Small area prevalence estimates using two surveys

William W. Davis[1], Charuta Soman[2], Zhaohui Zou[3]

National Cancer Institute, Bethesda, MD[1]
IMS Health, Philadelphia, PA[2]
Information Management Services, Silver Spring, MD[3]

Abstract

We develop a model based approach to estimate small area (county) prevalence using information from two surveys assumed to be an in person area survey and a telephone survey. Here, we demonstrate the ability to estimate parameters generated from the assumed model using a Monte-Carlo EM algorithm.

The proposed estimation method is investigated using simulated data. We assess the accuracy of the parameter estimates and also the accuracy of the county, state, and national estimates. The estimation accuracy of slope and intercept parameters of the model is good while some of the elements of the covariance matrix estimates are biased. The small area estimates are approximately unbiased. However, the coverage of the true values is less than the nominal value – most likely caused by the inaccurate covariance estimation. The model allows a test of whether non-response bias is present in the telephone survey estimates.

Keywords: non-response, non-coverage, Metropolis algorithm, generalized linear mixed model

1. Introduction

While telephone surveys can obtain large samples at a small cost, the direct estimates could suffer from non-response and/or non-coverage bias. Groves and Couper (1998) and Goyder *et al.* (2002) showed that an individual's socio-economic status (SES) is related to survey response propensity. Van Goor and Rispen (2005) studied the impact of the "middle class bias" (too few low and high SES individuals in the survey) of Dutch telephone surveys. The results of an additional (in-person survey) could be used to reduce or eliminate these potential telephone survey biases using a statistical model. We provide a model-based approach to small area estimation (SAE), where the small areas are counties, using two surveys. For an excellent summary of SAE methodology using a single survey see Rao (2003).

This work is closely related to Elliott and Davis (2005) and Raghunathan *et al.* (2007) which developed techniques for SAE of prevalence using two surveys. Our work is similar to Raghunathan's, with the following major differences:

- We use a binomial model for the sampling distributions of the individual sources.
- We assume an additional data source; namely, an estimate of the telephone response rate at the small area level.
- We use a sampling rather than a Bayesian estimation approach.

Now we define in more detail the data sources obtained from the two surveys. We assume that the in-person survey determines whether the respondent has a working telephone in the household. Then, for the in-person survey, for telephone owners in county *i*, we define $x_i, i = 1, 2, ..., a$ as the direct estimate (using the statistical weights and sample design) of the true population prevalence, θ_i, where *a* is the number of sampled areas of the "A" total areas (counties). We let $v_{x,i}$ denote the sampling variance of x_i based on the sample size, $n_{x,i}$ and let the effective sample size be $n'_{x,i} = x_i(1-x_i)/v_{x,i}$. Similarly, we let y_i, $n_{y,i}$, and $n'_{y,i}$ denote the prevalence estimate, the sample size, and the effective sample size, respectively, based on the in-person survey for those who reported not owning a telephone in county *i*.

For the telephone survey, we let z_i, $n_{z,i}$, and $n'_{z,i}$ denote the prevalence estimate, sample size, and effective sample size respectively. Table 1 summarizes the notation developed for the SAE of prevalence giving the direct estimate, the sample size and the effective sample size for the three data sources obtained from the two surveys.

Table 1. Notation for Survey Estimates in small area i

	Direct Estimate	Sample Size	Effective sample size
In-person: telephone	x_i	$n_{x,i}$	$n'_{x,i}$
In-person: non-telephone	y_i	$n_{y,i}$	$n'_{y,i}$
Telephone survey	z_i	$n_{z,i}$	$n'_{z,i}$

To include the telephone survey non-response, we assume that the number of completed calls, w_i, and the estimated number of eligible reporting units in the sample, $n'_{w,i}$, are available for all small areas.

In section 2 we provide a statistical model based on the sources from the two surveys. In section 3 we describe the model-based estimation methodology while in section 4 we show the estimation results for the parameters and geographic areas of interest.

2. Statistical Model

2.1 Telephone response assumptions

We assume that within each small area the adult telephone owners can be stratified into those who will respond and those who will not respond to the telephone survey (e.g., Cochran, 1977, Chapter 13). The true proportion of those who will respond in area *i* is labelled ρ_i. For the binary outcome, the true proportions may differ within the two strata; the proportions in the responding and non-responding strata are labelled as $\theta_{i,r}$ and $\theta_{i,n}$ respectively. Then, summing over the two strata, the true proportion, for the binary outcome is

$$\theta_i = \rho_i \theta_{i,r} + (1 - \rho_i)\theta_{i,n} \qquad (1)$$

The difference between the proportions in the responding strata from the entire population can be written

$$\theta_i - \theta_{i,r} = (1 - \rho_i)\left(\theta_{i,n} - \theta_{i,r}\right) \qquad (2)$$

Equation (2) exhibits the well-known fact that the non-response estimation bias is proportional to the proportion of non-responders and also to the difference in proportions on the binary outcome between the non-responding and responding strata. Table 2 provides a summary of the notation (showing the population proportion and outcome population proportion for the two strata and the entire population) for the households with telephones using estimates from the two surveys where θ_i satisfies (1).

