Notes on Implementation of the SETE Scoring Model

Richard Herrington, PhD
UNT ACUS Research and Statistical Support

The model that is being used to create predicted means for each course is essentially a one-way mixed effects ANOVA (sometimes referred to as a "hierarchical linear model" or a "multilevel ANOVA" model. These kinds of models are very popular amongst population demographers, survey sampling specialists, spatial mapping settings, etc., for the reason that they give reasonable predictions whenever "small area estimation" problems are present. These are settings where the information in a region (either geographic, temporal, or logical) is sparse (low numbers of sampled units); however in contrast, neighboring regions might have large numbers of sampled units.

There are several reasons for choosing this kind of model for small area estimation problems.

If this were the ONLY estimation problem in these small area settings (i.e. varying response rates), then it might be sufficient to do some kind of inverse probability weighting (IPW) to reduce the effect of the varying response rates across the sampled strata or clusters. However, even if the problem were only one of wildly varying response rates, and all that was used was IPW methodology, modeled results would produce very "sensitive" predictions for course means (very poor precision - high variability - wide confidence intervals) - since the "case weights" used to do the IPW would in some cases be ONLY based on a few observations is some the really small areas that being estimated.

However, in these small area estimation settings there are other issues involved as well. Regional characteristics (ecological or contextual information) are often associated with these widely varying response rates. In technical terms, we say there are "within-strata" or "within-cluster" correlations present (for our purposes - "within course units"), that, if not taken into account, give poor estimates of prediction variability. This is sometimes referred to as the effect of "intra-class correlation" or "contextual effects". Even if one has "balanced data" (e.g. equal numbers of responses within courses), the intra-class correlation within strata or clusters can be a source of precision problems for estimates of the strata or cluster means. This is sometimes referred to as the "multilevel" structure of data (or equivalently, "hierarchical structure").

The regional characteristics that can produce "bias" in the estimated population effects (both point estimates and range estimates, i.e. confidence intervals or prediction intervals) are potentially:

1) course level or faculty level effects (e.g. course size, age of faculty, ranking, years at institution, etc.)

2) student level effects (expected grade in the course, mean GPA entering the course, actual assigned course grade, age, gender, etc.)

A major influence in our ability to estimate a "population" course mean accurately is the number of respondents within a course. In some courses there can be a very limited amount of information present (e.g. 2 or 3 respondents out 20 people enrolled in the course - e.g. 3/20); and in the same sampled data, perfect or nearly perfect responding rates - e.g. 20/20, or 19/20). Several issues come to the forefront at this point:

1) How does one calculate "useful" range estimates of future values of the course mean? By useful we mean reasonably narrow ranges of our expected values of future course mean estimates. Courses with very small response rates will, from the modeled data, give VERY wide range estimates, IF the range estimates can even be estimated.

2) How does one equate the amount of information in a course with a very small response rate, with a course that has a nearly perfect response rate? For example, is it reasonable to say that a course average of 4 based on 3/20 responses is similar to a course average of 4 based on 20/20 responses?

3) In those instances where course information is VERY small (2/20), how can we (or can we at all?) produce a useful estimate range estimate, RATHER than NO estimate at all?

It is for these reasons that in the small area estimation literature, linear mixed effects models and generalized linear mixed effects models are used with quite a bit of success. Mixed effects models allow the calculation of "shrinkage estimators" that can allow relatively unbiased point estimates, and range estimates where the strata sample sizes are widely differing.

For example, in our case, the implementation of the linear mixed effects model in R (package "lmer"), produces "empirical bayes based point estimates" of the course means. This is based on the simple weighting scheme:

empirical bayes estimate of course mean = course_mean * (course_reliability) + population_mean * (1 - course_reliability)

Here the course reliability is an estimate of THE single course's reliability (that takes into account the course size and the internal consistency of the course responses). The "shrinkage effect" (course means shrinking back toward the population mean) occurs whenever we have very little information on a particular course). For example, a course with a weighted mean of 4 and a reliability of .25 has shrunken mean of:

predicted course mean = 4 * (.25) + 3.3 (1-.25) = 3.475

In comparison, another course that has a weighted mean of 4 with a reliability of .75 has:

predicted course mean = 4 * (.75) + 3.3 (1-.75) = 3.475 = 3.825

This reliability is a function of the course sample size so, that smaller response rates (giving less information) get shrunk back toward the population mean, where we have greater certainty about a point estimate for an arbitrary (random) course mean from the population. This is an example of what is referred to in multilevel modeling as "borrowing strength across groups".

