Further Developments in a Hierarchical Bayes Approach to Small Area Estimation of Health Insurance Coverage: State-Level Estimates for Demographic Groups[*]

Mark Bauder, Steven Riesz, and Donald Luery
U.S. Census Bureau, Washington, DC 20233

Abstract

Fisher et al. (2006) developed a hierarchical Bayes model to estimate the number of people without health insurance within demographic groups for states. The Centers for Disease Control and Prevention are interested in estimates of women without health insurance by demographic groups in families that earn less than 200% of the poverty line. Our approach jointly models direct estimates from the Annual Social and Economic Supplement to the Current Population Survey (CPS ASEC), and Census 2000 Sample Data, tax, food stamp, and Medicaid data, using a multivariate, hierarchical approach. We have improved the preliminary model in Fisher et al. (2006) by adding census data, improving the mean and variance models for the direct estimates and the administrative records data, and developing a raking procedure. In addition, for variance estimation, we have developed a method that takes into account the variance of the direct estimates that are used in the raking procedure.

Key Words: uninsured, SCHIP, CDC, Fay-Herriot model, multivariate model, administrative records.

1. Introduction

Policymakers are interested in estimates of health insurance coverage. Our current research is funded in part by the Centers for Disease Control and Prevention (CDC). The CDC provides free screening services for breast and cervical cancer to low-income, uninsured, and under-served women through the National Breast and Cervical Cancer Early Detection Program (NBCCEDP). The CDC wants to estimate percent eligible (the proportion of low-income women eligible for the screening) and participation rates (of the people eligible, the proportion screened) for the screening services, for various demographic groups within states and counties. The estimates that we are developing will allow the CDC to make estimates of percent eligible and participation rates with an acceptable level of precision.

Evaluation of many other federal programs can benefit from state-level health insurance estimates. For example, the Centers for Medicare and Medicaid Services (CMS) could use estimates of uninsured children at or below 200% of poverty, by race and ethnicity, to help states target medical services towards at-risk children. At the county-level, Medicaid administrators would have an additional tool for distributing the state's federal allotment to counties where the need for medical services may be greatest. In general, federal, state, and county programs that deal with health issues could calculate eligibility and participation rates of services for the uninsured by income categories.

In order to meet the needs of policymakers, the Census Bureau's Small Area Health Insurance Estimates (SAHIE) program is developing model-based small area estimates of health insurance coverage at the state level for areas defined by age, race, Hispanic origin, sex, and income to poverty ratio (IPR), and at the county level for areas defined by age, sex, and IPR. This paper focuses on the state-level estimates. Small area estimates for 2001 based on the model described in this paper can be found at the SAHIE web site, http://www.census.gov/hhes/www/sahie/index.html.

2. The problem and our approach

To meet the needs of the CDC at the state level, we need to estimate the number of women who are eligible for screening services through the NBCCEDP. More specifically, for each state, the CDC would like estimates of the number of women that are below 200% and 250% of poverty and are uninsured, for certain age categories and race/ethnicity categories. The age categories are 18-64 (for cervical cancer screening), and 40-64 and 50-64 (for breast cancer screening). The race/ethnicity categories that the Census Bureau

[*] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

can reliably provide are White non-Hispanic, Black non-Hispanic, and Hispanic. We need to make estimates for below 200% of poverty and below 250% of poverty because states have different eligibility criteria for the screening services.

In order to make these estimates, we use an area level model (Rao 2003) in which the areas are defined by cross-classifications of state, age, race/ethnicity, sex, and IPR categories. There are 51 states including the District of Columbia, 5 age categories (0-17, 18-39, 40-49, 50-64, 65+), 4 race/ethnicity categories (White non-Hispanic, Black non-Hispanic, Hispanic, and Other non-Hispanic), 2 sex categories, and 3 IPR categories (0-200%, 200-250%, and >250% IPR). Note that we have chosen these particular categories so that we can make all of the estimates of interest using one model. We have included the 65+ age group even though it has virtually 100 percent insurance coverage because two of the administrative records data for income - tax exemptions and food stamps - include this age group. We do not include this age group in the insurance coverage model.