Table 2. Notation for telephone households in county *i*

Strata	Pop. Prop.	Outcome Pop. Prop.
Responders	ρ_i	$\theta_{i,r}$
Non-responders	$1-\rho_i$	$\theta_{i,n}$
All	*1*	θ_i

2.2 Sampling distribution assumptions

The observed data is labelled $\underline{v} = \{\underline{v}_i\}$, where $\underline{v}_i = (w_i, x_i, y_i, z_i)$. We assume that the distributions of the components of $\underline{v}_i$ are binomial (*bin*) with the following parameters (defined in tables 1 and 2):

$$w_i \sim bin(n'_{w,i}, \rho_i) \qquad (3a)$$

$$n'_{x,i}x_i \sim bin(n'_{x,i}, \theta_i) \qquad (3b)$$

$$n'_{y,i}y_i \sim bin(n'_{y,i}, \phi_i) \qquad (3c)$$

$$n'_{z,i}z_i \sim bin(n'_{z,i}, \theta_{i,r}) \qquad (3d)$$

where ϕ_i is the proportion of the binary outcome for those not owning telephones. We define P_i as the proportion of individuals owning a telephone in area *i*. In small area *i*, the main parameter of interest – the proportion having the binary characteristic, λ_i, is

$$\lambda_i = P_i\theta_i + (1-P_i)\phi_i \qquad (4)$$

We assume logistic models for the 4 binomial parameters of (3a)-(3d) with random effects to account for correlation within small areas. We assume a model

$$\underline{\psi}_i = \underline{\alpha}_i + Z_i\underline{\beta} \qquad (5)$$

where $\underline{\alpha}_i^T = (\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \alpha_{i4})$, $\underline{\beta}^T = \left(\underline{\beta}_\rho, \underline{\beta}_{\theta,r}, \underline{\beta}_\phi, \underline{\beta}_{\theta,n}\right)$, $\underline{\psi}_i^T = \left(L\rho_i, L\theta_{ir}, L\phi_i, L\theta_{in}\right)$, where "L" is the *logit* operator defined by $L\rho_i = \text{logit}(\rho_i) = \ln(\rho_i/(1-\rho_i))$ with similar definitions for $L\theta_{ir}, L\phi_i,$ and $L\theta_{in}$, and where *T* denotes transpose . Also we define $Z_i^T = \underline{U}_i^T \otimes I_4$, where $\otimes$ denotes the Kronecker product, and $\underline{U}_i$ is a q-dimensional vector of county covariates. We assume that $\underline{\alpha} = \{\underline{\alpha}_i\}$ are independently distributed Gaussian random vectors, $\underline{\alpha}_i \sim N(\underline{\mu}, \Sigma)$ with mean vector $\underline{\mu}^T = (\mu_1, \mu_2, \mu_3, \mu_4)$ and covariance matrix $\Sigma = (\sigma_{ij})$.

2.3 Likelihood

With $\underline{\Theta} = \left(\underline{\beta}, \underline{\mu}, \Sigma\right)$ the density $p(\underline{v}, \underline{\alpha} \mid \underline{\Theta})$, of the data, $\underline{v}$, and the random effects, $\underline{\alpha}$, is given by

$$p(\underline{v}, \underline{\alpha} \mid \underline{\Theta}) = p(\underline{v} \mid \underline{\alpha}, \underline{\beta})\, p\left(\underline{\alpha} \mid \underline{\mu}, \underline{\Sigma}\right)$$
$$= \prod_i p\left(\underline{v}_i \mid \tilde{\underline{\alpha}}_i\right) p(\underline{\alpha}_i \mid \underline{\mu}, \Sigma) \qquad (6)$$

where $\tilde{\underline{\alpha}}_i = \left(\underline{\alpha}_i, \underline{\beta}\right)$ and *p(.|.)* denotes a generic probability density that is defined by its arguments. By the independence assumption,

$$p(\underline{v}_i|\tilde{\underline{\alpha}}_i) = p(w_i|\tilde{\underline{\alpha}}_i)\, p(y_i|\tilde{\underline{\alpha}}_i)\, p(z_i|\tilde{\underline{\alpha}}_i)\, p(x_i|\tilde{\underline{\alpha}}_i) \qquad (7)$$

In (7), the density of w_i, y_i, and z_i are obtained from the standard logistic regression model; for example,

$$\ln\left(p(w_i|\tilde{\underline{\alpha}}_i)\right) = c + w_i\left(\alpha_{i1} + \underline{\beta}_{\rho}^{T}\underline{U}_i\right) - n'_{wi}\ln\left(1+e^{\underline{\beta}_{\rho}^{T}\underline{U}_i+\alpha_{i1}}\right)$$

while the density of x_i is obtained from

$$p(x_i|\tilde{\underline{\alpha}}_i) \propto \exp\{n'_{x,i}\left(x_i \operatorname{logit}(\theta_i) + \ln(1-\theta_i)\right)\} \qquad (8)$$

where θ_i can be expressed in terms of $\tilde{\underline{\alpha}}_i$ through (1) and (5). The model specified in equations (6)-(8) is not a generalized linear mixed model (GLMM) - due to the nonlinear dependence of θ_i on the parameters of (1).

3. Estimation

We apply an extension of the EM algorithm (Dempster *et al.*, 1977), the Monte-Carlo Newton Raphson algorithm (McCullogh, 1997; Booth and Hobert, 1999). The general idea is to treat the random effects as missing data; the advantage of this approach is that the components of $\underline{v}$ are independent given the random effects.

