Small Area Estimation for the National Household Survey of Drug Abuse

Ralph E. Folsom and Jun Liu, Research Triangle Institute
P. O. Box 12194, Research Triangle Park, NC 27709-2194

1. Introduction

In this paper, we summarize our planned approach for producing small area estimates of drug and alcohol use for selected U.S. States and metropolitan statistical areas (MSAs). The small area statistics of primary interest are population prevalence rates of illicit drug and alcohol dependency as ascertained from responses to National Household Survey of Drug Abuse (NHSDA) questionnaire items. We plan to produce separate rates for 32 demographic subpopulations defined by the three way cross-classification of sex with four age intervals {[12, 17]; [18,25]; [26,34]; and [35 plus]} and four race/ethnicity groups [Hispanics, Whites, Blacks, and Other Races] where the three non-hispanic categories will exclude hispanics. While we expect most of our small area statistics for the other races category to be suppressed due to excessive mean squared error estimates, the inclusion of this other race category will permit us to report statistics for Hispanics, Whites and Blacks employing the same race/ethnicity definitions typically used in government data sources.

2. Small Area Estimation Strategy

Our small area estimation strategy will begin by predicting ***block group level*** drug or alcohol dependency/use rates for all the 1990 block groups in a state or MSA small area by demographic domain (the 32 sex by age by race/ethnicity subpopulations). In addition to a vector $X3_{ijkd}$ of person level indicators for domain-*d*, our predictors for state small area-*i* include:

$X2_{ijk} \equiv$ Block group (ijk) and associated tract level 1990 census variables.

and

$X1_{ij} =$ County (ij) drug and alcohol related arrest, treatment, and death rates.

We propose to use a logistic regression predictor denoted by $\hat{\pi}_{ijkd}$ for the drug or alcohol depency/use rate of domain-*d* in block group-*k* for MSA/County-*j* in state-*i*.

State specific small area estimates are planned for the 25 states listed in table 1. The associated counts of distinct MSA/County units, block groups, and responding persons presented in table 1 represent the numbers of units for which NHSDA respondent data and associated 1990 census data have been linked. These table 1 numbers of NHSDA survey units are pooled counts aggregated across the 1991, 1992, and 1993 editions of the NHSDA.

We planned initially to limit the number of states for which separate small area estimates were produced, to those states with four or more distinct MSA/County primary sampling units (PSUs) represented in the pooled 1991, 1992, and 1993 surveys. This limitation is motivated by our plan to estimate nested random effects for each reported small area. These state specific random effect estimates adjust for prediction error in the fixed effect portion of the model. We expect that the purely 'synthetic' small area estimates that would result from a fixed effect logistic model without the estimated state random effects would have significantly larger prediction errors. Since the state specific random effects are based on observed deviations between a direct survey data based estimate for the small area and the associated 'synthetic' or fixed effect model predictor, the mean-squared-error's (MSE's) of these random effects and the associated mse's for our state estimates will depend on the variance of the direct survey estimates.

As indicated in table 1, we have included three states for separate small area estimation that did not satisfy our initial 'four or more' MSA/County unit rule of thumb. We have allowed these exceptions based on a speculation that their between County/MSA variance contributions may be small compared to their block group and person level variance contributions. In this case, their substantial respondent and block group sample sizes will justify the exception. For the nineteen additional states, that were represented in our pooled 1991 through 1993 NHSDA sample by three or fewer MSA/County PSUs, we plan to produce small area statistics for four regional residuals. Specifically, we will produce estimates for four regional residual groups consisting of all those states in each of the four census regions that are not covered by state specific statistics. In addition to the 25 separate state small areas and four regional residuals, there are twenty-seven MSA primary sampling units with forty-eight or more sample block groups and four hundred plus respondents that are candidates for separate small area estimation.