Our model is similar to the Fay-Herriot model (Fay and Herriot 1979), however, it differs in several respects: (1) we use a Bayesian model, (2) we model two direct estimates (an estimate of the number of people in income categories, and an estimate of the proportion of people with insurance) from the CPS ASEC, instead of one direct estimate, and (3) we model Census 2000 Sample Data estimates and administrative record data, instead of using them as predictors in a regression model. The administrative data that we use are tax exemptions, Food Stamp participation, and data from the Medicaid/SCHIP programs.

Fisher (2003) and Fisher and Gee (2004) addressed a fundamental assumption in regression models like the Fay-Herriot model that the predictors are measured without error. In their "errors-in-variables" approach, they modeled the predictors of poverty for the Small Area Income and Poverty Estimates program assuming that they possess non-negligible variances. This was extended to small area estimates of insurance coverage in Fisher et al. (2006).

In the basic area level model, let $\widehat{N}_i$ be the direct estimate of N_i, the small area population value to be estimated. The basic area level model is given by

$$\begin{aligned}\widehat{\theta}_i &= g(\widehat{N}_i) = \theta_i + e_i \\ \theta_i &= \mathbf{z}_i^T \beta + v_i\end{aligned}$$

where the sampling errors e_i are independent and the v_i are independent and identically distributed area-specific random effects. The predictors $\mathbf{z}_i^T = \left(\mathbf{x}_i^T, \mathbf{A}_i^T\right)$ may include both administrative data $\mathbf{A}_i = \left(A_{1i}, \ldots, A_{pi}\right)^T$ and other area-specific auxiliary data $\mathbf{x}_i$. Our approach models both the direct estimates and the administrative data as possibly nonlinear regressions of the N_i and the N_i are modeled by a generalized linear model

$$\begin{aligned}\widehat{N}_i &= N_i + e_i \\ g\left(N_i\right) &= \mathbf{x}_i^T \gamma + v_i \\ A_{ji} &= h_j\left(N_i\right) + u_{ji}, j = 1, \ldots, p\end{aligned}$$

where the sampling errors e_i are independent, the v_i are independent and identically distributed area-specific random effects, and the u_{ji} are independent random effects for the administrative data.

We model two direct estimates because we have auxiliary data that are related to either income or insurance coverage. We then combine these two quantities to get our estimate of interest. The two quantities are (1) $N_{IPR}[i,a,r,s,k]$, the number of people in state/age/race/sex/IPR cell, and (2) $P_{IC}[i,a,r,s,k]$, the proportion of people in state/age/race/sex/IPR cell that are insured. We combine these two quantities to get $N_{IC}[i,a,r,s,k] = P_{IC}[i,a,r,s,k] N_{IPR}[i,a,r,s,k]$, the number of insured. The number of uninsured can be obtained from $N_{UI}[i,a,r,s,k] = N_{IPR}[i,a,r,s,k] - N_{IC}[i,a,r,s,k]$. Throughout the paper, we will use i to index state, a to index age, r to index race, s to index sex, and k to index IPR.

We made the following improvements to the model initially developed in Fisher et al. (2006). The model with the improvements is fully described in the next section.

- model the CPS ASEC estimates in income categories, $\widehat{N}_{IPR}[i,a,r,s,k]$, as totals instead of proportions,
- model the sampling variances instead of using parameters from the CPS ASEC generalized variance function,
- increase the number of predictors in the income and insurance logistic models,
- model the administrative data as totals instead of proportions,
- include Census 2000 Sample Data,
- allow the expectation and variance parameters to vary by groups and the variances to change nonlinearly with size, and
- rake to national CPS ASEC direct estimates.

3. The current model

3.1 The first part of the model: income

3.1.1 The CPS ASEC direct estimate of the number in income categories

The CPS ASEC direct estimates are averages of estimates from three ASEC surveys - 2001 through 2003. These surveys collect income for calendar years 2000, 2001, and 2002. The average income estimates are centered at 2001.

