SUBSTANCE ABUSE IN STATES: A METHODOLOGICAL REPORT ON MODEL BASED ESTIMATES FROM THE 1994-1996 NATIONAL HOUSEHOLD SURVEYS ON DRUG ABUSE

Ralph Folsom, Babu Shah, Akhil Vaish, Research Triangle Institute
Ralph Folsom, RTI, P. O. Box 12194, Research Triangle Park, NC 27709

Key Words: Small Area Estimation, Gibbs Sampling, Survey Weighting

Introduction

This paper summarizes the 1994-1996 small area estimation project completed recently for the Substance Abuse and Mental Health Services Administration. The main purpose of this project was to correct deficiencies in the previous NHSDA small area estimation exercise which used the penalized quasi-likelihood (PQL) estimation approach. In the current project, hierarchical Bayes estimation methods employing Gibbs sampling were used. A secondary purpose was to test the new methodology on pooled 1994-1996 data. This methodological test was not designed to published small area estimates but to begin preparation for the 1999 NHSDA. Beginning in 1999 SAMHSA will require that small area prevalence estimates be produced on a tight schedule for states and age groups for up to 20 drug use related binary outcomes.

In the current test, we produced prevalence estimates for sixteen drug use related outcomes by age group for all 50 states and the District of Columbia. Our new pseudo-Bayes estimation methodology retains the innovative features for our previous approach. These innovations include fitting Logistics mixed models with fixed predictions a four levels of hierarchy; namely,

- person level demographics,
- Census Block group level demographic population projections for 1996,
- 1990 Census Tract level demographic and socioeconomic status (SES) variables, and
- Intercensal County level variables including drug related arrest, treatment and death rates.

Microdata modeling of the 1/0 binary outcomes also allowed for the efficient inclusion of interactions between the person level demographic indicators and the block group, census tract, and county level predictors. This microdata modeling of prevalence at the block group level for our 32 age by gender by race/ethnicity demographic domains allowed the State SAEs to be built up from their block group level contributions. Finally, our method of employing survey weights in the PQL estimation equations yielded SAEs for States and their national aggregates that were design consistent. For States with large samples and for national aggregates of SAEs, this made our results self-calibrating to the robust design based estimates.

Pseudo-Hierarchical Bayes Estimation

Our new pseudo-Bayes estimation methodology was designed to remove the limitations of our in 1991-1993 approach that were noted in the Methodology Report produced by Folsom and Judkins (1997); namely

- Computational limitations of our PQL algorithm required the assumption that the age group specific random effects were independent of each other. Specifically, if

$$\text{logit}\,[\text{Prob}(y_{aijk}=1\,|\,\eta_i,\nu_{ij})] = x_{aijk}\beta_a + \eta_{ai} + \nu_{aij}$$

is the logistic mixed model for the probability that age group-a member-k of PSU-j in State-i has $y_{aijk}=1$, then to speed up convergence we were obliged to subset the data by age group to minimize the size of the associated β vector.

- Our 1991-1993 approach to survey weighting was also not robust against unequal selection probabilities that are informative in the sense of being correlated with the survey outcomes in a fashion not fully accounted for by the model specification.
- The first order PQL solution method we employed relies on the random effect variance components being small for its estimates of the components to be accurate. Otherwise, the associated variance component biases lead to biased SAE point and interval estimates.
- Perhaps most importantly, our 1991-1993 approach to SAE interval estimation ignored the uncertainty in the variance components. Malec, et. al (1993) show that this source of inflation in SAE interval estimates can be sizable.

To overcome these limitations, we developed a survey weighted full hierarchical Bayes (FHB) algorithm (PROC GIBBS) for fitting our large logistic mixed models. PROC GIBBS has the following benefits relative to available FHB software for the logistic model like BUGS and MLwiN.

- Uses survey weights so that SAE point estimates closely match direct survey weighted prevalence

estimates for areas with large samples, like California and the entire USA.

- PROC GIBBS's survey weighting also causes the associated posterior predictive density intervals for SAEs to match fairly closely the design based confidence intervals for large sample areas or aggregates.
- Unlike BUGS or MLwiN , our GIBBS procedure is able to fit our four age group specific models simultaneously. Note that these models can have upwards of 200 fixed effects, close to 1,000 random effects, and 20 variance/covariancecomponents. We never got BUGS or MLwiN to complete a single gibbs cycle on such models.

