ELSEVIER

Journal of Statistical Planning and Inference 112 (2003) 89–98

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Estimation of small-area proportions using covariates and survey data

Michael D. Larsen*

Department of Statistics, The University of Chicago, 5734 S. University Avenue, 606 Chicago, IL 60637, USA

Abstract

Data gathered in surveys often are used to estimate characteristics for subsets of the survey population. If the sample from a subset is small, then a traditional design-based survey estimator may have unacceptably large variance. Small-area estimation reduces the variance of estimators by "borrowing strength" across subsets. Here we compare estimators based on two models, one that uses simple geographic clustering and demographic data and one that uses more elaborate covariate information that relates subsets to one another. Data are from a survey conducted by the Gallup Organization. The methods incorporate survey weight information and are appropriate for rare events. Empirical Bayes estimation techniques are used. Covariates for the second model are selected in a step-wise manner until addition of another covariate does not yield a decrease in an objective criterion.

MSC: 68D05; 62F10; 62F99

Keywords: Empirical Bayes; Jackknife; Mean squared error; Small-area estimation; Variable selection

1. Introduction

Data gathered in surveys often are used to estimate characteristics of subsets (also called subpopulations or domains) of the survey population. If the sample from a subset is small, then a traditional design-based survey estimator may have unacceptably high variability. The Gallup Organization has conducted telephone household surveys to study the prevalence of alcohol and drug use in various states. State administrators

* Tel.: +1-773-702-9095; fax: +1-773-702-9810.
E-mail address: larsen@galton.uchicago.edu (M.D. Larsen).

0378-3758/03/$ - see front matter
PII: S0378-3758(02)00325-7

would like to have estimates of prevalence within counties. However, sample sizes often are planned and budgets allocated for producing accurate estimates in larger administrative regions.

Small-area estimation methods can be used to reduce the variance of estimators by "borrowing strength" across subsets. Although simple synthetic estimators tend to produce very stable estimates, they do not necessarily reflect observed differences in subsets of the population, even when sample sizes are not so small. Some composite estimators combine the direct and synthetic estimators, but in an arguably arbitrary manner.

Here we compare estimators based on two models, one that uses simple geographic clustering and demographic data (the model of Chattopadhyay et al., 1999) and an extension that uses more elaborate county-level covariate information. The methods incorporate survey weight information and are appropriate when trying to estimate the rate of occurrence of a rare event. Empirical Bayes estimation techniques (see, e.g., Efron and Morris, 1973; Fay and Herriot, 1979; Ghosh and Lahiri, 1987) are used. A procedure similar to step-wise variable selection in linear regression and using an objective criterion is used to choose covariates in the second model.

Small-area estimation methods are reviewed in Ghosh and Rao (1994). Malec et al. (1999), extending Malec et al. (1997) to account for unequal selection probabilities, specify a Bayesian hierarchical model that includes a logistic regression model for expected proportions. Farrell et al. (1997) use empirical Bayes procedures to estimate logistic regression parameters related to expected small area proportions in their model. Other recent work in small area estimation includes development of methods for spatial and temporal modeling of counts (see, e.g., Wakefield and Elliott, 1999) and means of quantitative variables through nested error regression models (see, e.g., Singh et al., 1998).

Section 2 presents notation and some estimators of small-area proportions. Section 3 specifies two models and estimators. It also discusses estimation of parameters, mean square error, and, for the second model, variable selection. Section 4 describes the Gallup Organization survey and displays results using various estimators. Section 5 is a conclusion.

2. Notation and basic estimators

The notation here follows that of Chattopadhyay et al. (1999) and is designed based on a Gallup Organization telephone survey of one adult per sample household. The households are located in counties, which are grouped into planning regions in a state. Let n_i be the sample size in the ith planning region, $i=1,\ldots,I$ $(n=\sum_{i=1}^{I} n_i)$. Suppose there are J_i counties in the ith planning region $(i=1,\ldots,I)$. The samples in each region are post-stratified according to K demographic groups. There are n_{ijk} observations within the kth demographic group in the jth county belonging to the ith planning region $(i=1,\ldots,I;\ j=1,\ldots,J_i;\ k=1,\ldots,K)$. Let S_{ij} indicate the demographic groups from which individuals have completed surveys in the jth county within the ith region $(i=1,\ldots,I;\ j=1,\ldots,J_i)$. That is, $k\in S_{ij}$ if $n_{ijk}>0$.