We use a Metropolis algorithm to produce random draws from the conditional distribution of $\underline{\alpha} \mid \underline{v}$ and use the Monte Carlo approximation to the required expectations of the EM algorithm. We choose the candidate distribution as $p(\underline{\alpha} \mid \underline{\mu}, \Sigma)$ so the acceptance function has a simple form (McCullogh, 1997). If $\underline{\alpha} = \left(\underline{\alpha}_1, \ldots, \underline{\alpha}_A\right)$ denotes an ordering of previous draw from the distribution of $\underline{\alpha} \mid \underline{v}$, we generate a new value for $\underline{\alpha}_i^*$ using the candidate distribution, $\underline{\alpha}_i^* \sim N(\underline{\mu}, \Sigma)$. If we denote $\underline{\alpha}^* = \left(\underline{\alpha}_1, .., \underline{\alpha}_{R-1}, \underline{\alpha}_i^*, \underline{\alpha}_{R+1}, ..., \underline{\alpha}_A\right)$, then we accept $\underline{\alpha}^*$ as the new value with probability

$$A_i\left(\underline{\alpha}, \underline{\alpha}^*\right) = \min\{1, \frac{p\left(\underline{v} \mid \underline{\alpha}^*, \underline{\beta}\right)}{p\left(\underline{v} \mid \underline{\alpha}, \underline{\beta}\right)}\} \qquad (9)$$

otherwise, we retain the previous value $\underline{\alpha}$. Equation (9) involves only the general linear model portion of the model; in fact,

$$A_i\left(\underline{\alpha}, \underline{\alpha}^*\right) = \min\{1, \frac{p\left(\underline{v}_i \mid \underline{\alpha}_i^*, \underline{\beta}\right)}{p\left(\underline{v}_i \mid \underline{\alpha}_i, \underline{\beta}\right)}\} \qquad (10)$$

where the ratio of the densities in (10) can be computed using (6)-(8).

Adding the Metropolis step into the EM algorithm gives a Monte Carlo EM algorithm as follows:

1. Set *m=0* and choose starting values for the parameters $\underline{\Theta}^{(0)} = \left(\underline{\beta}^{(0)}, \underline{\mu}^{(0)}, \Sigma^{(0)}\right)$
2. Generate N values, $\underline{\alpha}^{(1)}, \ldots, \underline{\alpha}^{(N)}$ from $p\left(\underline{\alpha} \mid \underline{v}, \underline{\Theta}^{(m)}\right)$ using the Metropolis algorithm. Then choose
   a. $\underline{\beta}^{(m+1)}$ as $\max\left(N^{-1}\sum_{k=1}^{N}\ln\left(p\left(\underline{v} \mid \underline{\alpha}^{(k)}, \underline{\beta}\right)\right)\right)$
   b. $\underline{\mu}^{(m+1)} = \overline{\underline{\alpha}} = (AN)^{-1}\sum_{i=1}^{A}\sum_{k=1}^{N}\underline{\alpha}_i^{(k)}$
   c. $\Sigma^{(m+1)} = (AN)^{-1}\sum_{i=1}^{A}\sum_{k=1}^{N}\left(\underline{\alpha}_i^{(k)} - \overline{\underline{\alpha}}\right)\left(\underline{\alpha}_i^{(k)} - \overline{\underline{\alpha}}\right)^T$
   d. Set *m=m+1*.
3. If convergence is achieved, declare $\hat{\underline{\Theta}} = \underline{\Theta}^{(m+1)}$ as the maximum likelihood estimate (mle); otherwise return to step 2.

3.1 Slope estimation

To carry out the maximization with respect to $\underline{\beta}$ in 2(a) above, we use the Taylor expansion

$$\frac{\partial}{\partial\underline{\beta}}\sum_{k=1}^{N}\ln\left(p(\underline{v}|\underline{\alpha}^{(k)}, \underline{\beta})\right) \cong \frac{\partial}{\partial\underline{\beta}}\sum_{k=1}^{N}\ln\left(p(\underline{v}|\underline{\alpha}^{(k)}, \underline{\beta})\right)\Big|_{\underline{\beta}=\underline{\beta}_0} + \frac{\partial^2}{\partial\underline{\beta}\partial\underline{\beta}^T}\sum_{k=1}^{N}\ln\left(p(\underline{v}|\underline{\alpha}^{(k)}, \underline{\beta})\right)\Big|_{\underline{\beta}=\underline{\beta}_0}\left(\underline{\beta} - \underline{\beta}_0\right) \qquad (11)$$

Equation (11) leads to the Newton Raphson recursion

$$\underline{\beta}_{(j+1)} = \underline{\beta}_{(j)} + H_{(j)}^{-1}\underline{c}_{(j)} \qquad (12)$$

where

$$H_{(j)} = -\frac{1}{N}\frac{\partial^2}{\partial\underline{\beta}\partial\underline{\beta}^T}\sum_{k=1}^{N}\ln\left(p(\underline{v}|\underline{\alpha}^{(k)}, \underline{\beta})\right)\Big|_{\underline{\beta}=\underline{\beta}_j} \qquad (13)$$

$$\underline{c}_{(j)} = \frac{1}{N}\frac{\partial}{\partial\underline{\beta}}\sum_{k=1}^{N}\ln\left(p(\underline{v}|\underline{\alpha}^{(k)}, \underline{\beta})\right)\Big|_{\underline{\beta}=\underline{\beta}_j} \qquad (14)$$

The algorithm to determine $\underline{\beta}^{(m+1)}$ for step 2(a) above can be stated as follows:

A1. Set *j=0* and $\underline{\beta}_{(0)} = \underline{\beta}^{(m)}$

A2. Calculate $H_{(j)}$, $\underline{c}_{(j)}$, and $\underline{\beta}_{(j+1)}$ from (12)-(14).

A3. If convergence is obtained, declare the limiting value as $\underline{\beta}^{(m+1)}$ and continue to step 2(ii) above. Otherwise set *j=j+1* and return to step A2.