To specify our GIBBS algorithm we assume that the State and PSU level random effect vectors η_i and v_{ij} with age group specific elements are four variate normal with null mean vectors and 4x4 covariances matrices D_η and D_v respectively. We assign an improper uniform prior to the age specific fixed coefficient vectors β_a, and use proper inverse Wishart priors for D_η and D_v. If we let $v_o$define the prior degrees-of-freedom parameter in the Wishart prior for D_η. then we define $D_{\eta o}$ as the prior mean matrix for D_η. With the prior degrees-of-freedom for D_η set to v_o=6 and p, the rank of D_η, equal to four, the prior density for $T_\eta = D_\eta^{-1}$ is proportional to $\det(T_\eta)^{1/2} \exp\{-(0.5)\,\mathrm{tr}\,(D_{\eta o} T_\eta)\}$. Since the prior mean of $(T_\eta)^{-1}$ is $[D_{\eta o} \div (v_o - p - 1)]$ and $(v_o - p - 1) = 1$, we see that $D_{\eta o}$ is indeed the prior mean for D_η. With 46 states plus the District of Columbia in our sample and assuming that the associated η_i were observed for i=1, ..., 47, we can define

$$\hat{D}_{mle|\eta} \equiv \left(\sum_{i=1}^{47} \eta_i \eta_i^T \div 47 \right) \qquad (1)$$

and then specify the conditional (given the η_i) posterior mean for D_η as

$$E\left(D_\eta \mid \underset{\sim}{\eta}, v_o = 6, D_{\eta o}\right) = \hat{D}_{mle|\underset{\sim}{\eta}} + \left(D_{\eta o} - \hat{D}_{mle|\eta}\right) \div 48 \quad (2)$$

Therefore conditional on observing the η_i state level random effects we see that the conditional posterior mean for D_η is the associated mle plus a term t hat contracts the mle matrix with the prior matrix all divided by one plus the number of states in sample (47).

Note that I have defined the conditional mle for D_η in equation (1) without using survey weights. Since states were largely included with certainty in the pooled 1994-1996 NHSDA, weighting is not an issue for D_η. For the between PSU covariance matrix D_v we considered using the PSU weights to form the conditional pseudo-mle analog of equation (1). Since the $(CV)^2$ of our PSU weights exceeds 3 we decided that the associated loss in effective sample size for estimating D_v would be too serve; that is, for our example, the elements of D_v would have covariances with sample size divisors of 49 effective PSUs instead of 196. Using the unweighted priors for our random effects and the inverse Wishart priors for the D matrices yields conditional posteriors that have the inverse Wishart form. Therefore, we can draw Gibbs samples directly from these inverse Wishart conditionals for D_η and D_v.

For the age group specific fixed coefficient vectors β we have assumed a flat prior. We can therefore sample from the conditional posterior given the survey weighted pseudo-mle that SUDAAN would produce by conditioning on the random effects from the previous Gibbs cycle. Since our age group specific sample sizes were quite large (13.5k) we used the asymptotic normal version of the conditional posteriors to select Gibbs samples for these parameter vectors. Note that conditional on the random effects, these age specific β vectors are independent of each other and can be sampled independently. Taking advantage of this result speeded up our algorithm substantially.

Conditioning on the state and PSU level random effects, we use the asymptotic covariance matrix for β_a that is produced by SUDAAN for an unequal probability single state sample from the age group. While this sandwich type asymptotic covariance matrix accounts for survey weight variation and lack-of-fit in the propensity model, it otherwise may not fully account for sample design induced clustering effects, particulary those that are not explicitly modeled by random effects. It is clear, for example, that our decision to give up on including area segment associated random effects in our NHSDA logistic models could result in some underestimation of SAE mean squared errors. Our sample sizes in these final stage clusters were small ($\bar{m}$ = 9 or 10) and solutions typically could not be obtained for the low prevalence outcomes. In spite of this shortcoming, for California, the state with the largest sample where our SAE point estimates and the direct survey estimates are always close to each other, our pseudo-Bayes posterior prediction intervals were also close to the survey design based confidence intervals. This result will be illustrated in the subsequent validation section.

Turning to the Gibbs Cycles for the random effects, I will illustrate our approach on the state level random effects. The basic idea is to use the survey weighted conditional pseudo-likelihood for η along with its multivariate normal prior to form the kernel distribution for the Metropolis-Hastings step of the Gibbs sampler. For this to work properly we have to scale the survey weights in the pseudo-likelihood so that the correct asymptotic covariance matrix is achieved for η. We then

use an adjusted form of the asymptotic normal conditional for η to select the initial draws in the Metropolis-Hastings algorithm. Using the asymptotic covariance for the conditional pseudo-mle as a guide, the required effective sample size adjustment to the age group weights uses the $\bar{\delta}$ and cap Δ quantities defined in Figure 1. The $\tilde{\Delta}$