The response (0 or 1) from the lth person from the kth demographic group in the jth county in the ith planning region ($i=1,\ldots,I;\ j=1,\ldots,J_i;\ k\in S_{ij};\ l=1,\ldots,n_{ijk}$) is denoted by y_{ijkl}. The survey sampling weight for this person is w_{ijkl}. A vector of covariates $x'_{ij}=(x_{ij1},\ldots,x_{ijp})$ is measured in county j. The proportion of the population from census estimates at the time of the survey in the kth demographic group is a_{ijk}. A parameter of interest is π_{ij}, the proportion of the jth county within the ith planning area ($i=1,\ldots,I;\ j=1,\ldots,J_i$) with a certain characteristic.

As in Chattopadhyay et al. (1999), the direct estimator of π_{ij} is

$$\widehat{\pi_{ij}}^{\mathrm{D}}=\frac{\sum_{k\in S_{ij}}\sum_{l=1}^{n_{ijk}}w_{ijkl}\,y_{ijkl}}{\sum_{k\in S_{ij}}\sum_{l=1}^{n_{ijk}}w_{ijkl}}=\frac{\sum_{k\in S_{ij}}(\sum_{l=1}^{n_{ijk}}w_{ijkl})\widehat{\pi_{ijk}}^{\mathrm{D}}}{\sum_{k\in S_{ij}}(\sum_{l=1}^{n_{ijk}}w_{ijkl})},$$

where $\widehat{\pi_{ijk}}^{\mathrm{D}}$ is the direct estimator of the prevalence in the kth demographic group in the jth county in the ith area. A synthetic estimator that implicitly assumes proportions for demographic groups are the same within counties in a planning region is $\widehat{\pi_{ij}}^{S}=\sum_{k\in S_{ij}}a_{ijk}\widehat{\pi_{ik}}^{\mathrm{D}}$, where $\widehat{\pi_{ik}}^{\mathrm{D}}$ is a direct survey estimator of π_{ik}, the prevalence in the kth demographic group in the ith planning region. A composite estimator is $\widehat{\pi_{ij}}^{\mathrm{C}}=\sum_{k\in S_{ij}}a_{ijk}\widehat{\pi_{ijk}}^{\mathrm{D}}+\sum_{k\notin S_{ij}}a_{ijk}\widehat{\pi_{ik}}^{\mathrm{D}}$ in which the prevalence in the kth demographic group in the jth county in the ith region is estimated directly ($\widehat{\pi_{ijk}}^{\mathrm{D}}$) if possible and by the region-wide estimate if necessary ($\widehat{\pi_{ik}}^{\mathrm{D}}$). This estimator does not necessarily produce estimates between $\widehat{\pi_{ij}}^{\mathrm{D}}$ and $\widehat{\pi_{ij}}^{\mathrm{S}}$. See Chattopadhyay et al. (1999) for discussion of these estimators.

3. Empirical Bayes estimators

The above estimators make strict choices about the use of direct versus pooled estimators and correspond to implicit models. In order to motivate estimators that make a more subtle compromise between alternatives, models are proposed below. The first model was originally presented in Chattopadhyay et al. (1999), and the second model is an extension that incorporates county-level covariate information.

3.1. Model 1 (Chattopadhyay et al., 1999)

A. Given the π_{ijk}'s, the y_{ijkl}'s are uncorrelated Bernoulli random variables with parameter π_{ijk} for $i=1,\ldots,I;j=1,\ldots,J_i;k=1,\ldots,K;l=1,\ldots,n_{ijk}$.
B. Marginally, the π_{ijk}'s are uncorrelated with $E(\pi_{ijk})=\mu_{ik}$; $\mathrm{Var}(\pi_{ijk})=d\mu_{ik}^2$ ($i=1,\ldots,I;j=1,\ldots,J_i;k=1,\ldots,K$).

In Model 1, the proportions in demographic groups can vary across counties in a region. In our application, we set $d=\frac{1}{3}$, which would be the value of d is the distribution of π_{ijk} were Uniform$(0,2\mu_{ik})$. The uniform prior distribution would imply that the proportions π_{ijk} are of similar small value in demographic group k across counties j within region i.