Now, we provide expressions for $\underline{c}_{(j)}$ and $H_{(j)}$, where the parameter $\left(\rho_i^{(k)}, \theta_{i,r}^{(k)}, \phi_i^{(k)}, \theta_{i,n}^{(k)}\right)$ is computed from (5) with $\left(\underline{\alpha}, \underline{\beta}\right) = (\underline{\alpha}^{(k)}, \underline{\beta}_j)$. Now,

$$\underline{c}_{(j)} = N^{-1} \sum_i \left(\sum_{k=1}^{N} W_i(\underline{\alpha}^{(k)}, \underline{\beta}_j) \tilde{\underline{w}}_i^{(k)} \right) \otimes \underline{U}_i \qquad (15)$$

where

$$W_i(\underline{\alpha}^{(k)}, \underline{\beta}) = K + L_i^{(k)} \qquad (16)$$

with $K=Diag(1,1,1,0)$, $L_i^{(k)} = \left(0_{4x3} \;\; \tilde{\underline{h}}_i^{(k)}\right)$ where 0_{4x3} is a 4x3 matrix with all entries equal to zero, $\tilde{\underline{h}}_i^{(k)} = \left(\theta_i^{(k)}(1-\theta_i^{(k)})\right)^{-1} \underline{h}_i^{(k)}$ with

$$\underline{h}_i^{(k)T} = [\rho_i^{(k)}\left(1-\rho_i^{(k)}\right)\left(\theta_{i,r}^{(k)} - \theta_{i,n}^{(k)}\right), \rho_i^{(k)}\theta_{i,r}^{(k)}\left(1-\theta_{i,r}^{(k)}\right),$$
$$0, \left(1-\rho_i^{(k)}\right)\theta_{i,n}^{(k)}\left(1-\theta_{i,n}^{(k)}\right)]$$

and the residual vector is defined by

$$\tilde{\underline{\omega}}_i^{(k)T} = \left(\tilde{w}_i^{(k)}, \tilde{z}_i^{(k)}, \tilde{y}_i^{(k)}, \tilde{x}_i^{(k)}\right) = \qquad (17)$$
$$\left(w_i - n_{w,i}^{'}\rho_i^{(k)}, n_{z,i}^{'}\left(z_i - \theta_{i,r}^{(k)}\right), n_{y,i}^{'}\left(y_i - \phi_i^{(k)}\right), n_{x,i}^{'}\left(x_i - \theta_i^{(k)}\right)\right)$$

Also, we have

$$H_{(j)} = A^{-1}\sum_i \Gamma_i^{(+)} \otimes \underline{U}_i \underline{U}_i^T \qquad (18)$$

where

$$\Gamma_i^{(+)} = \sum_{k=1}^{N} \Gamma_i^{(k)}$$

where the 4x4 matrices, $\Gamma_i^{(k)}$, are defined by

$$\Gamma_i^{(k)} = D_i^{(k)} + F_i^{(k)} - \tilde{G}_i^{(k)} \qquad (19)$$

where

$$D_i^{(k)} = Diag\left(n_{w,i}^{'}\rho_i^{(k)}\left(1-\rho_i^{(k)}\right), n_{z,i}^{'}\theta_{i,r}^{(k)}\left(1-\theta_{i,r}^{(k)}\right), n_{y,i}^{'}\phi_i^{(k)}\left(1-\phi_i^{(k)}\right), 0\right)$$

$$F_i^{(k)} = \left[n_{x,i}^{'}\theta_i^{(k)}\left(1-\theta_i^{(k)}\right) + \tilde{x}_i^{(k)}\left(1-2\theta_i^{(k)}\right)\right]\tilde{\underline{h}}_i^{(k)}\tilde{\underline{h}}_i^{(k)T},$$

$$\tilde{G}_i^{(k)} = \tilde{x}_i^{(k)}\left(\theta_i^{(k)}\left(1-\theta_i^{(k)}\right)\right)^{-1} G_i^{(k)}$$

where $G_i^{(k)} = \left(g_i^{(k)}(l,m)\right)$ is a 4x4 symmetric matrix with the following non-zero elements

$$g_i^{(k)}(1,1) = \left(\theta_{ir}^{(k)} - \theta_{in}^{(k)}\right)\rho_i^{(k)}\left(1-\rho_i^{(k)}\right)\left(1-2\rho_i^{(k)}\right)$$

$$g_i^{(k)}(1,2) = \theta_{ir}^{(k)}\left(1-\theta_{ir}^{(k)}\right)\rho_i^{(k)}\left(1-\rho_i^{(k)}\right)$$

$$g_i^{(k)}(1,4) = -\theta_{in}^{(k)}\left(1-\theta_{in}^{(k)}\right)\rho_i^{(k)}\left(1-\rho_i^{(k)}\right)$$

$$g_i^{(k)}(2,2) = \theta_{ir}^{(k)}\left(1-\theta_{ir}^{(k)}\right)\rho_i^{(k)}\left(1-2\theta_{ir}^{(k)}\right)$$

$$g_i^{(k)}(4,4) = \left(1-\rho_i^{(k)}\right)\theta_{in}^{(k)}\left(1-\theta_{in}^{(k)}\right)\left(1-2\theta_{in}^{(k)}\right)$$

In equation (17), $H_{(j)}$ depends on j because the quantities are computed with $\left(\underline{\alpha}, \underline{\beta}\right) = (\underline{\alpha}^{(k)}, \underline{\beta}_j)$.