Figure 1. Effective Sample Size Adjustment Factors

$$\bar{\delta}W_{ia} \equiv \sum_{jk} W_{ijk}\hat{\pi}_{ijk}(1-\hat{\pi}_{ijk}) \div W_{i++}$$

$$\bar{\delta}WW_{ia} \equiv \sum_{jk} W_{ijk}^2\hat{\pi}_{ijk}(1-\hat{\pi}_{ijk}) \div \left(\sum_{jk} W_{ijk}^2\right)$$

$$\bar{\Delta}WW_{ia} \equiv \sum_{jk} W_{ijk}^2(y_{ijk}-\hat{\pi}_{ijk})^2 \div \left(\sum_{jk} W_{ijk}^2\right)$$

$$\tilde{\Delta}WW_{ia} \equiv \left[\left(\sum_{i=1}^{47} \bar{\Delta}WW_{ia}^2\right) \div \left(\sum_{i=1}^{47} \bar{\delta}WW_{ia}\right)\right]\left(\bar{\delta}WW_{ia}\right)$$

quantity involves a smoothed estimate of the informative weighting effect calculated as a combined ratio over all 47 states. Given these definitions, the effective sample size for age group-a in state-i is given by equation (3).

$$em_{ia} = m_{ia} \div \left[\left(1+CVW_{ia}^2\right)\left(\tilde{\Delta}WW_{ia} \div \bar{\delta}W_{ia}\right)\right] \qquad (3)$$

These effective sample sizes are used to produce scaled versions of the survey weights that sum to em_{ia} within state-i by age-group-a combinations. This weight scaling leads to the kernel log density definition shown in equation (4). The weight adjustments for the PSU level random effects follow the same prescription.

$$\ell n\left\{F(\eta_i|\,)\right\} = -(1/2)\eta_i^T D_\eta^{-1}\eta_i + \sum_{jk} w_{ijk}\left\{y_{ijk}\ell n[\pi_{ijk}(\eta_i|\,)] + (1-y_{ijk})\ell n[1-\pi_{ijk}(\eta_i|\,)]\right\} \qquad (4)$$

Application of PROC GIBBS to the 1994-1996 NHSDA

Turning to our application of PROC GIBBS to the pooled 1994-1996 NHSDA survey data, we produced prevalence estimates for sixteen binary outcomes related to drug use, dependency, and treatment need. These three survey years of the NHSDA represented 46 states plus the District of Columbia and included 196 MSA or single county PSUs and 4,030 area segments for a total of 53,825 respondents. We fit age specific random effects for states and PSUs but not for segments. As indicated previously, the segment sample sizes of nine or ten were not large enough to support estimating segment level random effects since most of our binary outcomes have small prevalences, less than 5%.

For independent variables we had four major groups. The categorical variables included the personal level demographics along with several categorical classifications of the PSUs, segments, or tracts that the respondents lived in. The 36 county level variables were intercensal values obtained from various sources. We also used 26 tract level variables from the 1990 decennial census and 12 block group level variables based on 1996 intercensal population projections purchased from Claritas, Inc.

For variable selection, we group our independent variables into six initial groups and used SAS logistic with sample size sealed weights and backwards elimination to identify individual variables significant at three levels; i.e., 1%, 3%, and 5%. We then looked at measures of fit related to concordance and discordance for user and nonuser pairs. Most of the selected models were the 1% significance versions with the total number of fixed predictors ranging from 60 up to 175. We then ran PROC GIBBS on each of the sixteen outcome specific models. To speed convergence to stationary chains we only updated the fixed coefficient vectors β_a every eighth cycle. By examining the D_η and D_v matrices we determined that long run chains with a burn in of 500 cycles and a sample run of 10,000 yielded good comparability between the posterior distributions observed in the first and second halves of the chain. These results led us to use the 1,250 PROC GIBBS sample cycles with distinct β_a vectors to produce our small area estimates. Specifically, with θ_c denoting the set of fixed and random effect estimates from PROC GIBBS cycle-c and -b indexing census block groups in state-i, we formed prevalence estimates $\pi_{ibd}(\theta_c)$ for each of the thirty-two age by gender by race/ethnicity subpopulations indexed by-d. Since the NHSDA sampling rates within block groups are negligible, these π_{ibd} predicted prevalences can be weighted by associated population projections N_{ibd} and aggregated to yield state level SAEs for the numbers of persons with the attribute and for the associated prevalences. For block groups-b that are not in NHSDA sample PSUs, we sampled a value of v_{ij} from the conditional posterior for v_{ij} which is four variate normal with null mean vector and covariance matrix $D_v(c)$ at cycle-c. We purchased the block group v level population projections N_{ibd} from Claritas, Inc.

Validation of SAEs

To validate our small area estimation methodology we first examined the extent to which our survey weighting approach was successful in causing the regional and national aggregates of our SAEs to be close to their associated Horvitz-Thompson prevalence estimators. Figure 2 displays, for the indicator of past month any illicit drug use, a set of signed relative bias estimates for weighted and unweighted SAE regional, national, and California aggregates by age group. In this figure, the Horvitz-Thompson prevalence estimate is treated as the benchmark. It is clear that the survey weighted versions are, as a rule, substantially closer to the Horvitz-Thompson benchmarks than their unweighted counterparts.