Chattopadhyay et al. (1999) empirical Bayes estimator $\widehat{\pi_{ij}}^{\mathrm{EB1}}$ of π_{ij} is $\sum_{k\in S_{ij}} \times a_{ijk}(\widehat{B_{ijk}}\widehat{\pi_{ijk}}^{\mathrm{D}} + (1-\widehat{B_{ijk}})\widehat{\mu_{ik}}) + \sum_{k\notin S_{ij}} a_{ijk}\widehat{\mu_{ik}}$, where $\widehat{B_{ijk}} = d\widehat{\mu_{ik}}^2/(d\widehat{\mu_{ik}}^2 + c_{ijk}(\widehat{\mu_{ik}} - (d+1)\widehat{\mu_{ik}}^2))$, $c_{ijk} = \sum_{l=1}^{n_{ijk}} w_{ijkl}^2/(\sum_{l=1}^{n_{ijk}} w_{ijkl})^2$, and $\widehat{\mu_{ik}} = \widehat{\pi_{ik}}^{\mathrm{D}}$. The mean squared error (MSE) of the empirical Bayes estimator $\widehat{\pi_{ij}}^{\mathrm{EB1}}$ is derived and a jackknife estimation strategy is developed and applied in Chattopadhyay et al. (1999).

Model 1 reflects a clustering of counties by region. The model and data guide the level of compromise between direct and synthetic estimators in each demographic group in each county in each region. Model 2 below uses additional covariate information to guide the compromise between direct and synthetic estimators.

3.2. Model 2

A. Conditional on π_{ijk}, y_{ijkl}'s are uncorrelated Bernoulli random variables with parameter π_{ijk} for $i = 1,\ldots,I; j = 1,\ldots,J_i; k = 1,\ldots,K; l = 1,\ldots,n_{ijk}$.
B. Marginally, π_{ijk}'s are uncorrelated with $E(\pi_{ijk}) = (\exp(\alpha_{ik} + x'_{ij}\beta))/(1+\exp(\alpha_{ik} + x'_{ij}\beta)) = \pi_{ijk}(\alpha_{ik},\beta) = \mu_{ijk}$, say, and $\mathrm{Var}(\pi_{ijk}) = d\mu_{ijk}^2$.

In Model 2, α_{ik} is the effect due to the kth demographic group in the ith planning area and $\beta' = (\beta_1,\ldots,\beta_p)$ is a vector of regression coefficients. In absence of any covariates, Model 2 reduces to Model 1.

According to Model 2, given π_{ijk}, the $\widehat{\pi_{ijk}}^{\mathrm{D}}$'s are mutually independent, $E(\widehat{\pi_{ijk}}^{\mathrm{D}}|\pi_{ijk}) = \pi_{ijk}$, and $\mathrm{Var}(\widehat{\pi_{ijk}}^{\mathrm{D}}|\pi_{ijk}) = c_{ijk}\pi_{ijk}(1-\pi_{ijk})$ for $i = 1,\ldots,I, j = 1,\ldots,J_i, k = 1,\ldots,K$. The linear Bayes (see, e.g., Hartigan, 1969) estimator of π_{ij}, under Model 2 and a squared error loss function, is $\widehat{\pi_{ij}^{\mathrm{B2}}} = \sum_{k\in S_{ij}} a_{ijk}(B_{ijk}\widehat{\pi_{ijk}}^{\mathrm{D}} + (1-B_{ijk})\mu_{ijk}) + \sum_{k\notin S_{ij}} a_{ijk}\mu_{ijk}$, where $B_{ijk} = d\mu_{ijk}^2/(d\mu_{ijk}^2 + c_{ijk}(\mu_{ijk} - (d+1)\mu_{ijk}^2))$. When $k \in S_{ij}$, the proportion π_{ijk} is estimated by a weighted combination of the direct survey estimator and an estimator that involves the covariate information through μ_{ijk}.

Since limited assumptions about the prior distribution of π_{ijk} and no distributional assumptions about β and α_{ik} were made, a criterion is proposed here for determining values to use in place of β and α_{ik}. Under Model 2, with respect to the unconditional distribution determined by both parts of the model, $E(\widehat{\pi_{ijk}}^{\mathrm{D}}) = \pi_{ijk}(\alpha_{ik},\beta) = \mu_{ijk}$ and $\mathrm{Var}(\widehat{\pi_{ijk}}^{\mathrm{D}}) = d\mu_{ijk}^2 + c_{ijk}(\mu_{ijk} - (d+1)\mu_{ijk}^2)$. The criterion