Returning to a description of our small area estimation strategy, we will first produce domain-*d* specific estimates for small area-*i* by combining our block group-*ijk* level prevalence rate predictors $\hat{\pi}_{ijkd}$

together using projected 1992 population counts of domain-*d* members in block group-*ijk*. These 1992 population projections for all U.S. block groups by age, race/ethnicity, and sex have been purchased from a commercial vendor. Denoting these domain specific population projections by $\hat{N}^{'92}_{ijkd}$, the associated state-*i* prevalence rate is computed as follows

$$\hat{\pi}_{id} \equiv \left(\sum_{j \epsilon \Omega_i} \sum_{k \epsilon \Omega_{ij}} \hat{N}^{'92}_{ijkd} \; \hat{\pi}_{ijkd} \right) \div \hat{N}^{'92}_{i++d} \qquad (1.0)$$

where Ω_i denotes the set of all MSA/County units in state-*i* and Ω_{ij} depicts the set of all 1990 block-groups in MSA/County unit-*ij*. The denominator count $\hat{N}^{'92}_{i++d}$ in equation (1.0) symbolizes the total projected population for domain-*d* in state-*i*. The associated drug or alcohol dependency/use rate per person aged 12 or older for state-*i* is therefore

$$\hat{\pi}_i \equiv \left(\sum_{d=1}^{32} \hat{N}^{'92}_{i++d} \; \hat{\pi}_{id} \right) \div \hat{N}^{'92}_{i+++} \qquad (2.0)$$

The age interval, race/ethnicity group, and sex specific estimates will be produced similarly. In the next section we present the set of block group, tract, and county level variables that are being considered as regressors. We also list a set of eleven binary survey outcome measures that will serve as dependent variables for our logistics models.

3. Candidate Regressors and Dependent Variables

Figure 1 lists seventeen groups of candidate regressor variables that have been linked to our 1991 through 1993 survey data. Both block group and tract level versions of these variables have been obtained. Since these variables are long form sample (~8%) variables, the tract level versions have less sampling error. On the other hand, the block group versions are closer geographic matches to our second stage area segments which are typically comprised of two or three census blocks. We plan to preferentially include the tract level variables and add any block group versions that are also significant.

We have obtained county level regressors from three sources. The first of the county level sources is the FBI's Uniform Crime Reports data base for 1991. From this source, we have formed arrest rates per 1000 persons for illegal drug possession and for sales/manufacture by drug category. We have also included county level 1991 total violent crime arrest rates. Our second source of county level regressors combines data from the 1991 and 1992 National Drug Abuse Treatment Unit Survey (NDATUS) conducted by the Substance Abuse and Mental Health Services Administration. From this source we have formed 1991 and 1992 average treatment rates per 1000 county residents for ***alcohol alone*** and for ***illicit drugs alone or both drugs and alcohol***. Finally, we have obtained for 1990 a set of alcohol related death rates per 1000 county residents. The source of these death rates is the National Center for Health Statistics.

The person level binary dependent variables for which we will fit separate models and produce associated small area estimates include:

1.) Dependent on ***any illicit drug*** (not on alcohol).
2.) Dependent on ***alcohol*** (not on any illicit drug).
3.) Dependent on both ***alcohol*** and ***any illicit drug***.
4.) Past Month alcohol user.
5.) Past Month any illicit drug user.
6.) Past Month user of any illicit drug other than Marijuana.
7.) Past Month Cocaine User.
8.) Past Month Cigarette User.
9.) Past Year treatment for alcohol abuse (only)
10.) Past Year treatment for drug abuse only or for drug and alcohol abuse.
11.) Past Year arrest for a nontraffic offence.

In the following section we present the explicit form of our nested random effects logistic regression model.

4. The Logistic Model With Nested Random Effects

If we let y_{ijkl} denote one of the zero-one dependent variables listed above for responding person-*l* in block group-*ijk*, and define $X_{ijkl} \equiv (1, X1_{ij}, X2_{ijk}, X3_{ijkl})$ as the vector of county (X1), block group (X2), and person level regressors (X3, plus significant X1⊗X3 and X2⊗X3 interactions), then we employ the following model for the probability that y_{ijkl}=1 given the regressors X_{ijkl}, the associated fixed effect coefficients β, and the nested random effects $\eta_{ijk} = \eta_{1i} + \eta_{2ij} + \eta_{3ijk}$:

$$\textbf{\textit{Prob}}(y_{ijkl} = 1 \,|\, X_{ijkl}\; \beta + \eta_{ijk}) = \pi_{ijkl}$$

where (3.0)

$$\pi_{ijkl} \equiv [1 + \exp\{-(X_{ijkl}\; \beta + \eta_{ijk})\}]^{-1}$$

The random effects for state-*i* (η_{1i}), for MSA/County unit-*ij* (η_{2ij}), and for block-groups-*ijk* (η_{3ijk}) are assumed to be independent gaussian random variables with zero means and variances that are inversely proportional to the average survey weight for the associated cluster. Specifically, if the W_{ijkl} denote initial survey analysis weights W_{0ijkl} divided by the scale factor