Bayesian Approaches to Mixed-Effects Model Estimation

Now, how do we get range estimates (i.e. prediction intervals) for courses whenever sample sizes for some courses are extremely small (say n=2)? Whenever sample sizes are extremely small for strata or clusters, problems can arise in the optimization phase of the modeling process. In mixed effects modeling software, confidence intervals (for our purposes, prediction intervals) are frequently calculated by solving the inverse of a set of second-order partial differential equations that are based on the log-likelihood function of the model.

Solving the "zero points" of the "second derivatives" amounts to finding the maxima or minima of this set of equations. Optimization techniques that use "hill climbing" methods can fail when the likelihood function are not properly convex (has multiple optima, or is not smooth, i.e. continuous). This can happen when the data is sparse and/or when the model has marginal quality of fit. The so called "Hessian Matrix" gives the standard errors for the parameters of the model, and these standard errors allow confidence intervals to be calculated for the parameters of the model. In a linear mixed effects model, it is the predicted "random-effects" (the course means) that we are interested in calculating (and their corresponding confidence intervals - or prediction intervals).

The problem with highly irregular data (e.g. sparse and ragged) is that the optimization of the hessian matrix may not reliably converge with varying data sets under varying conditions. In recent years (last 15 years approx.), a common approach to dealing with these estimation problems in small area statistics, is to use simulation based approaches to model estimation (Markov Chain Monte Carlo - "MCMC"). These models frequently have a Bayesian formulation, and once the model has been estimated, the posterior distribution can be used to simulate draws from the Bayesian posterior distribution for the model parameters, and these simulated model parameters can be used to generate outcome values for the response variable that are consistent with the assumed model and estimated population parameters.

For our model, this would be the overall, fixed effect population mean -- the population "grand course mean" and the random effect estimates for each of the courses. The estimated "random effect" would be the extent to which a course mean will vary away from the grand mean of all the courses due to "true variance" -- in other words, this random effect will be a function of the degree of intra-class correlation that is present within a course (e.g. varying away from the population mean, for a specific sample size with a specific intra-class correlation).

Without going into too much detail, the bayesian approach that is taken here (and as is implemented in the R package "lme4" with function "mcmcsamp") is to use MCMC simulation methods to draw "simulates" for the random effects (i.e. predicted course means) from the posterior distribution of the model parameters, as is estimated by the "lmer" function in package "lme4", in R. The function that implements this is "mcmcsamp". "Non-informative" priors are assumed for the model parameters (i.e. a relatively flat probability distribution for each of the parameter's population means and variances).

Based on draws from the prior distribution, an iterative "markov-chain" is formed, that uses bayes rule to combine parameter simulates from the prior distribution, and likelihood values from the fitted model with the data, that (in an iterative fashion) eventually terminates on the posterior probability distribution for the model parameters. This posterior distribution represents our best estimation as to what the population parameters are based on an initial estimate (ignorant guess - flat priors) that has been updated with data and fit with a likelihood function (using baye's rule) --a kind of "average " estimate of what we thought we would get and what we actually got. A lot of current packages use some form of Gibb's sampling (a particular variety of MCMC sampling) to get estimates of the posterior distribution. R package lme4's function "mcmcsamp" allows us to generate simulates from the "posterior predictive distribution" of the fitted model linear mixed effects model.

The payoff for using an MCMC based simulation approach is that one can get robust (always get an estimate) and reasonably accurate and useful confidence intervals (narrow with fairly good "coverage"), that incorporate ALL the different sources of uncertainty into that prediction interval estimates (e.g. parameter uncertainty, sample uncertainty -- i.e. sample size, and model uncertainty). The drawback is that since it is based on an optimization method that uses simulation, getting arbitrary degrees of numerical accuracy requires increasing the number of simulation draws. Another, way of saying this is that, rerunning the simulation can give slightly differing results.

This can be remedied by INCREASING the number of simulates so that the average estimate is replicable to a given order of numerical precision. I tried to balance the number simulates necessary to achieve a given degree replicability in the MCMC results -- that is, I tried BALANCE the run-time with the amount of Monte-Carlo variation that one would see upon replicating the results. I did this by exploring the variation in the quantiles (10% and 90%) and the point estimates as a function of

the number of simulates. Obviously, increasing the number of simulates increases the amount of time necessary to get "an answer". To help with this, I combined MCMC methods with "Ensemble Averaging" methods. Background reading on this method can be found at:

http://www.scholarpedia.org/article/Ensemble_learning

Using Ensemble Averaging Methods (e.g. Bagging) to Decrease Variability and Bias (this is similar too but not necessarily the same as using bootstrap resampling to calculate confidence intervals for model parameters)

Bagging (bootstrap aggregation) is a method that uses resampling methods (i.e. bootstrap) to refit a model multiple times, for the purposes of producing a weighted composite of the different model predictions. In simple bagging, one would resample with replacement from the original data B times; refit the model B times based on the B resamples; produce B prediction estimates from the B model fits; and finally, averaging (equal weighting) the B prediction estimates. Bagging has been shown to decrease bias and variance whenever the model fit is unstable (e.g. model fit based on sparse or ragged data, or data based on a "sensitive" nonlinear model fitting methodology).