$$Q(\alpha_{ik},\beta) = \sum_{i,j} \frac{(\widehat{\pi_{ij}}^{\mathrm{D}} - \sum_k a_{ijk}\mu_{ijk})^2}{\sum_k a_{ijk}^2[d\mu_{ijk}^2 + c_{ijk}(\mu_{ijk} - (d+1)\mu_{ijk}^2)]}$$

is like a sum of standardized squared residuals. The values $\widehat{\alpha_{ik}}$ and $\hat{\beta}$ that minimize $Q(\alpha_{ik},\beta)$ with respect to α_{ik} and β will be used as estimates of α_{ik} and β. An empirical Bayes estimator of π_{ij} then is $\widehat{\pi_{ij}}^{\mathrm{EB2}} = \sum_{k\in S_{ij}} a_{ijk}(\widehat{B_{ijk}}\widehat{\pi_{ijk}}^{\mathrm{D}} + (1-\widehat{B_{ijk}})\widehat{\mu_{ijk}}) + \sum_{k\notin S_{ij}} a_{ijk}\widehat{\mu_{ijk}}$, where $\widehat{\mu_{ijk}} = \pi_{ijk}(\widehat{\alpha_{ik}},\hat{\beta})$ and $\widehat{B_{ijk}} = B_{ijk}(\widehat{\mu_{ijk}})$.

3.3. MSE of the estimator

The mean square error (MSE) of the estimator $\widehat{\pi_{ij}}^{\mathrm{B2}}$ can be derived following arguments used in Chattopadhyay et al. (1999). It can shown that with respect to Model 2 that $E(\widehat{\pi_{ij}}^{\mathrm{B2}}) = \sum_{k\in S_{ij}} a_{ijk}\mu_{ijk} + \sum_{k\notin S_{ij}} a_{ijk}\mu_{ijk} = E(\pi_{ij})$, where $\pi_{ij} = \sum_{k=1}^{K} a_{ijk}\pi_{ijk}$. Furthermore, $\mathrm{Var}(\widehat{\pi_{ij}}^{\mathrm{B2}}) = \sum_{k\in S_{ij}} a_{ijk}^2 B_{ijk} d\mu_{ijk}^2 = \mathrm{Cov}(\widehat{\pi_{ij}}^{\mathrm{B2}}, \pi_{ij})$. Thus, it follows that

$$\begin{aligned}
\mathrm{MSE}(\widehat{\pi_{ij}}^{\mathrm{B2}}) &= E(\widehat{\pi_{ij}}^{\mathrm{B2}} - \pi_{ij})^2 \\
&= \mathrm{Var}(\widehat{\pi_{ij}}^{\mathrm{B2}}) + \mathrm{Var}(\pi_{ij}) - 2\mathrm{Cov}(\widehat{\pi_{ij}}^{\mathrm{B2}}, \pi_{ij}) \\
&= \mathrm{Var}(\pi_{ij}) - \mathrm{Var}(\widehat{\pi_{ij}}^{\mathrm{B2}}) \\
&= d\left(\sum_{k\in S_{ij}} a_{ijk}^2(1 - B_{ijk})\mu_{ijk}^2 + \sum_{k\notin S_{ij}} a_{ijk}^2\mu_{ijk}^2\right).
\end{aligned}$$

If estimates of unknown parameters are plugged into the formula above, it will tend to underestimate $\mathrm{MSE}(\widehat{\pi_{ij}}^{\mathrm{EB2}})$. The uncertainty due to estimation of the unknown parameters (α_{ik} and β) must be taken into account (see, e.g., Prasad and Rao, 1990; and Lahiri and Rao, 1995). One way to include the uncertainty due to estimation of α_{ik} and β is through the formula $\mathrm{MSE}(\widehat{\pi_{ij}}^{\mathrm{EB2}}) = \mathrm{MSE}(\widehat{\pi_{ij}}^{\mathrm{B2}}) + E(\widehat{\pi_{ij}}^{\mathrm{EB2}} - \widehat{\pi_{ij}}^{\mathrm{B2}})^2$.