We have the CPS ASEC direct estimate of the number in cell $[i, a, r, s, k]$, $\widehat{N}_{IPR}[i, a, r, s, k]$. We assume that $\widehat{N}_{IPR}[i, a, r, s, k]$ is normally distributed and unbiased:

$$\widehat{N}_{IPR}[i, a, r, s, k] \sim N\left(N_{IPR}[i, a, r, s, k],\ v_{\epsilon,IPR}[i, a, r, s, k]\right),$$

where we model the sampling variance, $v_{\epsilon,IPR}$, by

$$v_{\epsilon,IPR} = \lambda_1[r, k]POP[i, a, r, s]^{1+\lambda_2}\frac{q[i, a, r, s, k](1 - q[i, a, r, s, k])}{S[i]}$$

where $q[i, a, r, s, k]$ is the proportion of those in the i^{th} state who are in cell $[i, a, r, s, k]$, *i.e.* $q[i, a, r, s, k] = N_{IPR}[i, a, r, s, k]/POP[i]$ ($POP[i]$ is the demographic population estimate for state i), $S[i]$ is the number of households in the three CPS ASEC samples in state i, $POP[i, a, r, s]$ is a demographic population estimate for cell $[i, a, r, s]$, and the λ's are parameters to be estimated. The multiplicative variance parameter (λ_1) differs by race by IPR and λ_2 does not vary. The sampling errors are assumed to be independent across state/age/race/sex/IPR cells.

3.1.2 Proportions in income categories

We model $P_{IPR}[i, a, r, s, k]$, the proportion in IPR category k in state/age/race/sex cell $[i, a, r, s]$. It is important to note that we have the demographic population estimate $POP[i, a, r, s]$, which we consider to be known without error, therefore $P_{IPR}[i, a, r, s, k]$ is related to $N_{IPR}[i, a, r, s, k]$ by

$$N_{IPR}[i, a, r, s, k] = P_{IPR}[i, a, r, s, k]POP[i, a, r, s].$$

We model $P_{IPR}[i, a, r, s, k]$ as a 3-category logistic model with normally distributed model error:

$$P_{IPR}[i, a, r, s, k] \;=\; \frac{\exp(\mu_{IPR}[i, a, r, s, k]\ )}{\sum_{k=1}^{3}\exp\left(\mu_{IPR}[i, a, r, s, k]\right)}$$

where μ_{IPR} follows a normal linear model $\mu_{IPR}[i, a, r, s, k] = X[i, a, r, s, k]\beta_{IPR} + u_{IPR}[i, a, r, s, k]$. $X[i, a, r, s, k]$ is a row vector of predictors and β_{IPR} is a vector of regression coefficients. The model error, u_{IPR}, is normally distributed $u_{IPR}[i, a, r, s, k] \sim N(\ 0, \nu_{\mu,IPR})$, where $\nu_{\mu,IPR}$ is a constant to be estimated.

The predictors in the vectors $X[i, a, r, s, k]$ are:

- indicators of age, race and sex groups each interacted with indicators of the IPR categories (*i.e.*, there is a fixed effect for each age by IPR, race by IPR, and sex by IPR group)
- age by sex by IPR interactions
- age by race by IPR interactions
- interactions of a tax nonfiler rate with IPR.

The tax nonfiler rate is defined as $\frac{POP-TAX}{POP}$ where POP is the demographic population estimate for a domain, and TAX is the number of IRS exemptions for that domain.

3.1.3 The Census 2000 Sample Data estimates

We model the Census 2000 Sample Data estimates of the numbers in income categories as

$$CEN[i, a, r, s, k] \sim N\left(\widetilde{CEN}[i, a, r, s, k], v_{CEN}[i, a, r, s, k]\right)$$

where $\widetilde{CEN}[i, a, r, s, k] = \alpha[r]N_{IPR}[i, a, r, s, k]$ and $v_{CEN}[i, a, r, s, k] = \lambda_1[r]\left(\widetilde{CEN}[i, a, r, s, k]\right)^{\lambda_2[r]}$. The α's and λ's are estimated. Note that the parameters are allowed to differ by race. Here and below, we refer to the α's as expectation parameters, and the λ's as variance parameters.

