MODEL-BASED APPROACHES TO SMALL AREA ESTIMATION WITH BINARY DATA

Charles E. McCulloch, Departments of Statistical Science and Biometrics,
Cornell University, 434 Warren Hall, Ithaca, NY 14853

Keywords: finite population sampling, penalized quasi-likelihood, best prediction

1. Introduction

Small area estimation is a long-standing problem in survey sampling which arises in a variety of contexts, including accurate estimation of quantities for municipalities or census divisions, estimation for areas which are small in spatial extent, or, more generically estimation of stratum level effects. Special approaches to small area estimation are needed for surveys in which (at least some) strata have very small sample sizes. In such a case, direct estimates of strata effects (using only the data from each stratum) are likely to be highly inefficient and techniques which "borrow strength" across strata may be advantageous.

Let me begin with an example which is richly featured and serves to illustrate some of the difficult decisions involved in choosing a satisfactory analysis for such data. Habitat Conservation Plans (HCPs) are agreements between non-federal landowners and the U.S. Government which allow incidental taking of endangered species as long as the taking of such species is minimized and mitigated. These were authorized in 1982 under a modification to the Endangered Species Act. From their implementation in 1982 until September 1997, 225 HCPs had been approved. Many conservation activists and independent scientists have charged that HCPs are not based on sound science and are not aiding in the recovery of endangered species (Mann and Plummer, 1997).

To address this concern, a group of ecologists headed by Peter Kareiva selected a sample of 43 of the plans and embarked on a systematic assessment of the use of scientific reasoning in HCPs (Mann and Plummer, 1997). Because of the magnitude of the effort, sampled HCPs were distributed to eight universities for assessment. To further complicate the issue, each HCP covers from one to many species. For HCPs with a single species that species was evaluated. For HCPs with multiple species, a sample of the species was evaluated. To fix ideas, I will focus on a single question among the multitude which were assessed. For each species covered by the HCP, I will consider the following question: "Is there an unambiguous plan to change the HCP strategy in response to new monitoring information?"

For the HCP example the individual HCPs form our strata and we have very few (often only one) subunit per stratum. Small area estimation in this context means estimation for each of the HCPs of the proportion of species for which there is an unambiguous plan.

The goals of this paper are to describe some mixed models appropriate for the analysis of binary survey data and compare and contrast estimation methods for those models. The estimation methods considered are maximum likelihood (ML), generalized estimating equations (GEEs), penalized quasi-likelihood (PQL), and Bayes. I will make a few comments in Section 2 but otherwise not attempt to discuss the broader issue of model- versus designed-based inference.

2. Model-based versus design-based inference

There is a long-standing debate on of the merits of model- versus design-based inference in survey sampling. For an excellent discussion paper in the context of small area estimation see Ghosh and Rao (1994). I would like to acknowledge these fundamental differences and the fact that there are distinct advantages to each approach without spending significant amounts of time arguing them. I will instead mostly concentrate on comparing methods of model-based estimation.

However, I do want to point out several features of the HCP example which make model-based inference moderately attractive. First, small area estimation, given the fundamental idea of borrowing strength across strata, is a situation where model-based inferences are, perhaps, somewhat more compelling. Second, while the 43 sampled HCPs are obviously selected from the finite population of 225 HCPs, we might be interested in regarding the 43 as a sample from the (conceptually) infinite population of HCPs which could be accumulated if policies and situations were to stay the same. Using a model-based approach is natural for this latter case and, even for the finite population case, makes for a more convenient comparison of the finite and infinite population cases.

Finally, there is the post-sampling complication of assessment of the HCPs in eight groups. Viewed one way, this introduces a (necessarily model-based) correlation among all the assessments within a group, which must be accommodated. Viewed another way, we would certainly want to make inferences beyond the particular raters who assessed the plans and would like to regard school as a random effect.

3. Methods of model-based analyses and comparisons

Suppose we have decided on a model-based approach. There are several competing methods for fitting such models, including maximum likelihood, GEEs, penalized quasi-likelihood, and Bayes methods. What are the advantages and disadvantages of each?