To compare our state level SAEs against external benchmarks, we secured a set of 1993 person level food stamp participation rates for States. Using the NHSDA questionnaire reports of food stamp participation to derive a person level indicator, we then produced state level SAEs using our modeling approach. When we contrasted our SAEs with the benchmark we obtained a correlation of 0.91 whereas the direct Horvitz-Thompson prevalence estimates had a correlation of 0.67.

Finally, we conducted a bootstrap subsampling exercise to create five subsamples from California that mimic the design of the 1999 NHSDA as it applies to the 42 small sample states and the District of Columbia. In the 1999 NHSDA, these small sample states will have samples of 96 area segments drawn in sets of eight from each of twelve substate regions referred to as field interviewer or FI regions. Dwellings will be drawn from these segments and persons selected from the dwellings so as to yield 900 interviews from each small sample state with 300 responses targeted from each of three age groups; namely 12 through 17 year olds, 18 through 25 year olds, and persons aged 26 and older. Having drawn five such independent subsamples from our pooled 1994-1996 California sample we then refit weighted and unweighted solutions to the past month any illicit drug use model with five replicated versions of the California data in the sample. We then used the PROC GIBBS parameter solutions to form separate SAEs based on each of the five California data sets as if California had split into five new States. Figure 3 shows that our SAEs performed well with the corresponding relative root mean square error estimates (RMSQ) substantially smaller than for the design based estimates without an increase in the Rel Bias. The average interval width for the SAEs is just 58 percent of the average Horvitz-Thompson estimator confidence interval.

References

Folsom, R.E. and Judkins, D.R. (1997). Substance Abuse in States and Metropolitan Areas: Model Based Estimates from the 1991-1993 National Household Survey on Drug Abuse--Methodology Report. Prepared for the Substance Abuse and Mental Health Services Administration.

Malec, D., Sedransk, J., and Tompkins, L. (1993). "Bayesian Predictive Inference for Small Areas for Binary Variables in the National Health Interviews Survey," in *Case Studies in Bayesian Statistics*: eds. C. Gatsonis, J.S. Hodges, R.E. Kass, and N.D. Singpurualla, New York: Springer-Verlag, 377-389.

Figure 2. Relative Bias of Weighted and Unweighted SAE Aggregates

	Relative Bias (Weighted)					Relative Bias (Unweighted)				
	12-17	18-25	26-34	35+	Total	12-17	18-25	26-34	35+	Total
Total United States	-5.27%	-2.14%	-2.31%	-5.12%	-3.55%	-1.70%	7.49%	11.61%	37.60%	15.66%
North East Region	-10.83%	3.05%	-0.44%	8.65%	3.50%	-7.94%	15.21%	21.37%	84.65%	33.86%
South Region	-10.16%	-5.94%	-6.50%	-7.45%	-6.54%	-8.02%	-2.15%	5.69%	26.02%	7.87%
North Central	5.91%	2.35%	2.30%	-5.80%	0.13%	3.45%	19.87%	18.57%	34.95%	20.67%
West Region	-6.20%	-4.10%	-2.51%	-10.18%	-7.66%	5.86%	4.53%	6.43%	27.12%	9.55%
California	-3.50%	-3.53%	2.90%	-7.94%	-4.44%	2.76%	3.46%	8.20%	20.28%	8.00%

Figure 3. Relative Root Mean Squared Errors (RMSQ) and Relative Biases (Rel Bias) for California Bootstrap Sample*

		Design Based Estimate				SAE Based Estimate			
Age Group	State	P_d	L95	U95	Width	P_m	L95	U95	Width
Total	CA	7.08%	6.17%	8.11%	1.94%	7.00%	6.20%	7.88%	1.68%
Total	CA1	6.64%	4.91%	8.93%	4.02%	7.05%	5.74%	8.55%	2.81%
Total	CA2	6.92%	4.78%	9.91%	5.13%	6.58%	5.33%	8.03%	2.70%
Total	CA3	9.09%	6.41%	12.75%	6.34%	8.03%	6.53%	9.74%	3.22%
Total	CA4	7.09%	4.79%	10.38%	5.58%	6.97%	5.60%	8.54%	2.94%
Total	CA5	8.30%	6.25%	10.94%	4.69%	8.05%	6.51%	9.82%	3.31%
				Avg Width	5.15%			Avg Width	2.99%
	Rel RMSQ	15.17%				9.18%			
	Rel Bias	7.52%				3.65%			

*These RMSQ and Rel Bias calculations use the full sample Horvitz-Thompson prevalence estimate of 7.08% as the benchmark.