The second partial derivation matrix defined in (13) is the sample information matrix and its expectation is used in the Fisher scoring algorithm, which differs from the Newton-Raphson algorithm by the using the expected rather than the sample information matrix in (12). Since the residuals defined in (17) satisfy $E\left(\tilde{x}_i^{(k)}\right) = 0$, the Fisher scoring algorithm replaces (19) with $\Gamma_i^{(k)} = D_i^{(k)} + F_i^{(k)}$ where $F_i^{(k)} = n_{x,i}^{'}\theta_i^{(k)}\left(1-\theta_i^{(k)}\right)\tilde{\underline{h}}_i^{(k)}\tilde{\underline{h}}_i^{(k)T}$. Under suitable regularity conditions, the variance covariance of the mle $Var(\underline{\hat{\beta}}) \cong H^{-1}$ -- the inverse of the Hessian matrix.

3.2 Small area estimation (SAE)

We use the estimates of the final step of the algorithm to estimate the county prevalence rate, λ_i, for the i^{th} county. We estimate $\underline{\Psi}_i$ by

$$\underline{\hat{\Psi}}_i = \underline{\hat{\alpha}}_i + Z_i\underline{\hat{\beta}} \qquad (20)$$

where $\underline{\hat{\beta}}$ is the m.l.e. of $\underline{\beta}$ and $\underline{\hat{\alpha}}_i = N^{-1}\sum_{k=1}^{N}\underline{\alpha}_i^{(k)}$ is obtained from the final step of the algorithm. Then, $\underline{\hat{\xi}}_i = \left(\hat{\rho}_i, \hat{\theta}_{ir}, \hat{\phi}_i, \hat{\theta}_{in}\right)$ is obtained from $\underline{\hat{\Psi}}_i$ using the 1-1 transformation between $\underline{\xi}_i$ and $\underline{\Psi}_i$. Finally, using (4) we estimate λ_i through

$$\hat{\lambda}_i = P_i\hat{\theta}_i + (1-P_i)\hat{\phi}_i \qquad (21)$$

where

$$\hat{\theta}_i = \hat{\rho}_i\hat{\theta}_{i,r} + (1-\hat{\rho}_i)\hat{\theta}_{i,n} \qquad (22).$$

Using a first order Taylor expansion we approximate the variance of the estimate as

$$Var\left(\hat{\lambda}_i\right) \cong \underline{\kappa}_i^T Var\left(\underline{\hat{\Psi}}_i\right)\underline{\kappa}_i \qquad (23)$$

with

$$\underline{\kappa}_i^T = \left(\partial\lambda_i/\partial\underline{\Psi}_i\right)^T_{\underline{\Psi}_i=\underline{\hat{\Psi}}_i} = [P_i(\theta_{ir} - \theta_{in})\rho_i(1-\rho_i),$$

$$P_i\rho_i\theta_{ir}(1-\theta_{ir}),(1-P_i)\phi_i(1-\phi_i),P_i\theta_{in}(1-\theta_{in})]_{\underline{\xi}_i=\hat{\underline{\xi}}_i} \quad (24)$$

and

$$Var\left(\hat{\underline{\Psi}}_i\right)=Var\left(\hat{\underline{\alpha}}_i+Z_i\hat{\underline{\beta}}\right)\cong N^{-1}\hat{\Sigma}+Z_iH^{-1}Z_i^T \quad (25)$$

where *N* random effects are generated for each county, Z_i is defined in (5) and *H* is the Hessian defined in (13).

3.3 State and National estimation

We are also interested in prevalence estimates for areas larger than counties (i.e., states and the nation) by age and gender categories: for example, male current smoking for those 18 years and older. We denote the number of people in the i^{th} county in the age gender category as W_i and the normalized population as $W_i^{'}=W_i/\sum_i W_i$. Then, the prevalence estimate $\hat{\lambda}_R$ for an area *R* composed of counties can be obtained as the average of the county estimates

$$\hat{\lambda}_R=\sum_{i\in R}W_i^{'}\hat{\lambda}_i \quad (26)$$

and its variance can be approximated by

$$Var\left(\hat{\lambda}_R\right)=\sum_{i\in R}\left(W_i^{'}\right)^2Var\left(\hat{\lambda}_i\right) \quad (27)$$

where $Var\left(\hat{\lambda}_i\right)$ is obtained from (23).

3.4 Non-response bias estimation

A goal is to determine if there is significant non-response bias; for example, the "middle class bias" described by Goyder *et al.* (2002). From (5) we have

$$\text{logit}(\theta_{i,r})-\text{logit}(\theta_{i,n})=\left(\alpha_{i,2}-\alpha_{i,4}\right)+\underline{U}_i^T\left(\underline{\beta}_{\theta,r}-\underline{\beta}_{\theta,n}\right)$$

where $\alpha_{i2}-\alpha_{i4}\sim N(\mu_2-\mu_4,\sigma_{22}+\sigma_{44}-2\sigma_{24})$. It follows that there is significant non-response bias if either $\mu_2-\mu_4\neq 0$ or if $\underline{\beta}_{\theta,r}-\underline{\beta}_{\theta,n}\neq\underline{0}$. Standard normal theory statistical tests for these hypotheses can be carried out using estimates of Σ and $Var(\hat{\underline{\beta}})$ respectively. To assess the impact of the individual covariates, it is also useful to carry out statistical tests of the components of $\left(\underline{\beta}_{\theta,r}-\underline{\beta}_{\theta,n}\right)$.