3.1 A model

First we need to describe a prototype model against which to frame the discussion. Consider the HCP example where the response is yes or no to the question about the existence of a response plan for new monitoring information, clearly a binary variable. To acknowledge the binary nature of the response, we will need to assume a Bernoulli distribution as the marginal distribution for the data. Let Y_{ijk} be 1 if the response to the question was yes for species k in HCP j from school i. We therefore have

$$Y_{ijk} \sim \text{Bernoulli}(p_{ijk}).$$

The predominant way to model such a setting is to build a mixed model using a random effect for stratum (Ghosh and Rao, 1994). Unsampled strata or subunits are then regarded as values to be predicted using the model.

For our HCP example we will also need a random effect for school. A convenient way to model school and HCP effects and to broadly allow the inclusion of covariates is to use a generalized linear mixed model. For a generalized linear model the next step is to decide how to link the probability of a "yes" response with the school, HCP and covariate effects. A possible, though by no means exclusive way is to assume that a linear mixed model applies to the logit of p_{ijk}:

$$\text{logit}(p_{ijk}) = x'_{ijk}\beta + s_i + h_{j(i)},$$

where x_{ijk} is a vector of fixed covariates (which might be observation specific), s_i are the school effects, and $h_{j(i)}$ are the plan-nested-within-school effects. These latter two are going to be assumed to be random effects. So to them we assign a distribution, which we will choose to be normal, though others are possible:

$$s_i \sim i.i.d.\ N(0,\sigma_s^2)$$
$$h_{j(i)} \sim i.i.d.\ N(0,\sigma_h^2).$$

If we are to perform a Bayesian analysis, then we would need prior distributions for β (typically a normal distribution if the random effects and errors are assumed to be normally distributed) and prior distributions for the hyperparameters in the distributions of the β's, the s_i, and the $h_{j(i)}$ (i.e. for σ_s^2). We now consider methods of estimation for this model.

3.2 Maximum likelihood estimation

A very common method of estimation for linear mixed models is maximum likelihood (ML) or variants like restricted maximum likelihood (REML) either of which is typically based on the assumption of normally distributed random effects and errors. For example, the package SAS fits such models using PROC MIXED and has either ML or REML options. Likewise, for many generalized linear models, maximum likelihood is also the method of choice. For example, a logistic regression or Poisson regression model is invariably fit using ML. What about generalized linear mixed models? Unfortunately, for the model of Section 3.1 the likelihood cannot be written in closed form.

When the model has a single random effect or two nested random effects (our HCP example has schools and plans nested within schools), it is relatively easy to numerically evaluate the integrals in the likelihood. For example, with a single random factor (e.g. strata in the simplest small area estimation situation) the likelihood is a product of one-dimensional integrals. One can then maximize the likelihood numerically to find ML estimates and to perform likelihood ratio tests.

When there is a single, normally distributed random effect, the likelihood is the product of integrals of the form:

$$\int_{-\infty}^{+\infty} g(x)\exp\{-x^2\}dx\,.$$

These can be accurately evaluated using Gauss-Hermite quadrature that approximates the integral with a summation:

$$\int_{-\infty}^{+\infty} g(x)\exp\{-x^2\}dx \approx \sum_i w_i g(x_i)\,.$$

The weights, w_i, and the evaluation points, x_i, are given in references, i.e., Abramowitz and Stegun (1964).

If the ML estimates can be calculated numerically, then inference using them would proceed using the usual asymptotic approximations:

- ML estimates are asymptotically normal, with SEs coming from second derivatives of the log likelihood.
- Tests would be based on the likelihood ratio test, comparing -2loglikelihood for nested models.
- Results for testing variance components are the same as the linear mixed model. (Being careful as to the large sample distribution of the likelihood ratio statistic!).
- Best predicted values would be estimated by calculating E[random effect|data] and plugging in ML or REML estimates for the unknown variance parameters. In general, the conditional expected values can't be evaluated in closed form either.

This last step can be problematic because it makes it difficult to calculate SEs for the best predicted values that incorporate the extra variability associated with estimating the variance components.

In summary, ML estimation

- Has known large sample properties,
- Can be used with likelihood ratio tests,
- ◊ Can be hard to compute for many generalized linear mixed models,
- ◊ Must have its small sample performance assessed for any particular model,
- ◊ Has difficult to assess SEs for prediction.

ML estimation is not widely available, but there are some special purpose packages, for example MIXOR (available free from www.uic.edu/~hedeker/mix.html) and SABRE (available from www.cas.lancs.ac.uk/software/sabre3_1/sabre.html/) that perform such computations.