3.1.4 Tax exemptions

The numbers of tax exemptions in 2001 are aggregated at a higher level. We model them as follows:

$$TAX[i, a_T, k_T] \sim N\left(\widetilde{TAX}[i, a_T, k_T], v_{TAX}[i, a_T, k_T]\right)$$

where a_T and k_T are the tax age and tax IPR categories, with $a_T = 1, 2$ and $k_T = 1, 2$. Here $a_T = 1$ represents age 0-17, $a_T = 2$ represents age 18 and over, $k_T = 1$ represents IPR $\leq 200\%$, and $k_T = 2$ represents IPR $> 200\%$. The mean of the distribution of the number of exemptions is given by

$$\widetilde{TAX}[i, a_T, k_T] = \sum_{a,r,k} \alpha[a_T, r, k_T] N_{IPR}[i, a, r, s, k]$$

where the sum is to the appropriate a_T and k_T level. The variance is $v_{TAX}[i, a_T, k_T] = \lambda_1[a_T, k_T] POP[i, a_T]^{\lambda_2}$. $POP[i, a_T]$ is the demographic population estimate for the ${a_T}^{th}$ tax age category within the i^{th} state. We estimate λ_1 but constrain λ_2 to be fixed at 1.7. λ_2 could not be reliably estimated but by setting it to a central value, 1.7, the diagnostics for the tax exemption data became acceptable.

3.1.5 Food Stamp participation

We model 2001 Food Stamp participation using only the 0-200% IPR category because people in households with income near or below the poverty line are eligible for Food Stamps. The detailed requirements can be found at http://www.fns.usda.gov/fsp/applicant_recipients/eligibility.htm. In particular $FS[i] \sim N\left(\widetilde{FS}[i], v_{FS}[i]\right)$ where $\widetilde{FS}[i] = \alpha \sum_{a,r,s} N_{IPR}[i, a, r, s, 1]$ and $v_{FS}[i] = \lambda \widetilde{FS}[i]$. The parameters α and λ are estimated.

3.2 The second part of the model: insurance

3.2.1 The CPS ASEC direct estimate of the proportion insured

We assume that the CPS ASEC direct estimate of the proportion insured, $\widehat{P}_{IC}$, follows

$$\widehat{P}_{IC}[i, a, r, s, k] \sim N(P_{IC}[i, a, r, s, k],\ v_{\epsilon,IC}[i, a, r, s, k]).$$

The sampling variance follows $v_{\epsilon,IC}[i, a, r, s, k] = \lambda f_i \frac{P_{IC}[i,a,r,s,k](1-P_{IC}[i,a,r,s,k])}{\widehat{N}_{IPR}[i,a,r,s,k]}$. This variance form is based on the generalized variance function (GVF) used for calculating CPS ASEC variances.[1] The f_i are state factors accounting for differences in sampling rates among states. The GVF is of the above form, except that N_{IPR} is in the denominator, and λ is a constant that has been determined. We chose to estimate λ rather than plug in the GVF factor because the GVF may not be appropriate for all levels of aggregation. We chose to use $\hat{N}_{IPR}$ in the denominator instead of N_{IPR} because otherwise the variance model for $\widehat{P}_{IC}$ would affect the estimates from the income model.

3.2.2 Proportions insured

We assume that the proportion insured follows a logistic model:

$$\text{logit}(P_{IC}[i, a, r, s, k]) = X[i, a, r, s, k]\beta_{IC} + u_{IC}[i, a, r, s, k]$$

with $u_{IC}[i, a, r, s, k] \sim N(0, \nu_{\mu,IC})$. The constant $\nu_{\mu,IC}$ is a parameter to be estimated. The $X[i, a, r, s, k]$'s are row vectors of predictors, and β_{IC} is a vector of regression coefficients. The predictors in the X vectors include (1) state, age, race, sex and IPR indicators, and (2) age by IPR, race by IPR, sex by IPR, age by sex, and race by sex interactions.