An estimator of $\mathrm{MSE}(\widehat{\pi_{ij}}^{\mathrm{EB2}})$ is given by

$$\mathrm{mse}(\widehat{\pi_{ij}}^{\mathrm{EB2}}) = \mathrm{mse}_J(\widehat{\pi_{ij}}^{\mathrm{B2}}) + E_J(\widehat{\pi_{ij}}^{\mathrm{EB2}} - \widehat{\pi_{ij}}^{\mathrm{B2}})^2, \tag{1}$$

where $\mathrm{mse}_J(\widehat{\pi_{ij}}^{\mathrm{B2}}) = d(\sum_{k\in S_{ij}} a_{ijk}^2(1 - \widehat{B_{ijk}})\widehat{\mu_{ijk}}^2 + \sum_{k\notin S_{ij}} a_{ijk}^2\widehat{\mu_{ijk}}^2) - (J_i - 1)/J_i \times \sum_{u=1}^{J_i}(\mathrm{mse}_{(-u)}(\widehat{\pi_{ij}}^{\mathrm{B2}}) - \mathrm{mse}(\widehat{\pi_{ij}}^{\mathrm{B2}}))$ and $\mathrm{E}_J(\widehat{\pi_{ij}}^{\mathrm{EB2}} - \widehat{\pi_{ij}}^{\mathrm{B2}})^2 = \frac{J_i-1}{J_i}\sum_{u=1}^{J_i}(\widehat{\pi_{ij(-u)}}^{\mathrm{EB2}} - \widehat{\pi_{ij}}^{\mathrm{EB2}})^2$. The calculation of the two terms, $\mathrm{mse}_J(\widehat{\pi_{ij}}^{\mathrm{B2}})$ and $\mathrm{E}_J(\widehat{\pi_{ij}}^{\mathrm{EB2}} - \widehat{\pi_{ij}}^{\mathrm{B2}})^2$, is accomplished by leaving out each county one-by-one and reestimating α_{ik} and β. In the above,

$$\widehat{\pi_{ij(-u)}}^{\mathrm{EB2}} = \sum_{k\in S_{ij}} a_{ijk}^2(\widehat{B_{ijk(-u)}}\widehat{\pi_{ijk}}^{\mathrm{D}} + (1 - \widehat{B_{ijk(-u)}})\widehat{\mu_{ijk(-u)}}) + \sum_{k\notin S_{ij}} a_{ijk}\widehat{\mu_{ijk(-u)}}$$

and

$$\mathrm{mse}_{(-u)}(\widehat{\pi_{ij}}^{\mathrm{B2}}) = d\left(\sum_{k\in S_{ij}} a_{ijk}^2(1 - \widehat{B_{ijk(-u)}})\widehat{\mu_{ijk(-u)}}^2 + \sum_{k\notin S_{ij}} a_{ijk}^2\widehat{\mu_{ijk(-u)}}^2\right),$$

where $\widehat{\mu_{ijk(-u)}}$ is obtained as is $\widehat{\mu_{ijk}}$ except that the uth county is deleted and $\widehat{B_{ijk(-u)}} = B_{ijk}(\widehat{\mu_{ijk(-u)}})$.

Shao and Tu (1995) provide a review of the Jackknife. Jiang et al. (1998) provide justification for these estimates of the MSE of $\widehat{\pi_{ij}}^{\mathrm{EB2}}$.

3.4. Selection of covariates for Model 2

Typically there are many covariates available on the counties. For models with a given number of covariates, here we choose the set of covariates that produce the lowest value of the objective criterion $Q(\alpha_{ik}, \beta)$. In order to reduce the number of models that have to be fit, covariates are added to the model in a manner analogous to step-wise variable selection in linear regression until the value of the Q function stops decreasing appreciably.

4. An Example

4.1. Gallup survey and covariates

The data analyzed here come from a Gallup Organization household survey on the use of alcohol and illegal drugs among civilian, non-institutionalized adults in a particular state. Based on assessments of use and dependence levels, the state can project treatment needs. Sample sizes were chosen to achieve a desired level of accuracy in estimating prevalence at the state level and then divided among the planning regions. Independent random digit dialing samples using Casady and Lepkowski's (1993) truncated method were generated for each planning region, as described in Chattopadhyay et al. (1999).

In order to allocate the sample to the regions, optimal sample sizes were computed based on an index formed from drug treatment admission rates in the counties in the planning regions. Additional sample was allocated in each region to the 18–45 age group. Sampling weights w_{ijkl} were calculated to compensate in estimation for disproportional sampling fractions relative to current census population estimates.