3.3 Generalized estimating equations

GEEs are a computationally less demanding method than ML estimation. They are applicable (mainly) to longitudinal data, where I define

Longitudinal data = data collected on a subject on two or more occasions, and the number of occasions is typically small compared to the number of subjects.

GEEs work most easily for models specified on the unconditional distribution. In contrast, we have been specifying models that are conditional on the random effects.

For our HCP example, the identification with longitudinal data is that strata are the equivalent of subjects and subunits within a stratum are analogous to the repeated measures on the subject. We could specify:

$$\begin{aligned} \mathrm{E}[Y_{ijk}] &= p_{ijk} \\ \mathrm{logit}(p_{ijk}) &= x'_{ijk}\beta \end{aligned} \qquad (1)$$

and use this mean specification along with an empirically estimated correlation structure. This may look the same as our model in Section 3.1, but it is not quite. In Section 3.1, the conditional probability of a yes is assumed to follow a logistic form, whereas in (1), it is the unconditional probability that is modeled as a logistic form. More importantly for small area estimation is that the model (1) does not explicitly include any random effects to facilitate the estimation for a small area. Therefore, to use GEE estimation for small areas we must identify an approximate model of the form given in (1) which corresponds to the model of Section 3.1 (Zeger, Liang, and Albert, 1988). We then estimate the model in the form of (1) but have to do further work to get estimates of the random effects variances and covariances, which are needed to estimate the best predicted values. Methods are described in Zeger, Liang and Albert (1988, Section 3.2).

A key feature of GEE estimation and the reason it is commonly used for longitudinal data is that it uses an empirical estimate of the within stratum correlation. This empirical estimate is built up from the replication across independent blocks of data. Since it is empirical is does not depend on strong model assumptions and is robust in that sense. Situations where the data do not break up into a large number of relatively small and independent blocks are not amenable to the use of GEEs (Diggle, Liang and Zeger, 1994, p.77). For example, the HCP example would only break up into eight blocks (for the eight schools), each of which would be of a different size and configuration of plans and species within plans. Hence it would not be amenable to GEE estimation.

GEE estimation is available in several common statistical packages, for example, SAS (in PROC GENMOD), SUDAAN, STATA, and S-Plus. In summary GEEs

- Have robust standard errors,
- Are often relatively efficient,
- ◊ Work best when the data can be broken up into a relatively large number of blocks, each with

relatively homogenous arrangement of a small number of observations

◊ Are easiest for marginal models, not random effects models.

3.4 Bayes estimation

A number of authors (e.g., Datta and Ghosh, 1991; Hulting and Harville, 1991) have argued for the superiority of Bayes methods over frequentist based methods. This seems to revolve mainly around the difficulty of assessing prediction error with estimated variance components within frequentist based methods. I regard the decision as more basic and more philosophical and do not want to raise the long-standing Bayes versus frequentist debate. However, I do have some caveats on the use of Bayes procedures.

Some statistical workers have wanted to "have their cake and eat it too" in the sense that they want to take advantage of the straightforward way in which Bayes procedures can be developed and can handle problems like estimated variance components but do not want to inject (possibly subjective) information in the form of a proper prior distribution for the parameters. The usual solution is to hypothesize flat, improper, non-informative, or diffuse prior distributions and to "let the data speak for themselves."

This can cause problems. Flat and other improper priors for variance components can cause the posterior to fail to exist, rendering the Bayes methodology useless. This has been demonstrated in both the linear mixed model (Hobert and Casella, 1996) and mixed models for binary data like the one described in Section 3.1 (Natarajan and McCulloch, 1995). This is not to say that *all* improper priors lead to improper posteriors, but just that care needs to be taken when improper priors are considered.

A suggestion for the avoidance of improper priors is to use diffuse priors instead. For example, one might choose a flat prior truncated within some range, or a normal distribution prior whose variance is quite large. Unfortunately, these can lead to problems also. For Bayes estimation for the model of Section 3.1, a typical way to calculate Bayes estimates is through the use of the Gibbs sampler. Natarajan and McCulloch (1998) show that there are data sets for which the Gibbs sampler either breaks down because the prior is too "close" to improper or it converges, but to a posterior that is influenced by the prior. Said another way, there is no happy middle ground were the analysis is not influenced by the choice of the prior but the Gibbs sampler still works. Worse yet, in some of the situations in which the Gibbs sampler fails it gives no obvious signs that something is amiss. It can even be run for the improper posterior situation within giving any obvious warning.