[1] See the "Source and Accuracy Statement" section in U.S. Census Bureau (2006). Note that in the document, the GVF has a small quadratic term that we ignore here.

3.2.3 Medicaid/SCHIP data

We model the 2001 Medicaid/SCHIP data via its distribution conditional on the number insured, N_{IC}. However, $N_{IC} = P_{IC} N_{IPR}$, therefore parameters from both the income and the insurance parts of the model appear in the model for the Medicaid/SCHIP data. We assume

$$MED[i,a,s] \sim N\left(\widetilde{MED}[i,a,s], v_{MED}[i,a,s]\right).$$

The mean is given by

$$\widetilde{MED}[i,a,s] = \sum_r \alpha[a,r] N_{IC}[i,a,r,s,1]$$

where $N_{IC}[i,a,r,s,1] = P_{IC}[i,a,r,s,1] N_{IPR}[i,a,r,s,1]$, the number insured in the IPR category 0 - 200%. The variance is $v_{MED}[i,a,s] = \lambda[a]\widetilde{MED}[i,a,s]$. We constrain the α's to be the same for all age groups except 0 to 17, and the same with the λ's. We also constrain the α's to be the same for White non-Hispanic and Other non-Hispanic race categories.

4. Model selection

4.1 Selecting predictors for the regression models

For the regression parts of the income and insurance models, we considered as possible predictors interactions of the categorical variables, as well as a continuous variable derived from tax exemptions. We decided which predictors to keep in the model by looking at approximate 95% confidence intervals for the regression coefficients. The confidence intervals were constructed by taking the posterior mean of the regression coefficient plus or minus two times the posterior standard deviation. We considered a predictor significant if the 95% confidence interval for its regression coefficient does not include 0. We generally kept a predictor in the model if it was significant. We applied this test to classes of predictors so that if one predictor in a class were significant, we kept all of the predictors in that class.

4.2 Selecting parameterizations

We sometimes allowed the expectation and variance parameters (the α's and λ's above) to vary by groups. In some cases, exploratory data analysis showed that some groups are predictive of the CPS ASEC direct estimate. These analyses suggested that we allow expectation parameters to vary by those groups. In other cases, we tried plausible parameterizations, and if the estimates of the parameters differed substantially, we generally allowed them to differ in the final model. The parameterizations are described in the model sections.

4.3 Model diagnostics

Our primary diagnostics for evaluating the form of the model are posterior predictive p-values (PPP-values) and standardized residuals. A PPP-value is defined as $Prob\left(T(y^{(rep)}, \theta^{(rep)}) \geq T(y^{(obs)}, \theta^{(rep)}) | data\right)$ for some function T, where $y^{(obs)}$ is the observed value of y, $\theta^{(rep)}$ is drawn from the posterior distribution of the parameter vector θ, and $y^{(rep)}$ is drawn from the posterior predictive distribution of y conditional on $\theta = \theta^{(rep)}$. We focus on two possibilities for T: $T_1(y,\theta) = y$ and $T_2(y,\theta) = (y - E(y|\theta))^2$. The former measures model fit with respect to the predicted mean of y, and the latter measures model fit with respect to the predicted variance of y. A large proportion of PPPs near 0 or near 1 indicates a poor model fit. We obtain standardized residuals by dividing the difference between an observed value and its predicted mean by its predicted standard deviation.

