Where do Random Effects Come From?

Now that you are (hopefully) convinced that you should probably be using random effects models we have to figure out how to get the actual random effect.

When you estimate a random effects model you do so by something like:

1) Estimating a fixed effects model with no intercept or covariates by
    a. Generalized Least Squares (GLS)
    b. Maximum Likelihood Estimation (MLE) via Laplace Approximation or linearization
        i. Also sometimes called penalized or marginal quasi likelihood
    c. Restricted MLE via Laplace Approximation or linearization
    d. numerical/stochastic integration
2) Extracting the coefficients (and possibly standard errors) from the indicator variables
    a. Either via a 2 stage model or an Expectation Maximization procedure
3) Stacking them in another variable called a random effect
    a. Literally just using something like predict
4) Incorporate them into the model
    a. Either via GLS or integrating them into a likelihood function and then taking the first and second derivatives and so on
5) You often then use an EM procedure to refine the random effects estimates

That list of things probably didn't make any sense to anyone and that's fine. There are literally entire graduate courses on campus that only go through parts of that list so we are just going to go over the basics ideas.

Direct Estimation of Random Effects by Fixed Effects

- If you estimate a fixed effects only model with indicator variables and then predict Y you end up with a random effect. This is a two stage approach to RE estimation that people used to use regularly. You can use it to difference the random effect out of the model—remove it from the residuals in a Generalized Least Squares setting or condition on it in a Maximum Likelihood setting so that it's out of the way for calculating other variables. You can also use the first stage estimates as a variable (or control function) in a second stage model.
    - This is ideal because it is quick and gets you the exact random effect. You just use matrix algebra or calculus to directly calculate the thing analytically.
- Linear regression and Poisson are pretty much the only times you can do this.
- You can't estimate fixed effects (without a lot of effort and under very particular conditions – see Hahn and Newey (2004) in most nonlinear or generalized linear models so you can't build the random effects from them.
    - There is no closed form solution to integrate out the random effect because it's nonlinear in the parameters the same as the X's. In other words, the calculus problem is impossible if you want an exact answer.
    - You can't use dummy/indicator variables to build the random effect manually because artificially increasing the dimensionality of a nonlinear likelihood function can induce bias very rapidly (as few as ten dummies). This is known as incidental parameters bias (Lancaster 2000) and it is **very** annoying.
        - It tends to get worse the further away from linear models you go. Tobit models and negative binomial models seem fine with dummies but binary outcomes can be horribly biased (Greene 2002, 2004, 2007) .

Linearization via Taylor Series

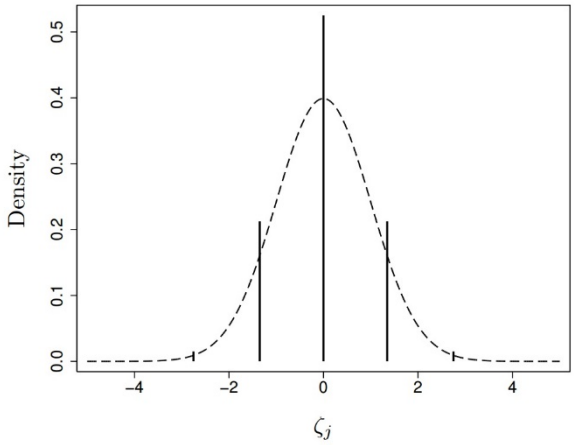
- If you don't have a linear outcome you can cheat and treat it like one anyway by using a high order Taylor Series expansion. It's basically like including age and age squared in a linear regression to account for the fact that age has a nonlinear effect. The Taylor Series does this with all model parameters and can include squared, cubic, quartic or whatever level terms you want. Eventually, it will get you accurate results from a linear model no matter what kind of outcome you are dealing with.
- This is used in the HLM program and in a number of other settings as either a first or second order Taylor Series. It's referred to as Penalized or Marginal Quasi Likelihood because it uses the Taylor Series to build fixed effects, then random effects, then incorporates those random effects into a Generalized Linear Model.
- It assumes the random effects are normally distributed and it almost always underestimates the variance.
    - First order Taylor Series underestimates by about 20% in some cases. Second order Taylor Series underestimates variance by 5%-10%.
    - If you have weirdly shaped random effects the performance is worse.

