there's a `laplace` option, equivalent to PQL-1, as a faster but less accurate alternative for exploratory work. The number of integration points can be varied by level.

Stata 14 can also fit these models using `meglm`.

Software notes (continued)

R's `lme4` package has a function `glmer()` to fit generalized linear multilevel models.

For random-intercept models the default is PQL, but it is possible to specify adaptive quadrature using the mode and curvature by specifying the number of integration points via the `nAGQ` argument, which defaults to one. I strongly recommend that you avoid the default and specify 7 or preferably 12 points as Stata does.

For random-coefficient models the only option available is PQL, which unfortunately means that maximum-likelihood results should be considered approximate and useful only for exploratory work. (As we will see later, however, these models can be estimated in R using Bayesian methods.)

Hospital Deliveries

Our first example will use data from Lillard and Panis on the decision to deliver a birth in a hospital or elsewhere, available in the datasets section as `hospital.dat`.

The dataset comprises 501 women with 1060 births. The outcome `hosp` is a binary indicator of hospital delivery with mean 0.297.

The predictors of interest are `loginc` or log-income, `distance` to the nearest hospital, and two indicators of the woman's education: `dropout` for less than high school and `college` for college graduates or higher (only 8.4% of the women).

A simple logit model suggests that all predictors have significant effects on the probability of hospital delivery, but the assumption of independence is not adequate with repeated observations on the same women.

A Random-Intercept Model

We therefore introduce a woman level random effect a_i and assume that conditional on that each woman's outcomes are independent with probability satisfying the logit model

$$\Pr\{Y_{ij} = 1 | a_i\} = \text{logit}^{-1}(a_i + \boldsymbol{x}'_{ij}\boldsymbol{\beta})$$

where $\boldsymbol{x}_{ij}$ represents the predictors for the j-th birth of the i-th woman and $a_i \sim N(0, \sigma^2_a)$ is the woman-specific random effect, assumed normally distributed.

As noted earlier the likelihood for this model has no closed form and must be evaluated using numerical integration. The computing logs show results using 12 quadrature points in Stata and R. Notably R's default choice of PQL does not converge with these data, but specifying `nAGQ=7` works fine.

Maximum-Likelihood Estimates

Here are estimates obtained using all the defaults in Stata

```
Integration method: mvaghermite                 Integration pts.  =         12

                                                Wald chi2(4)      =     110.06
Log likelihood  = -522.65042                    Prob > chi2       =     0.0000
 ------------------------------------------------------------------------------
        hosp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      loginc |   .5622009   .0727497     7.73   0.000     .4196141    .7047876
    distance |  -.0765915   .0323473    -2.37   0.018    -.1399911   -.013192
     dropout |  -1.997753   .2556249    -7.82   0.000    -2.498769   -1.496737
     college |    1.03363   .3884851     2.66   0.008     .2722135    1.795047
       _cons |   -3.36984   .4794505    -7.03   0.000    -4.309546   -2.430134
-------------+----------------------------------------------------------------
    /lnsig2u |   .4372018   .3161192                     -.1823805    1.056784
-------------+----------------------------------------------------------------
     sigma_u |   1.244335   .1966791                       .912844    1.696203
         rho |   .3200274   .0687907                      .2020988    .4665343
------------------------------------------------------------------------------
LR test of rho=0: chibar2(01) = 29.61                  Prob >= chibar2 = 0.000
```

We will discuss interpretation of the fixed effects as well as the standard deviation of the random effects. (We'll leave estimation of the random effects themselves to the next example.)

Subject-Specific and Population Average

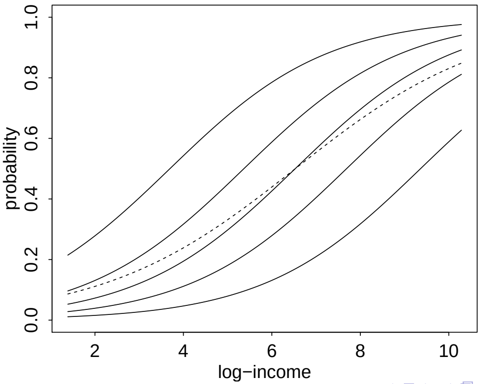
The fixed effects have a *subject-specific* interpretation. For example the coefficient of `college` means that the odds of delivering a birth in a hospital are multiplied by 2.81 when a woman has a college education, compared to what her odds would be with only a high school education but the same income, distance to the hospital, *and* unobserved characteristics as captured by a_i.