4.4 Diagnostics for the income model

For the income part of the model, we paid particularly close attention to the diagnostics for the CPS ASEC direct estimate. The average of the PPPs for the mean is 0.51, the mean of the PPPs for the variance is 0.51, and the average of the standardized residuals squared is 1.00. None of these results suggest that the model fits poorly. We plotted the PPP for the mean, the PPP for the variance, and the standardized residual against the log of the population, the log of the sample size, and the log of the posterior mean of $N_{IPR}[i,a,r,s,k]$. Figures 1, 2, and 3 show the plots of the standardized residuals. In the plots, there are no obvious patterns except in the plot of the standardized residual vs. the log of the posterior mean of $N_{IPR}[i,a,r,s,k]$. There is a group of points with small values of the posterior mean and negative standardized residuals which we should investigate further.

Diagnostics for the Census 2000 Sample Data estimates, tax exemptions, and Food Stamps are generally good. For the Census 2000 Sample Data estimates, the average of the PPPs for the mean is 0.43, the average of the PPPs for the variance is 0.43, and the average of the standardized residuals squared is 1.66. The first suggests that the means of the posterior predictive distribution of the Census 2000 Sample Data are too low, and the last two suggest that the predicted variances are too low. Figures 4 and 5 show plots of the standardized residual against the log of the population and the log of the posterior mean of $N_{IPR}[i, a, r, s, k]$, respectively. The plots appear to show a slight upward trend. The trend is most pronounced when the plot is restricted to points from the IPR > 250% category, as shown in Figure 6. Otherwise, the plots appear fairly featureless.

For the tax exemptions, the averages of the PPPs for the mean for each of the four cross-classifications of tax age and tax IPR range from 0.46 to 0.52. The averages of the PPPs for the variance for the four cross-classifications range from 0.47 to 0.49. The average of the standardized residuals squared range from 0.99 to 1.00.

The average of the PPPs for the mean for Food Stamps is 0.46, and the average of the PPPs for the variance is 0.47. These are consistent with a good model fit. The mean of the standardized residuals squared is 1.51, but the median is 0.42. There appear to be outliers in the standardized residuals that increase the average of the standardized residuals squared.

4.5 Diagnostics for the insurance model

We analyzed the diagnostics for the CPS ASEC direct estimate. The average of the PPPs for the mean is 0.48, the average of the PPPs for the variance is 0.49, and the average of the standardized residuals squared is 1.01. All of these values are consistent with a good model fit. The plots of the PPPs and the standardized residuals have some patterns in that the spread of the standardized residuals decreases as the population increases. This suggests that the predicted variance is too small for large values of the population.

For the Medicaid/SCHIP data, the average of the PPPs for the mean is 0.56, and the average of the PPPs for the variance is 0.53. The average of the standardized residuals squared is 1.40, with a median of 0.31. There appear to be outliers among the standardized residuals.

5. Raking to direct estimates

For each cross-classification of age, race/ethnicity, and sex, we control the estimate of the number insured to the national CPS ASEC estimate of the number insured. This is from the 2002 CPS ASEC which is used to estimate poverty for 2001. We do this for two reasons: (1) to make the small area estimates consistent with the national direct estimates, and (2) to correct for possible deficiencies of the model.

We control the estimates as follows: for each cross-classification of age, race/ethnicity, and sex, we sum the small area estimates of the number insured over states and income categories to get a national estimate. We then calculate a raking factor by dividing the CPS ASEC direct estimate by this national estimate. So the raking factor is $\widehat{N}_{IC}[a, r, s] \Big/ \sum_{i,k} N_{IC}[i, a, r, s, k]$.

We then get a raked estimate of the number insured for each cross-classification of state, age, race/ethnicity, sex, and IPR by multiplying the raking factor by the small area estimate of the number insured:

$$N_{IC}^{raked}[i, a, r, s, k] = \frac{\widehat{N}_{IC}[a, r, s]}{\sum_{i,k} N_{IC}[i, a, r, s, k]} N_{IC}[i, a, r, s, k].$$

We obtain the number uninsured by subtracting the raked number insured from the number in the IPR category, $N_{UI}^{raked}[i, a, r, s, k] = N_{IPR}[i, a, r, s, k] - N_{IC}^{raked}[i, a, r, s, k]$.