Again, I do not mean to imply that all diffuse priors cause problems, but merely that it is possible. This is unfortunate, because Gibbs samplers are used in such a context precisely because analytic results are hard to derive. If analytic results were available, we would be able to avoid the numerical problems and be forewarned about the nonexistence of posterior distributions.. Or if the Gibbs sampler behaved in an anomalous way for such problems we would at least have a warning that something was amiss.

Bayes estimation is not widely available in standard software though the package BUGS for constructing analyses is available from the web site http://www.biostat.umn.edu/mirror/methodology/bugs. In summary, Bayes procedures

- Are able to incorporate prior information,
- Can straightforwardly accommodate unknown variance components,

◊ Can encounter numerical problems.

3.5 Penalized quasi-likelihood

Quasi-likelihood estimation has gained wide popularity in the fitting of generalized linear models (McCullagh and Nelder, 1989). This popularity is rightly deserved since the validity of the estimation depends only on the mean to variance relationship and not on further model assumptions. Further, it is often fully or highly efficient compared to the optimal, model-based procedures, so the robustness to model variation comes at a small or zero price.

However, the strength of quasi-likelihood estimation is a weakness when it comes to random effects models. Since there is nothing specified in the model concerning the distribution of the random effects the methods must be modified for use in mixed models. Thus has arisen the idea of penalized quasi-likelihood (PQL) in an attempt to maintain the model robustness of quasi-likelihood with regard to the mean structure while building in minimal assumptions about the random effects. Roughly, to the quasi-likelihood is added a "penalty" term which forces the random effects to behave somewhat as if they were selected from a distribution.

Despite early promising work (Schall, 1991; Breslow and Clayton, 1994), PQL has not generated estimators with good properties. For binary data it can often give highly biased estimators (Rodriguez and Goldman, 1995; Breslow and Lin, 1994). They thus cannot be recommended in practice.

PQL is available via the SAS macro GLIMMIX and in the packages MLn, Varclus, and HLM. In summary, PQL

- Is computationally fairly easy,
- ◊ Does not work well for highly non-normal data (e.g. binary),
- ◊ Is mainly for normally distributed random effects.

4. Conclusions

The conclusions can be easily stated. Mixed models are relatively straightforward to specify for binary data by adding random factors to a generalized linear model. By adding in random strata effects, such a model can be quite useful for small area estimation.

If a model-based approach is taken to small area estimation with binary data, then some care is needed in estimation of the model. Maximum likelihood, proper prior Bayes procedures, and generalized estimating equations (where appropriate – see below) are the methods of choice. Penalized quasi-likelihood methods cannot be recommended in practice. Bayes methods with improper or diffuse priors should be used with care and generalized estimating equations should only be used when the data break up into a relatively large number of blocks, each with relatively homogenous arrangement of a small number of observations.

Acknowledgement: This work was partially supported by NSF DMS 96-25476

5. References

Abramowitz, M. and Stegun, I.A. (1964). Handbook of Mathematical Functions. National Bureau of Standards, Washington, D.C.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**: 28-36.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9-25.

Breslow, N.E. and Lin, X. (1994). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**:81-91.

Datta, G.S. and Ghosh, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Annals of Statistics* **19**: 1748-70.

Diggle, P., Liang, K.-Y., and Zeger, S.L. (1994). Longitudinal Data Analysis. Oxford University Press, Oxford.

Ghosh, M. and Meeden, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association* **81**: 1058-62.

Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science* **9**: 55-93.

Hulting, F.L. and Harville, D.A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association* **86**: 557-568.

Mann, C. and Plummer, M. (1997). Qualified thumbs up for habitat plan science. **278**: 2052.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models, 2nd Ed. Chapman and Hall, London.

Natarajan, R. and McCulloch, C.E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* **82**: 639-43.

Natarajan, R. and McCulloch, C.E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? In press in *Journal of Computational and Graphical Statistics*.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A.* **158**: 73-89.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**: 719-727.

Zeger, S.L., Liang, K.-Y., Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation. *Biometrics* **44**: 1049-60.