Contrast this with a *population-average* effect, which we can obtain by averaging the effect of college education over all women with given observed characteristics. For example at the mean `loginc` of 5.988 and the mean `distance` of 3.918, the probabilities for `college` 1 and 0 averaged over the distribution of a using Gauss-Hermite integration are 0.637 and 0.442, leading to an odds ratio of 2.21.

Population-average (or marginal) coefficients are smaller in magnitude than subject-specific (or conditional) coefficients.

Plotting SS and PA Effects

The figure below shows the predicted probability of hospital delivery as a function of log-income for women with high school education, who live at the average distance from a hospital, and have unobserved characteristics in percentiles 10, 30, 50, 70 and 90. We also show the predicted probabilities based on the population average model (dashed line).



Standard Deviation of Random Effects

A nice way to interpret the standard deviation σ_a is to write $a_i = \sigma_e z_i$ where z_i is a standard-normal random effect, so the model becomes

$$\text{logit}(\pi_{ij}) = \sigma_a z_i + \boldsymbol{x}'_{ij}\boldsymbol{\beta}$$

and σ_a can be interpreted as a regular logit coefficient for the standardized random intercept z_i

In our data $\hat{\sigma}_e = 1.244$. Thus, the odds of hospital delivery for a woman with unobserved characteristics one standard deviation above the mean are 3.47 times the odds of an average woman with the same log-income, distance to a hospital and education.

Similarly, the odds for a woman with unobserved characteristics one standard deviation below the mean are 71% lower than for the average woman with the same observed characteristics.

This parameter is also related to the intra-class correlation.

Latent Intra-Class Correlation

The intraclass correlation is best defined in terms of the latent variable formulation of the model shown earlier. For a logit model

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \pi^2/3}$$

because in a standard logistic distribution $\sigma_e^2 = \pi^2/3$. (In a probit model $\sigma_e^2 = 1$, and in a c-log-log model is it $\sigma_e^2 = \pi^2/6$.)

For the hospital delivery data the correlation between the propensity of a woman to deliver any two births in a hospital is

$$\hat{\rho} = \frac{1.244^2}{1.244^2 + \pi^2/3} = 0.32$$

This also means that 32% of the variance in the latent propensity to deliver a birth in a hospital can be attributed to women.

Manifest Intra-Class Correlation

In a 2003 paper with Elo we proposed looking at the correlation between actual binary outcomes, which depends on the covariates. Our method is described in MALMUS §10.9.3 and implemented in a Stata command called `xtrho`.

We calculate a two-by-two table of expected outcomes for two observations in the same group, which we do by integrating out the random effect at selected values of the linear predictor. At the median we get

	No	Yes	
No	0.6153	0.1454	0.7607
Yes	0.1454	0.0938	0.2393
	0.7607	0.2393	1.0000

The marginal probability that a median woman would deliver a birth in a hospital is 24%, and the joint probability for two births is 9%. The Pearson correlation is 0.20 and Yule's Q is 0.46. The odds ratio is 2.73.

Correcting Standard Errors for Clustering

Some researchers faced with repeated binary observations simply fit logit models and then adjust the standard errors for clustering using extensions of the Huber-White “sandwich” estimator. This approach is fine if you keep in mind two caveats:

1. You must realize you are fitting a population-average rather than a subject-specific model and interpret the parameters accordingly. As we have seen, the effect for a particular subject differs from the average effect in the population.
2. The estimates obtained using a logit model are not efficient because they ignore the correlation structure of the data. A better approach is to use generalized estimating equations (GEE), which produces efficient population-average estimates and correct standard errors.

Comparison of Estimates

The table below compares four estimates of the effect of college education and its standard error, using logit models, logit with corrected standard errors, GEE, and random effects

	Logit	Cluster	GEE	Multilevel
$\hat{\beta}$	0.8217	0.8217	0.8078	1.0336
s.e.	0.2611	0.2884	0.2980	0.3885

The first three methods estimate a population-average effect equivalent to an odds ratio of 2.24 (not unlike our result), and both correcting for clustering and using GEE inflate the standard error.

The estimated subject-specific effect corresponds to an odds ratio of 2.81 and is larger than the average effect (it also has a larger standard error).

The key point is that having clustered data affects not just the standard errors but the coefficients themselves.