The county-level covariates come from the U.S. Census and the Kids Count 1993 survey from the Center for the Study of Social Policy (1993). There are dozens of variables that could be used in the prediction equations. We illustrate our method by considering nine auxiliary variables that record various rates. We do not claim subject area expertise in our selections. The variables are unemployment rate (UNE), percentage of housing units that are vacant (VAC), percentage of population 18 years of age or younger (YOUNG), percentage of population classified as minority (MIN), percentage of children under 18 living in poverty as defined by the U.S. Bureau of the Census (POV), percentage of children in families headed by a person without a spouse in the home (SINGLE), percentage of youths 18 years of age or younger with status offenses, misdemeanors, and felonies (CRIM), percentage of families with related children who are AFDC recipients (AFDC), and percentage of births with prenatal care in first three months of pregnancy (PRENAT).

The main outcome y_{ijkl} considered here is whether or not a respondent is dependent on alcohol as defined by the National Technical Center's DSM-III-R criteria. People dependent on or abusing alcohol according to the criteria may be eligible for treatment. There are eight demographic groups used to stratify the population: a two-by-four cross of sex (Female, Male) with age (18–24, 25–44, 45–64, 65+).

Table 1
Objective criterion $Q(\alpha_{ik}, \beta)$ for models 1 and 2 with single covariates

Covariates	$Q(\alpha_{ik}, \beta)$
POV	22.7
AFDC	23.4
UNE	26.0
SINGLE	26.2
YOUNG	27.3
VAC	27.5
PRENAT	27.8
CRIM	27.9
MIN	28.3
None[a]	56.5

[a]No covariates corresponds to Model 1.

4.2. Results

The function Q is minimized in this example using the method of steepest descent. Multiple starting values lead to the same values of Q, which were checked to be minimum values by perturbing parameter values and recomputing Q. The variable reporting the percentage of children under 18 living in poverty as defined by the U.S. Bureau of the Census (POV) had the lowest value of Q: 22.74. Variables were standardized to have mean 0 and variance 1. The Q criterion evaluated on Model 1 was much higher. In Model 1, the value of μ_{ijk} is estimated to be $\widehat{\mu_{ik}} = \widehat{\pi_{ik}}$ in correspondence with model assumptions. The coefficient for the variable POV was negative (-0.50), which means that, other things being equal, the predicted percent classified as dependent on alcohol in a demographic group in counties in a planning region is reduced in counties with higher poverty rates relative to what the prediction would have been otherwise. This variable has correlation in counties with positive direct estimates with direct county-level estimates of alcohol dependence rates of -0.38 on the linear scale and -0.41 on the logistic scale. Table 1 presents results.

When a second covariate is added to variable POV in the second model, the Q criterion decreases, but not by much. The lowest value (21.9) is achieved by combining POV and SINGLE, the percentage of children in families headed by a person without a spouse in the home. The coefficients on POV and SINGLE are -0.44 and -0.16, respectively. Given the complexity of the model, for a demographic group in a county the negative coefficient on SINGLE does not mean simply that predictions of alcohol dependence decrease as the percentage of children in singly headed households increases.

Adding a third variable to these two did not decrease Q very much.

Table 2 displays estimates and square root of estimated mean square errors for the empirical Bayes estimator and estimated standard error for the direct survey estimator for 40 counties from the state. Direct survey estimates ($\widehat{\pi_{ij}}^{\mathrm{D}}$) are quite variable. The empirical Bayes estimator ($\widehat{\pi_{ij}}^{\mathrm{EB1}}$) based on the first model exhibits less variability. It