5.1 Accounting for the variance of the direct estimates

In our estimates of the variance of our raked small area estimates, we wanted to take into account the variance of the direct estimates $\widehat{N}_{IC}[a, r, s]$ that were used as controls in the raking. To do this, we did a separate run of the Markov Chain Monte Carlo simulation with the following change: for the direct estimates $\widehat{N}_{IC}[a, r, s]$ that were used as controls, we estimated their variances using the GVF method. We then treated the controls as random quantities with normal distributions and the estimated variances. In each iteration of the MCMC simulation, instead of using $\widehat{N}_{IC}[a, r, s]$, we used a draw from a normal distribution with mean $\widehat{N}_{IC}[a, r, s]$ and variance from the GVF method. This allowed the variances of the raked small area estimates to reflect contributions from the variances of the controls. This method assumes that the direct estimates $\widehat{N}_{IC}[a, r, s]$ are independent from the small area estimates of proportions $N_{IC}[a, r, s] \Big/ \sum_{i,k} N_{IC}[i, a, r, s, k]$.

6. Future research

Sampling variances and correlations. Further research into the variances of the direct estimates for both the income and insurance parts of the model is needed. One possibility is to investigate alternative functional forms for the variances. Another possibility is to estimate the variances outside of the model, using the CPS ASEC replication method and modeling to smooth the large variability of the variance estimates.

The models for the direct estimates assume that the sampling errors are independent. Because of household clustering in the CPS ASEC, we would expect significant correlations among the direct estimates. Research would be needed to include these correlations in the modeling of the direct estimates.

Modeling the Census 2000 Sample Data estimates. When we tried some alternate parameterizations in modeling the Census 2000 Sample Data estimates, we obtained some surprising results. When we allowed the variance parameters to differ over some demographic groups, we obtained very different estimated values for those parameters. We should investigate to determine the underlying reason for this. We also noted a trend in the plot of the standardized residuals against the predicted number insured. This suggests that we consider alternative functions or parameterizations for the mean of the Census 2000 Sample Data estimate. Further, there is reason to expect that the variance of the Census 2000 Sample Data has two components. One component is due to sampling error in the Census 2000 Sample Data and the other component is due to model error.

Modeling the administrative record data. We should attempt to improve the models for the administrative data: tax, Food Stamps and Medicaid/SCHIP. Some of the diagnostics showed outliers or trends that suggest areas for research. We especially should consider alternate forms and parameterizations for the variance functions. We should also consider alternatives for the expectation functions.

Raking. In order to obtain consistency with published national direct estimates at the age by sex and race by sex levels, we controlled the estimates to the national age by race by sex direct estimates for the number insured. We found that the variances of the direct estimates controlled to were not negligible, and investigation showed that when the variance of the control was not taken into account, the variances of the estimates could be substantially underestimated. We should also reconsider the value of forcing modeled estimates to match direct estimates, when the variances of the direct estimates are not negligible such as controlling to age by sex and race by sex estimates by two-way raking.

Other data sources. We should look for other sources of data that could be used as predictors in the IPR or IC logistic regression models (sections 3.1.2 and 3.2.2), or could be modeled conditional on IPR or IC numbers.

References

Fay, R.E., and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data", *Journal of the American Statistical Association*, 74, 269-277.

Fisher, R. (2003), "Errors-In-Variables Model for County Level Poverty Estimation", SAIPE Working Paper, Washington, DC, U.S. Census Bureau.

Fisher, R. and Gee, G. (2004), "Errors-In-Variables County Poverty and Income Models",*2004 American Statistical Association Proceedings of the Section on Government and Social Statistics.*

Fisher, R., O'Hara, B. and Riesz, S. (2006), "Small Area Estimation of Health Insurance Coverage: State-Level Estimates for Demographic Groups", *2006 American Statistical Association Proceedings of the Section on Government and Social Statistics.*

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

U.S. Census Bureau (2006), "Current Population Survey, 2006 Annual Social and Economic (ASEC) Supplement", available from http://www.census.gov/apsd/techdoc/cps/cpsmar06.pdf.