For our purposes, I use bagging to reduce the dependence of the linear mixed effect model estimates on the sparseness of data in certain course units. The idea is that by "perturbing the data" multiple times, our average prediction estimate will be more generalizable to sparse data sets that might occur randomly in future samples. For our purposes, I treat the strata (courses units) as fixed structure, and resample with replacement within each course, to produce a data set that has the same number of course units with the same sample sizes within each course unit -- that is, after resampling, the data set is different, but with the same number of rows across course units, and the same number of rows within course units, but with varying data overall from resample to resample. "B" "lmer" fits are produced, and "B" MCMC posterior prediction estimates are produced, with "B" estimates of the 10th percentile, "B" estimates of the 90th percentile, and "B" estimates of the expected posterior course mean. A simple average of the "B" estimates for the 10th percentile, an average of the "B" estimates of the 90th percentile, and an average of the "B" estimates of the expected posterior course average is produced.

Right now, B=30 estimates, and the number of MCMC draws are 200. I have explored increasing the number of MCMC draws, but find that 200 is adequate in producing replicate estimates (converged estimates). "B" could be increased to reduce the "Monte Carlo" variation in the quantile estimates, but there is a diminishing returns effect on increasing this as well.

Note that all the expected course means will vary naturally within the stated prediction intervals from simulation to simulation. But the Average of those means

should be stable for large course sample sizes. Note that courses that have extremely small n's (n=2 for example) will naturally vary from simulation to simulation WITHIN the confidence interval (just like for large sample sizes), however, the mean of those expected course means will vary slightly BECAUSE the sample size is so small -- but they don't vary by much. The exact estimate being perfectly replicble, isn't as important, as much as the fact that they should vary 90% of the time (from simulation to simulation) within the stated intervals.

Essentially, each time the MCMC estimate is run, it is like generating new data from a future sample. Of course the data will be slightly different. This more closely mimicks the real world.

So to summarize, we can increase the both the MCMC simulate draws (200 now) and the number of "bagged" model fits (30 now) and this will reduce the variation in the "mean of the means" (the bagged - averaged - course means). But we shouldn't worry too much about this for small course sizes, because there is nothing we can do about it, the point estimate on a particular run will suffice as a set of predictions for the WHOLE SET of courses.

Case Weighting

In addition to using "shrinkage methods" (as part of the linear mixed effects model), we also use a method that produces case weights such that the observed sampling rates are adjust for what is called "unit non-response" (for example in a class of 20 confirmed enrollees, if only 17 respond to the evaluation, then there are 3 unit non-responses). These "adjusted" case weights are then used in a post-stratification procedure where weighted resampling is done (using the case weights) to produce multiple weighted resamples where the effect of non-response has been (hopefully) removed as much as possible.

This technique is based on what is called "Response Homogeneity Groups" (or RHG for short) -- it is based on estimating the probability of responding to a survey based on auxiliary variables (course, faculty and student variables that are collected on ALL enrollees) -- and from estimated probability of response to the survey, we can adjust for the probability of non-response to the survey. This can be used IN CONJUNCTION with the linear mixed effects model fits and its shrinkage based estimators. This is a well-known technique that (if done properly) can significantly reduce bias in the course means (and increases precision) in the point estimates of the course means.

In each of these multiple, stratified samples, we estimate our linear mixed effects model that I outlined above. So when I talk about using B=30 bagged estimates of the linear mixed effects model, these are actually 30 weighted bootstrap drawn samples (with courses) -- using the RHG adjusted sample weights as the bootstrap

resampling weights -- the weights determine the probability that a particular case will appear in a particular resample).

For each, weighted bootstrap resamples, we will get "an estimate" of the predicted means for each course. Averaging across these predicted "course mean estimates", or "bagging", will produce a single set of course estimates that are less biased and have less variability associated with them THAN if we DIDN'T use the aggregation of these 30 predicted.