Table 2
Estimates and error estimates for 40 counties

	Direct		Model 1		Model 2 One predictor		Model 2 Two predictors		Observed groups	Sample size
	$\widehat{\pi_{ij}}^{D}$	(Est.se)	$\widehat{\pi_{ij}}^{EB1}$	$(\sqrt{MSE})$	$\widehat{\pi_{ij}}^{EB2}$	$(\sqrt{MSE})$	$\widehat{\pi_{ij}}^{EB2}$	$(\sqrt{MSE})$		
1	1.7	2.9	1.6	0.7	1.6	0.5	1.5	0.5	8	30
2	4.4	2.4	2.1	1.0	4.8	1.5	4.0	1.3	8	111
3	0.0	0.0	3.0	1.1	2.3	1.2	2.6	1.3	8	36
4	0.0	0.0	5.3	2.1	3.0	1.2	3.3	1.1	5	6
5	9.4	5.3	6.9	2.3	6.3	2.4	6.7	2.3	8	37
6	1.6	1.3	2.7	0.9	3.3	1.5	2.8	1.3	8	136
7	9.3	7.3	3.1	1.1	2.1	1.1	2.0	1.1	6	25
8	0.0	0.0	3.1	1.1	1.2	0.8	1.4	0.9	7	20
9	0.0	0.0	5.8	2.3	3.9	1.5	4.7	1.6	3	3
10	1.5	1.2	1.9	0.7	2.5	0.9	2.3	0.9	8	81
11	0.0	0.0	1.5	0.7	3.0	1.0	2.5	0.9	8	58
12	7.0	6.3	1.8	0.8	3.5	1.1	3.7	1.2	6	14
13	5.7	5.7	6.4	2.3	5.7	2.2	5.2	1.8	8	37
14	0.0	0.0	1.6	0.8	4.9	1.6	4.8	1.7	4	12
15	2.4	1.6	4.4	1.7	3.5	1.4	3.2	1.2	8	120
16	4.1	4.1	3.0	1.0	3.8	1.6	4.3	1.7	7	32
17	2.8	2.8	1.8	0.9	4.6	1.5	4.3	1.6	8	48
18	3.9	1.3	3.2	0.8	2.9	1.3	2.5	1.1	8	316
19	0.0	0.0	5.7	2.3	3.0	1.2	2.8	0.9	5	19
20	3.1	3.1	3.2	1.1	1.5	0.9	2.0	1.1	6	20
21	2.7	1.8	5.8	1.8	3.7	1.3	3.5	1.1	8	102
22	4.2	2.2	2.2	0.7	3.1	1.1	3.0	1.1	8	124
23	9.7	3.4	8.8	2.4	7.7	2.4	7.5	2.1	8	121
24	0.0	0.0	1.9	0.8	4.3	1.5	5.2	1.8	6	22
25	7.8	5.0	1.8	0.7	5.7	1.9	6.0	2.1	6	32
26	0.0	0.0	1.6	0.8	1.9	0.6	1.9	0.6	7	28
27	2.2	2.2	4.9	2.0	4.5	1.9	3.3	1.2	8	63
28	10.5	10.5	1.7	0.8	4.6	1.5	5.4	1.8	5	5
29	0.0	0.0	3.0	1.1	4.1	1.8	4.2	1.8	5	12
30	0.0	0.0	1.5	0.7	2.6	0.9	2.7	0.9	6	11
31	4.6	3.3	5.8	2.2	5.1	1.8	5.9	1.9	8	44
32	8.4	4.5	4.1	1.2	7.0	2.1	7.7	2.1	8	52
33	2.5	1.5	2.1	0.7	2.9	1.1	2.9	1.1	8	144
34	2.9	2.8	1.7	0.8	2.3	0.7	2.2	0.8	7	49
35	0.0	0.0	2.8	1.0	2.2	1.2	2.4	1.2	8	22
36	0.0	0.0	2.9	1.0	1.9	1.1	2.2	1.1	6	17
37	4.2	4.7	2.1	0.8	2.7	1.0	2.8	1.0	6	26
38	0.0	0.0	5.7	2.4	3.3	1.3	3.4	1.2	6	16
39	0.0	0.0	3.0	1.1	4.0	1.8	4.7	1.8	6	10
40	5.3	2.1	3.5	0.9	3.3	1.6	3.2	1.5	8	144

is arguable that the nonzero estimates for $\widehat{\pi_{ij}}^{EB1}$ are more believable than the estimated zeros for $\widehat{\pi_{ij}}^{D}$. Estimates using the synthetic ($\widehat{\pi_{ij}}^{S}$) and composite ($\widehat{\pi_{ij}}^{C}$) estimators are presented and their problems are discussed in Chattopadhyay et al. (1999).

The empirical Bayes estimator ($\widehat{\pi_{ij}}^{EB2}$) based on model two produces estimates that are different from those produced by the other estimators. The adjustment reflects the use of the covariate POV in the model. The estimated MSE's for the Model 2 estimates, however, are not lower than those for Model 1. The average MSE value is about the same for the two models. The source of the increase in the MSE's over Model 1 is the larger changes in the jackknife estimates of π_{ij}: $\widehat{\pi_{ij(-u)}}^{EB2}$. As some counties are removed from the calculation of α_{ik} and β, parameter estimates ($\widehat{\alpha_{ik}}$, $\widehat{\beta}$) change enough to cause large changes in the estimated proportions. Minimizing the Q criterion is not equivalent to minimizing MSE. The changes in proportions increase the value of the second term in (1).