$$\bar{W}_0 \equiv \sum_{i=1}^{a} \sum_{j=1}^{n_i} \sum_{k=1}^{r_{ij}} \sum_{l=1}^{m_{ijk}} W_{0ijkl} \div a\; n_i\; r_{ij}\; m_{ijk}$$

then we assume that

$$Var(\eta_{3ijk}) = (\sigma_3^2 \div \bar{W}_{ijk}) \qquad (4.0)$$

with

$$\bar{W}_{ijk} \equiv \Big(\sum_{l=1}^{m_{ijk}} W_{ijkl} \div m_{ijk}\Big)$$

Similarly, we assume that

$$Var(\eta_{2ij}) = (\sigma_2^2 \div \bar{W}_{ij}) \qquad (5.0)$$

with

$$\bar{W}_{ij} \equiv \Big(\sum_{k=1}^{r_{ij}} \bar{W}_{ijk} \div r_{ij}\Big)$$

Finally, we assume that the state level random effects have variance

$$Var(\eta_{1i}) = (\sigma_1^2 \div \bar{W}_i) \qquad (6.0)$$

with

$$\bar{W}_i \equiv \Big(\sum_{j=1}^{n_i} \bar{W}_{ij} \div n_i\Big)$$

To specify the survey design weighted empirical-Bayes solution for our nested random effects logistic model, we employ a design weighted version of Breslow and Clayton's (1993) working linear model. If we let $\hat{\beta}_g$ denote the g-th itteration estimate of the fixed-effect coefficients and $\hat{\eta}_g^T \equiv (\eta_{1g}^T, \eta_{2g}^T, \eta_{3g}^T)$ denote the g-th iteration estimate of the full set of sample state (η_{1g}), MSA/County (η_{2g}), and block-group (η_{3g}) random effects, then the working linear model for the g-th iteration is

$$Y_{gijkl} = X_{ijkl}\,\beta_g + Z_{ijk}\,\hat{\eta}_g + e_{gijkl}$$

where

$$e_{gijkl} \equiv (y_{ijkl} - \pi_{gijkl}) \div [W_{ijkl}\,\pi_{gijkl}\,(1-\pi_{gijkl})]$$

and

$$Z_{ijk} \equiv (z_{1i}\; z_{2ij}\; z_{ijk})$$

with the z_{1i}, z_{2ij}, and z_{3ijk} denoting vectors of one-zero indicator variables that respectively pick off the i-th element of η_{g1}, the ij-th element of η_{g2}, and the ijk-th element of η_{g3}. If we further define $\underset{\sim}{Y}_g$ and $\underset{\sim}{e}_g$ as the full sample column vectors of the Y_{gijkl} and e_{gijkl} variates, then the matrix form of our working linear model is

$$\underset{\sim}{Y}_g = X\hat{\beta}_g + Z\hat{\eta}_g + \underset{\sim}{e}_g \qquad (7.0)$$

Now, we can specify the approximate covariance matrix for $\underset{\sim}{Y}_g$ in the following mixed linear model form

$$Cov(Y_g) \approx V_g \equiv ZD_g\,Z^T + R_g \qquad (8.0)$$

where

$$R_g \equiv \mathrm{Diag}\big\{[W_{ijkl}^2\,\pi_{gijkl}\,(1-\pi_{gijkl})]^{-1}\big\}$$

and

$$D_g \equiv BLK\text{-}DIAG\;(D_{gt};\; t=1,2,3)$$

with

$$D_{gt} \equiv \sigma_{gt}^2\,\underset{\sim}{\bar{W}}_t^{-1}$$

where

$$\underset{\sim}{\bar{W}}_1 \equiv Diag(\bar{W}_i;\; i=1(1)a)$$

$$\underset{\sim}{\bar{W}}_2 \equiv Diag(\bar{W}_{ij};i=1(1)a,\; j=1(1)n_i)$$

and

$$\underset{\sim}{\bar{W}}_3 \equiv Diag(\bar{W}_{ijk};i=1(1)a,\; j=1(1)n_i,\; k=1(1)r_{ij})$$