3.5 Initial values for the estimates

We choose initial values for the algorithm using the marginal distributions of (5). For example, we estimated $\left(\underline{\beta}_{\rho},\sigma_{11}\right)$ using the GLMM model for w_i, which satisfies (3a) with

$$\text{logit}(\rho_i)=\alpha_{i,1}+\underline{U}_i^T\underline{\beta}_{\rho} \text{ where } \alpha_{i1}\sim N\left(\mu_1,\sigma_{11}\right)$$

We used the estimate of α_{i1} as an initial estimate of μ_1. We obtain estimates of these parameters using SAS Proc NLMIXED (SAS Institute Inc., 2004). Similarly, we obtain initial estimates for $\left(\mu_2,\underline{\beta}_{\theta,r},\sigma_{22}\right)$ using z_i with (3d) and for $\left(\mu_3,\underline{\beta}_{\phi},\sigma_{33}\right)$ using y_i with (3c). Ignoring non-response bias, we equated the initial values of $\left(\mu_4,\underline{\beta}_{\theta,n},\sigma_{44}\right)$ to those of $\left(\mu_2,\underline{\beta}_{\theta,r},\sigma_{22}\right)$.

4. Results

4.1 Simulation methodology

Twenty data sets were simulated from the model with the same value of $\underline{\Theta}$. The effective sample sizes were chosen to resemble a yearly telephone survey with a national sample size of 200,000, with equal state sample sizes, and county sample sizes proportional to population within each state. The overall telephone survey response rate was assumed to be 50% with state response rates ranging between 30% and 70%. The in-person survey was assumed to be of size 40,000 and to be concentrated in 800 counties.

The parameters (defined in Tables 3 and 4 below) were chosen so that the nationally weighted estimate would be approximately 20% - close to the national level of male current smoking. Also, we used the smoking rates for telephone and non-telephone households specified as in Raghunathan *et al.* (2007). We used the populations, W_i, corresponding to males 18 and older from Census 2000 in equations (26) and (27) to evaluate the state and national estimates.

We used the following five county-level covariates: u5, Per capita property taxes; u8, Percent of persons below poverty; u10, Civilian labor force unemployment rate; u15, Buying power index; and u18, population. These were selected from the 18 covariates used by Raghunathan *et. al.* (2007) and were selected to be related to county SES; thus, q = 5 so that U_i is a 5-dimensional vector.

The 20 datasets were generated as follows:

- Generate $\underline{\alpha}_i\sim N(\underline{\mu},\Sigma)$ for i=1,…,A
- Calculate $\underline{\psi}_i$ from (5),
- Calculate $\left(\rho_i,\theta_{ir},\phi_i,\theta_{in}\right)$ and θ_i from (1)
- Generate v_i using 3a-3d

The county estimates, $\hat{\lambda}_i$, were obtained from (21) and (22) while the true county values are obtained from (1)

and (4). Although $\underline{\Theta}$ is fixed, the random effects cause the true small area values to vary over the simulations.

The estimation algorithm was coded in S-Plus (1999). The run time of the algorithm was approximately 16 hours per simulated data set on a personal computer. Important determinants of the run time are the MAX number of iterations (500 used here) and the number of $\underline{\alpha}$ vectors generated for each county (5 used here).

4.2 **Estimation accuracy**

Table 3 summarizes the estimation accuracy for the model parameters, $\underline{\mu}$ and $\underline{\beta}$. The first two columns specify the parameter; while the remaining columns specify the true value, the standard error of the mean (Std. error), the Student t-statistic (T stat.), and the root mean square error (RMSE).

Table 3. Estimation accuracy for $\underline{\mu}$ and $\underline{\beta}$

Parameter		True value	Std. error	T stat.	RMSE
$\underline{\mu}$	μ_1	-0.1	0.0022	-2.88	0.012
	μ_2	-1.1	0.0037	1.93	0.018
	μ_3	-0.1	0.0339	-1.32	0.158
	μ_4	-0.9	0.0136	0.29	0.061
$\underline{\beta}_{\rho}$	u_5	-0.01	0.003	0.17	0.015
	u_8	-0.4	0.005	2.79	0.025
	u_{10}	0.5	0.004	-2.47	0.019
	u_{15}	0.1	0.003	3.61	0.017
	u_{18}	-0.3	0.003	2.40	0.015
$\underline{\beta}_{\theta,r}$	u_5	0.3	0.003	-3.01	0.019
	u_8	0.2	0.007	1.38	0.033
	u_{10}	-0.1	0.005	-1.44	0.023
	u_{15}	-0.1	0.007	1.98	0.032
	u_{18}	-0.05	0.003	0.34	0.015
$\underline{\beta}_{\phi}$	u_5	-0.25	0.028	1.01	0.130
	u_8	0.07	0.035	1.22	0.161
	u_{10}	0.1	0.025	-2.13	0.121
	u_{15}	0.1	0.041	0.63	0.186
	u_{18}	-0.2	0.028	0.51	0.128
$\underline{\beta}_{\theta,n}$	u_5	0.15	0.014	0.72	0.064
	u_8	-0.2	0.022	-2.22	0.110
	u_{10}	-0.2	0.011	2.96	0.061
	u_{15}	-0.1	0.027	-1.88	0.129
	u_{18}	-0.4	0.016	1.75	0.075

The largest effective sample sizes are obtained from the telephone survey, $n_{w,i}$, and $n'_{z,i}$; in general, the estimates obtained from these sources ((μ_1, μ_2), $\underline{\beta}_{\rho}$, and $\underline{\beta}_{\theta,r}$) have smaller standard errors and RMSEs than the estimates of ((μ_3, μ_4), $\underline{\beta}_{\phi}$, and $\underline{\beta}_{\theta,n}$).

Although the estimation accuracy is relatively good in table 3, there are more large T-statistics (in absolute value) than would be expected from the T-distribution with 20 degrees of freedom (d.o.f.). For example, 17% (=4/24) of the observed t-statistics are larger in absolute value than 2.85, which is the 1% percentage point of the Student distribution with 19 d.o.f.