Figure 1: Income: Standardized residual vs. log population

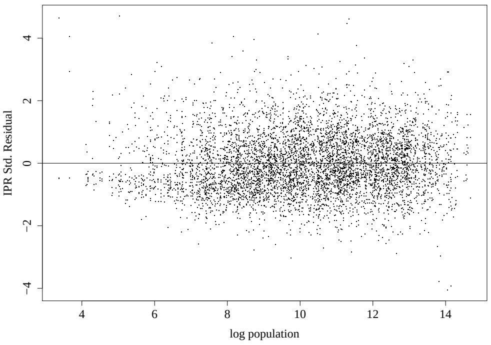


Figure 2: Income: Standardized residual vs. log sample size

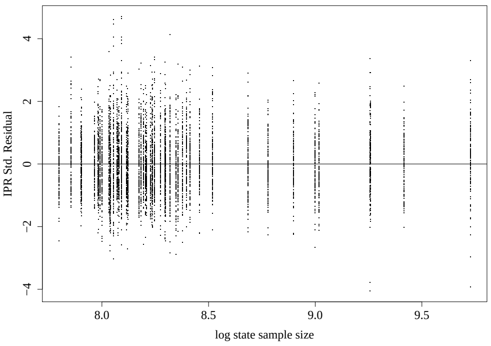


Figure 3: Income: Standardized residual vs. log posterior mean

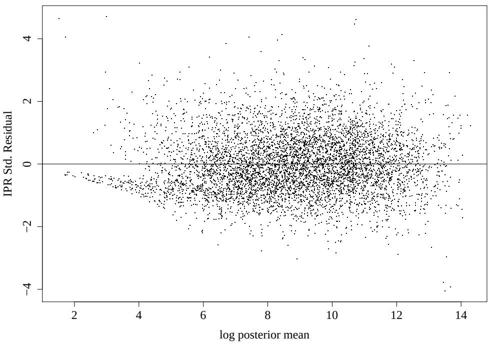


Figure 4: Census: Standardized residual vs. log population

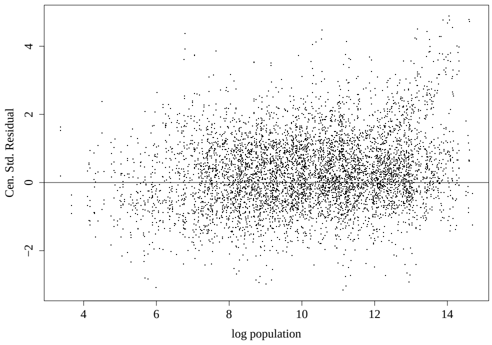


Figure 5: Census: Standardized residual vs. log posterior mean

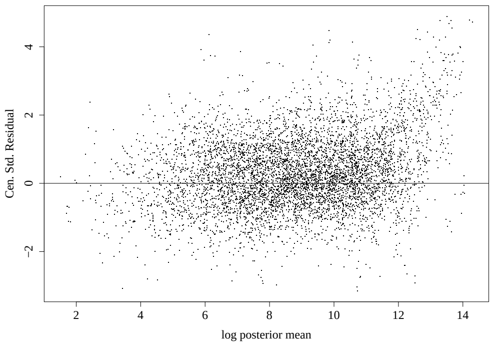
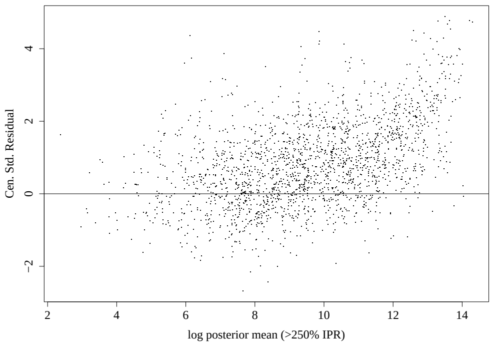


Figure 6: Census: Standardized residual vs. log posterior mean (>250% IPR)


Cen. Std. Residual
log posterior mean (>250% IPR)