This notation leads to the following survey design weighted empirical-Bayes or quasi-likelihood estimators for $\hat{\beta}$ and $\hat{\eta}$ at iteration g+1

$$\hat{\beta}_{g+1} = (X^T\,V_g^{-1}\,X)^{-1}\,(X^T\,V_g^{-1}\,\underset{\sim}{Y}_g)$$

and (9.0)

$$\hat{\eta}_{g+1} = D_g\,Z^T\,V_g^{-1}(\underset{\sim}{Y}_g - X\hat{\beta}_g)$$

Our nested random effects model leads to a particular structure for V_g such that the algebraic form for V_g^{-1} is known. This permits us to algebraically derive estimation formulae for the elements of $\hat{\beta}_{g+1}$ and $\hat{\eta}_{g+1}$. The only numerical matrix inversion required is for the $(X^T V_g^{-1} X)$ matrix whose elements are derived algebraically. Following Breslow and Clayton, we use an approximate restricted maximum likelihood (REML) algorithm to estimate the vector

$$\underset{\sim}{\sigma}^2_{g+1} \equiv (\sigma^2_{(g+1)1} \text{ , } \sigma^2_{(g+1)2} \text{ , } \sigma^2_{(g+1)3})$$

of variance components involved in D_{g+1}. As with the elements of $(X^T V_g^{-1} X)$ and $(X^T V_g^{-1} \underset{\sim}{Y}_g)$, we derive algebraic expression for the elements of our REML score function $U(\underset{\sim}{\sigma}^2_g)$ and the associated three by three element information matrix $I(\underset{\sim}{\sigma}^2_g)$.

To illustrate the effects of survey weighting on the fixed and random effects estimators, we observe that the $\hat{\beta}$ solution algorithm in Eq. (9.0) satisfies the following survey weighted score functions:

$$\sum_{(ijkl)\in s} W_{ijkl} \, X^T_{ijkl} \, [y_{ijkl} - \pi_{ijkl} \, (\hat{\beta}, \hat{\eta})] = \underset{\sim}{\phi}_{p+1} \qquad (10.0)$$

where $\underset{\sim}{\phi}_{p+1}$ denotes the *(p+1)* element null vector.

As is the case for survey weighted fixed effect logisic regression, these equations guarantee that the W weighted mean of the $\hat{\pi}$ predictors equals the W weighted mean of the data variable y for any domain represented by an indicator variable in X_{ijkl}. The survey weighted random effect estimates mimic the form of their unweighted analogs with the survey weights used for averaging. For a linear regression model with only one level of clustering we get the result

$$\hat{\eta}_{1i} = \left\{\hat{\sigma}^2_1 \div [\hat{\sigma}^2_1 + (\hat{\sigma}^2_2 \div m_i)]\right\} (\bar{y}_i - \bar{X}_i \hat{\beta}) \qquad (11.0)$$

where $\bar{y}_i$ and $\bar{X}_i$ are survey weighted (W_{ij} weighted) sample means. With more than one stage of clustering, the $\bar{y}_i$ and $\bar{X}_i$ analogs are inverse variance weighted, sacrificing sample deisgn consistency for variance reduction. In the following section, we present results on interval estimation for survey weighted small area statistics.

6. Interval Estimation for Small Area Statistics

Recalling the matrix form of our working linear model, the posterior covariance matrices for $\hat{\beta}$ and $\hat{\eta}$ given D know are:

$$Cov \; (\hat{\beta}, \; \hat{\beta}^T) = (X^T \, V^{-1} \, X)^{-1}$$

$$Cov \; (\hat{\eta}, \; \hat{\eta}^T) = (D \; - \; DZ^T \, PZD)$$

where

$$P \equiv [V^{-1} \; - \; V^{-1} \, X(X^T \, V^{-1} X)^{-1} \, X^T \, V^{-1}]$$

and

$$Cov(\hat{\eta}, \; \hat{\beta}^T) \equiv D(Z^T \, V^{-1} \, X) \; (X^T \, V^{-1} \, X)^{-1}$$

To quantify the uncertainty associated with our nested random effect logistic model estimators, we plan to linearize the logit transformed versions of the $\hat{\pi}_{id}$ statistics. For states where the fraction of the domain-*d* population that resides in sample counties is negligible, the linearized form of $\ln[\hat{\pi}_{id} \div (1-\hat{\pi}_{id})]$ is proportional to

$$\tau_{id} \equiv \bar{\chi}_{\Omega id} \, \hat{\beta} \; + \; \hat{\eta}_{1i} \qquad (12.0)$$

where $\bar{\chi}_{\Omega id}$ is the *weighted* mean of the X_{ijkd} regressor vectors over all block-groups-(ijk) in the state-i universe (Ω_i). The weights used to produce this $\bar{\chi}_{\Omega id}$ average are proportional to the quantities

$$\alpha_{ijkd} \equiv (N^{`92}_{ijkd} \; \div \; N^{`92}_{i++d}) \hat{\pi}_{ijkd} (1-\hat{\pi}_{ijkd})$$