Table 4 shows the estimation accuracy for results for the covariance matrix, Σ. The columns specify the parameter, the true value, the mean estimate over the simulations (Mean Estim.), the standard error of the mean (Std. error), the Student t-statistic (T stat.), and the root mean square error (RMSE). In general, the estimation accuracy of the covariance matrix was not as good as for the parameters of table 3. The estimates of the diagonal elements of Σ were biased as shown by the large t-statistics for each. However, the estimation results for the off-diagonal elements were good.

Table 4. Estimation accuracy for Σ

	True value	Mean Estim.	Std. error	T stat.	RMSE
σ_{11}	0.1	0.083	0.001	-18.9	0.017
σ_{12}	0.0	0.002	0.001	1.7	0.005
σ_{13}	0.0	0.019	0.012	1.6	0.055
σ_{14}	0.0	0.000	0.006	0.0	0.026
σ_{22}	0.1	0.048	0.002	-30.1	0.053
σ_{23}	0.0	0.000	0.011	0.0	0.050
σ_{24}	0.0	-0.023	0.004	-5.3	0.030
σ_{33}	0.8	0.849	0.068	0.7	0.307
σ_{34}	0.0	0.007	0.015	0.5	0.067
σ_{44}	0.1	0.139	0.008	5.1	0.051

Figure 1a shows the scatter plot of the small area estimate $\hat{\lambda}_i$ vs. the true value λ_i obtained from all 20 simulations for all counties, where the values are expressed as percentages (with the line that shows equality superposed). Figure 1b shows a similar scatterplot with values averaged (for true and estimated values) over the 20 simulations for each county.

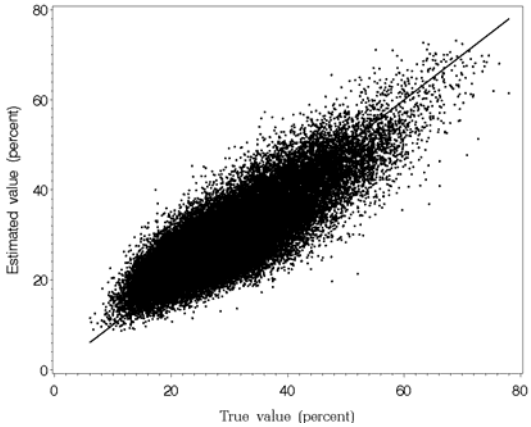

Figure 1a. True and estimated values for all counties and all 20 simulated datasets

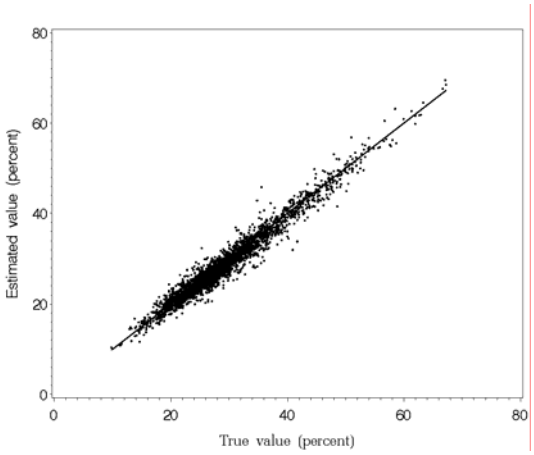

Figure 1b. True and estimated values for all counties averaged over all 20 simulated datasets

Figures 2a and 2b show scatterplots that are similar to figures 1a and 1b but at the state -- rather than the county level. Figure 2a shows the values for all states (including D.C.) and all simulations while figure 2b shows the average values for all states.

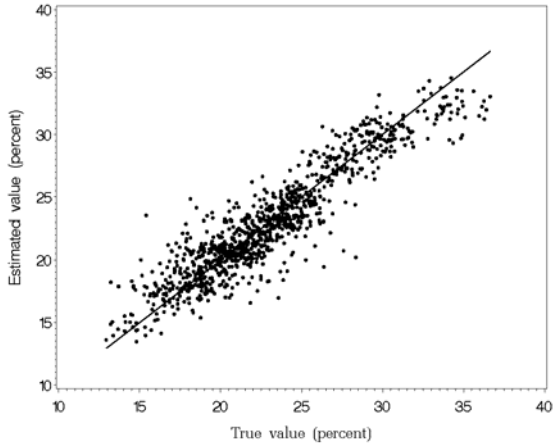

Figure 2a. True and estimated values for all states and all 20 simulated datasets

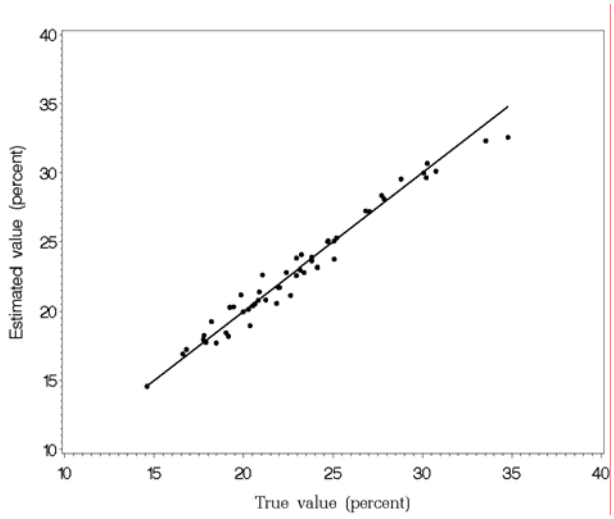

Figure 2b. True and estimated values for all states averaged over all 20 simulated datasets

Figure 3 shows the histogram of the national-level differences (estimated-true). All differences are less than 1 percent and they appear to be approximately uniformly distributed over this interval.