When the second covariate SINGLE is added to Model 2, the proportions change a little bit, but not much. The small change in estimates is consistent with the small decrease in the Q criterion. When SINGLE is included, estimated mean square errors decrease on average, but not by much, from what they were with one covariate. SINGLE apparently is not contributing much additional information about dependence rates.

5. Conclusion

Small-area estimation methods can be useful when estimates are desired for more than one small area, sample size is inadequate in some areas for direct estimation, and estimators that "borrow strength" across areas are acceptable. The empirical Bayes estimators presented here are based on models that have few distributional assumptions and incorporate survey weights. They compromise between synthetic estimators that tend to be very stable, but unresponsive to large sample sizes in some areas, and direct estimators that are quite variable with small sample sizes. The level of compromise in each county in the second model is determined in part through the influence of covariate information.

The example from the Gallup organization was used to illustrate the methods. Although adding covariates did not lead to large improvements in estimated mean square error in this example, the methods presented here do provide a procedure for including covariates and for assessing their contribution. Perhaps the factors of age and sex, which were included in defining demographic groups, and the clustering of counties into planning regions were much more relevant for the outcome of alcohol dependence than the covariates considered here.

Further study will include comparison of empirical Bayesian methods with hierarchical Bayesian methods that specify distributions for parameters α_{ik} and β. Additionally, the influence of the value of d will be studied.

Acknowledgements

The author would like to thank Partha Lahiri of the University of Nebraska for his significant contribution to this work, to Manas Chattopadhyay and John Reimnitz of the

Gallup Organization for their contributions to this work, and to the Gallup Organization for the data.

References

Casady, R.J., Lepkowski, J.M., 1993. Stratified telephone survey designs. Survey Methodol. 19, 103–113.

Center for the Study of Social Policy, 1993. Kids Count Data Book. Washington, D.C.

Chattopadhyay, M., Lahiri, P., Larsen, M., Reimnitz, J., 1999. Composite estimation of drug prevalences for sub-state areas. Survey Methodol. 25, 81–86.

Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors—an empirical Bayes approach. J. Amer. Statist. Assoc. 68, 117–130.

Farrell, P.J., MacGibbon, B., Tomberlin, T.J., 1997. Empirical Bayes estimators of small area proportions in multistage designs. Statist. Sinica 7, 1065–1083.

Fay, R., Herriot, R., 1979. Estimates of income for small places: an application of James–Stein procedures to census data. J. Amer. Statist. Assoc. 74, 269–277.

Ghosh, M., Lahiri, P., 1987. Robust empirical Bayes estimation of means from stratified samples. J. Amer. Statist. Assoc. 82, 1153–1162.

Ghosh, M., Rao, J.N.K., 1994. Small area estimation: an appraisal. Statist. Sci. 9, 55–93.

Hartigan, J.A., 1969. Linear Bayesian methods. J. Roy. Statist. Soc. Ser. B 31, 446–454.

Jiang, J., Lahiri, P., Wan, S., 1998. Jackknifing mean squared error of empirical best predictor. unpublished manuscript.

Lahiri, P., Rao, J.N.K., 1995. Robust estimation Of mean squared error Of small area estimators. J. Amer. Statist. Assoc. 90, 758–766.

Malec, D., Davis, W.W., Cao, X., 1999. Model-based small area estimation of overweight prevalence using sample selection adjustment. Statist. Med. 18, 3189–3200.

Malec, D., Sedransk, J., Moriarity, C.L., Leclere, F.B., 1997. Small area inference for binary variables in the National Health Interview Survey. J. Amer. Statist. Assoc. 92, 815–826.

Prasad, N.G.N., Rao, J.N.K., 1990. The estimation of mean squared errors of small area estimators. J. Amer. Statist. Assoc. 85, 163–171.

Shao, J., Tu, D., 1995. The Jackknife and the Bootstrap. Springer, Berlin.

Singh, A.C., Stukel, D.M., Pfeffermann, D., 1998. Bayesian versus frequentist measures of error in small area estimation. J. Roy. Statist. Soc. Ser. B 60, 377–396.

Wakefield, J., Elliott, P., 1999. Issues in the statistical analysis of small area health data. Statist. Med. 18, 2377–2399.