Approximation of an intractable integral via Laplace Approximation

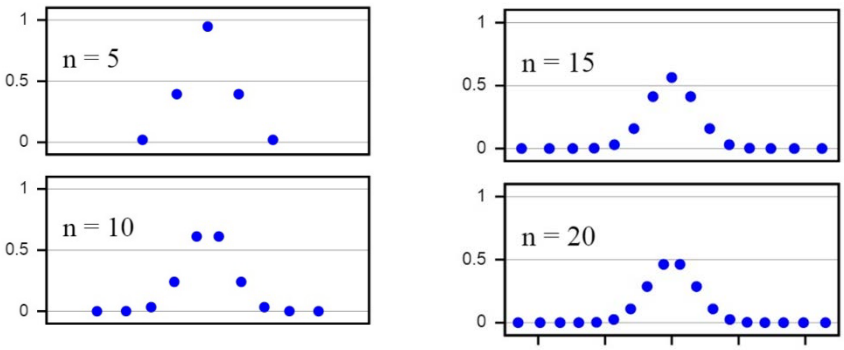
- Like the linearization approach it uses a first or second order Taylor Series approximation but it does it in a different way. You start with the Posterior Bayes Mode probably calculated by some linear model. You then use the Taylor Series Expansion around that mode and make the nonlinear random effect linear through approximation. You can then integrate it out of the model or estimate it or whatever you want to do with it.
    - The basic linearization approach above with the Taylor Series runs everything in the model as if it were a linear model. You are basically using a linear probability model with some arbitrary number of Taylor Series terms. The Laplace Approximation just does this with the random effect and then proceeds to use a standard generalized linear model (like a logit) conditional on the approximated random effect
- It's quick and reasonably accurate if the random effect is normally distributed or at least well behaved and symmetrical
- It has several of the same underlying flaws as the full linearization approach in that it's probably biased downward (e.g. it lies to you and tells you your random effects are smaller than they are).
- In a setting like a logit model this is bad because a misspecified random effect can bias everything else in the model. The more linear(ish) the model the less random effects misspecification hurts you (Litiere 2007).

Numerical Approximation via Gauss-Hermite Adaptive Quadrature

- Instead of trying to make the random effect fit a linear model so that you can integrate it out of a likelihood function most researchers use some variation on quadrature. This maps out the actual random effect directly. This can be done a lot of different ways (e.g. Simpsons Rule, the Trapazoid Rule, Gaussian quadrature, Clenshaw–Curtis quadrature).
    - Stata uses Gauss-Hermite Adaptive Quadrature so that's the one we are talking about
- You start with a particular value as a guess of some key point in the distribution (often the mode of the random effect) as estimated like you would from a Laplace approximation.
- You then go out a certain distance in either direction and pick another point (either equidistant as with Simpson's rule and the Trapezoid rule or based on Hermite polynomials of the random effects distribution).
- You fit a trapezoid between the two points and integrate inside the box using the straight lines to approximate the density under that section of the curve. You do this enough times and eventually you can draw a line between points that approximately fits the distribution of the latent random effect variable. If this sounds like approximating a distribution using something like a Taylor Series that's because it's the same basic idea of using a set of linear functions to approximate a nonlinear one. You just don't need a particular linearizing formula to approximate it with numerical quadrature.

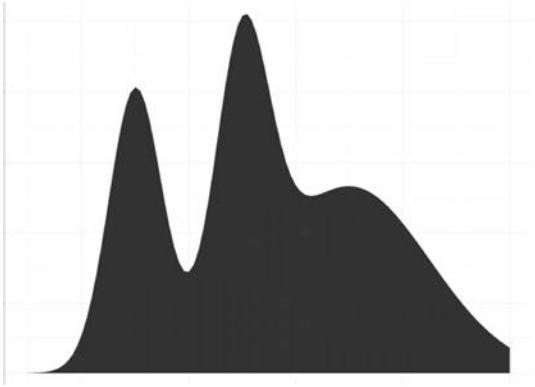


- Pick enough points and the distribution of the random effect will be approximated to as close to accurate as you want (you typically start with 7 and work your way up by odd numbers.) Some people (Rabe-Hesketh and Skrondal 2012) advocate dozens of quadrature points. The fewer points you pick the less accurate it will be if the random effect isn't normally distributed. You need a lot of points if it's not at least reasonably symmetrical.
- Quadrature methods tend to only work in low dimensions (a few random effects/coefficients). If you have more than that then it will take forever and probably just crash.