If we denote the sum of the α_{ijkd} over all block-groups in the state-i universe by $\hat{\pi}(1-\hat{\pi})_{\Omega id}$, then we can approximate the mean-squared-error of $\ln[\hat{\pi}_{id} \div (1-\hat{\pi}_{id})]$ given D by

$$mse\left\{\ln[\hat{\pi}_{id} \div (1-\hat{\pi}_{id})] \,|\, D\right\} = [\hat{\pi}(1-\hat{\pi})_{\Omega id} \div \hat{\pi}_{id}(1-\hat{\pi}_{id})]^2$$

$$\otimes \; mse(\tau_{id}|D) \qquad (13.0)$$

To approximate $mse(\tau_{id}|D)$ we can employ the posterior covariance matrices specified above with $\hat{D}$ replacing D. To account for the significant additional variation that may result from estimating the $\hat{\underset{\sim}{\sigma}}^2$ variance components in $\hat{D}$, we have derived a special case of Prashad and Rao's (1991) result for our three level nested random effects model.

References

Breslow, N.E. and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models", *JASA*, 88, 9-25.

Prasad, N.G.N. and Rao, J.N.K. (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators", *JASA*, 85, 163-171.

Table 1. Proposed State Small Areas

STATE GROUP	GROUP NAME	MSA/ COUNTIES	BLOCK GROUPS	RESPONDING PERSONS
0	TOTAL	182	8,603	84,927
1	CALIFORNIA	13	1,263	12,448
2	OHIO	12	268	2,044
3	TEXAS	11	534	5,502
4	FLORIDA	10	909	10,171
5	N. CAROLINA	9	246	2,070
6	PENNSYLVANIA	8	280	2,215
7	VIRGINIA	8	365	3,624
8	LOUISIANA	7	142	1,173
9	MISSOURI	6	133	1,161
10	NEW YORK	6	800	8,740
11	GEORGIA	5	120	1,082
12	ILLINOIS	5	727	8,130
13	INDIANA	5	115	983
14	KENTUCKY	5	137	1,329
15	MICHIGAN	5	178	1,205
16	NEW JERSEY	5	169	1,542
17	TENNESSEE	5	100	873
18	KANSAS	4	61	521
19	OKLAHOMA	4	72	561
20	OREGON	4	59	412
21	S. CAROLINA	4	61	458
22	WISCONSIN	4	55	475
23	MINNESOTA	3	83	720
24	WASHINGTON	3	74	697
25	WEST VIRGINIA	3	62	537

Figure 1 - Block Group and Tract Level 1990 Census Variables

1. Race x Hispanic

% White nonhispanic
% Black nonhispanic
% Hispanic
% Other

2. Education for persons 18 or older

% 0-8 years
% 9-12 years and no H.S. diploma
% H.S. graduate
% some college and no degree
% associate degree
% bachelors, graduate, or professional degree

3. Age

% 0-18 years
% 19-24 years
% 25-34 years
% 35-44 years
% 45-54 years
% 55-64 years
% 65 and over

4. Poverty

% families below poverty level

5. Public Assistance

% households with public assistance income

6. Disability

% persons 16-64 with a work disability

7. Household composition

% one-person households
% of households with female heads (no spouse present) with children under 18

8. Employment

% of men 16 years and older in the labor force
% of women 16 years and older in the labor force

9. Housing value - owner occupied units

Median value of owner occupied housing units

10. Housing rent - rental units

Median rents for rental units

11. Sex by marital status (persons 16 years and older)

% Females currently married and not separated
% Females separated, divorced, or widowed
% Females never married
% Males currently married and not separated
% Males separated, divorced, or widowed
% Males never married

12. Income

Median Household Income

13. Urbanicity

% of persons residing in an urban place

14. Urbanized Area

% of persons in an MSA urbanized area

15. Age of Housing

% of HUs built before 1939
% of HUs built from 1940 to 1949

16. High School Dropout Rate (Tract level only)

% of high school age children who have dropped out

17. Underclass Tract Indicator (Tract level only)

18. Hispanic Subpopulations

% of Hispanics that are Cuban
% of Hispanics that are Puerto Rican