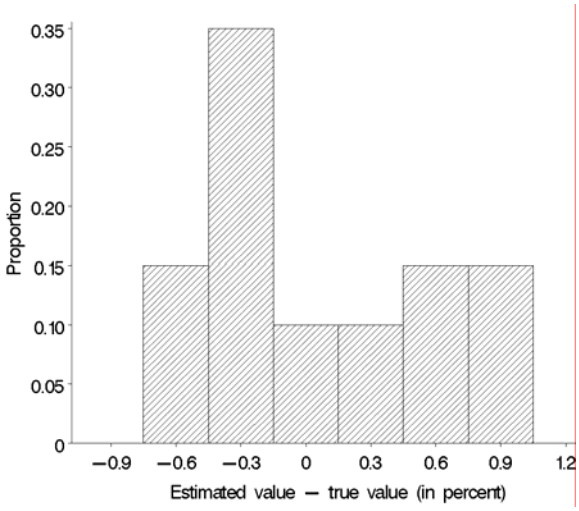

Figure 3. National level error histogram

In summary, none of the figures (1 through 3) indicate substantial estimation bias in the county, state or national estimates –even though Table 3 suggests that there may be a small estimation bias in some of the components of μ and β. The small parameter estimation biases may cancel out in the geographic estimates.

Table 5 provides statistics on the difference (estimated value - true value) for the county, state and nation. The RMSE difference is decomposed into mean difference and the standard deviation of the differences. In all cases, the mean difference is a negligible portion of

the RMSE. This supports the previous claim based upon graphics (Figure 1-3) that the small bias in the parameter estimates (Table 3) does not translate into appreciable bias of the geographic estimates.

Table 5. Difference between estimated and true values (in percent) for counties, states, and the nation

	RMSE	Mean difference	Std. dev. of differences
County all	4.96	-0.38	4.94
County avg.	1.58	-0.38	1.53
State all	1.77	-0.06	1.77
State avg.	0.76	-0.06	0.76
Nation all	0.51	0.07	0.51

The variance of county, state, and national estimates were calculated using equations (23) and (27). Then, the coverage of the 95% confidence intervals was estimated by determining the fraction of the true values that were contained in the interval $\hat{\lambda}_i \pm 1.96 * \sqrt{Var\left(\hat{\lambda}_i\right)}$.

Table 6 shows that the coverage probabilities for counties, states, and the nation. The estimated coverage probabilities are below the nominal value in all three cases. Since table 3 indicates that there is little if any bias in these estimates, the under-coverage probably arises from an under-estimate of the true variance.

Table 6. Coverage statistics for 95% CIs

	Number of estimates	Coverage percentage
County	62,280	84.1%
State	1,020	80.8%
National	20	85.0%

Table 7 shows coverage statistics for the non-response bias parameters. The table shows the true values with five of the six values non-zero. For each replication and each of the six parameters, a 95% confidence interval was calculated using the method sketched in section 3.4, and the number of intervals covering the true value was calculated.

Table 7. Coverage statistics for 95% confidence intervals for non-response bias parameters

Parameter vector	Parameter	True value	Coverage percentage
$\underline{\mu}$	$\mu_2 - \mu_4$	-0.2	100%
$\underline{\beta}_{\theta,r} - \underline{\beta}_{\theta,n}$	u_5	0.15	40%
	u_8	0.4	0%
	u_{10}	0.1	95%
	u_{15}	0.0	100%
	u_{18}	0.35	0%

The complement of the coverage percentage is an estimate of the power of the test. Although this is a small simulation, Table 7 suggests that there is very high power for detecting differences of the order of 0.35 or more in the (absolute value of the) components of $\underline{\beta}_{\theta,r} - \underline{\beta}_{\theta,n}$ (100% in both cases) and moderate power for detecting differences of the order of 0.15.

Acknowledgements

The authors wish to thank the following for assistance during this work: Ram Tiwari, Lan Huang, Eric "Rocky" Feuer, Kevin Dodd, Michael Elliott, Nat Schenker, Dawei Xie, Van Parsons, Trivellore Raghunathan, Ali Mokdad, Machelle Town, and Brenda Edwards.

References

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Statist. Soc. B*, 61, 265-285.

Cochran, W.G. (1977). *Sampling Techniques* (3rd ed.). John Wiley, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, 39, 1-38.

Elliott MR and Davis WW (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Applied Statistics*, 54, 595-609.

Goyder, J., Warriner, K., and Miller, S. (2002). Evaluating socio-economic status (SES) bias in survey nonresponse. *J. Off. Statist.*, 18, 1-12.

Groves, R.M., Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Assn.*, 92, 162-170.

Raghunathan TE, Xie D, Schenker N, Parsons V, Davis WW, Dodd K and Feuer EJ, (2007). Combining Information from Two Surveys to Estimate County Level Prevalence Rates of Cancer Risk Factors and Screening. *J. Am. Statist. Assn.* (to appear).

Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley, New York.

SAS Institute Inc. (2004). SAS/STAT 9.1 User's Guide. Cary, NC.

S-PLUS 2000 User's Guide (1999). Data Analysis Products Division, MathSoft, Seattle, WA.

Van Goor, H. and Rispens, S. (2004). A middle class image of society. *Quality and Quantity*, 38, 35-49.