Approximation via Stochastic Integration

- Numerical quadrature works by mapping the distribution of the random effect directly through some variation on geometric approximation. It draws rectangles under the points in the distribution and then sums up the area within the rectangles. There's another approach to mapping the distribution of a random effect where the math operates at a relatively low level by constant repetitive work.
- Essentially, you pick a starting value like the mode of the random effect and then you wander around and keep a record of the shape of the random effect based on its value at every single place you've been.
- This technique is called Metropolis-Hastings and is often used to map out an entire likelihood density in a kind of random walk. Usually with some kind of stepping function that makes it more likely to walk up hill to denser regions than downhill.
    - There are a lot of flavors of Metropolis-Hastings samplers plus Gibbs Samplers, Quasi-MCMC based on Halton or Sobol sequences, and simulated annealing. They all try to map out a distribution instead of using a mathematical function to approximate it.



- It has the benefit that it doesn't assume any particular distribution on the data so it doesn't care if your likelihood function is bizarre or if the random effect isn't normal.
- You trade computational time for the ability to map the actual distribution and then integrate it out as usual.
- If you have a random effects distribution like the one pictured (which can happen when you mix levels improperly or do not use the Mundlak specification) then linearization methods, the Laplace Approximation, and Quadrature will all fail miserably the way any software program has them coded. You can make them work to a better degree but only by programming your own estimator based on an exceptionally high order Taylor Series. Stochastic integration doesn't care if your distribution looks like that. It will map it and then give you the right distribution anyway.
- Note that the entire basis for stochastic integration is the ability to map out the distribution. In order to do this in any reasonable way we have to rely on sampling theory. You don't want to literally walk over every possible value in the PDF or PMF because it would take forever. So we use sampling theory and just take a random sample. Except that we can't take a random sample because we don't have a sampling mechanism for the different points. So we take a correlated sample based on a random walk and use time series methods to make it look like a random sample. In order to really learn about this method you have to learn a bit about time series estimation ☹

Note that with each of these methods most software might only use the random effects estimates as starting values in an EM approach. Expectation maximization is a general and very flexible approach to dealing with missing data in maximum likelihood. EM isn't always used but if it is then you can trust the estimates a bit more assuming the model was properly specified in the first place (i.e. Mundlak specification). The final model is probably more accurate than the method used would lead you to believe because of the EM.

Steps in EM

- E-Step: Generate some estimate of the random effect
- M-Step: fit a maximum likelihood model to maximize the likelihood
- E-Step: generate new predictions for the random effect based on the model
- M-Step: maximize based on the new data set
- Repeat until the estimates don't change

Comparison of Methods

Method	**Accuracy of Your Answer**
Linearization PQL/MQL	Uses generalized least squares to estimate the RE then usually drops it into another model. With most software it is typically based on a first or second order Taylor series with some EM thrown in. The way it is done in HLM it is more accurate than Laplace but often still wrong. Biased fixed effects and variances will be too small
Laplace	Based on linearization of the integral itself instead of using GLS. It's similar to the former method because of the Taylor Series so the estimates will be close. Usually only accurate if your RE is close to normally distributed.
Quadrature	Lots of flavors. Adaptive Gaussian is generally the best but it involves difficult math and is more likely to fail when your RE's aren't close to normal or you have too many dimensions. Can be made accurate to an arbitrary standard so long as it doesn't crash
MCMC	Lots of flavors. Some version of MCMC will eventually get you the right answer. Eventually. Most common is Metropolis-Hastings but there are faster and more accurate variants like Hamiltonian Monte Carlo. Can be made accurate to an arbitrary standard so long as autocorrelation problem can be adequately dealt with.

Method	**Time to Your Answer**
Linearization PQL/MQL	Pretty fast but can depend on micro iterations and EM algorithm. About the time it takes to run a complicated linear model
Laplace	About as fast as the former method. Slightly more if using EM on top.
Quadrature	Regular quadrature is slooooow. Adaptive quadrature is an order of magnitude faster in some cases. The more points or random effects in your model the slower it gets. In complex models slow might become never.
MCMC	Can be extremely slow. Anywhere from "Go out to dinner and check when you get back" to "Go on vacation and check when you get back". Hamiltonian Monte Carlo is much faster than traditional when you have messy likelihood functions. Quasi-MCMC based on Halton sequences is about an order of magnitude faster than Metropolis Hastings.