A Model for the County-Level Estimation of Insurance Coverage by Demographic Groups

Robin Fisher[*] and Mark Bauder[†]
Small Area Methods Branch
Data Integration Division
U.S. Census Bureau

September 13, 2006

1 Introduction

This paper describes technical details of a model to estimate the number of people without health insurance for small areas. These models were developed for a study to assess the feasibility of producing estimates suitable for the purposes of the Centers for Disease Control and Prevention's (CDC's) National Breast and Cervical Cancer Early Detection Program (NBCCEDP). For NBCCEDP the primary interest is in estimates of the numbers of women without health insurance within particular age groups and income groups defined by Income to Poverty Ratio (IPR). We produce estimates of the number uninsured by county, age group, sex, and IPR category. For a more detailed discussion of the project, see U.S. Census Bureau (2006).

The group eligible for screening differs by state and program. All are females without insurance coverage. Eligible age groups are 18 to 64 years of age, 40 to 64 years of age, and 50 to 64 years of age. Eligible income categories are those with IPRs of between 0 and 200% or between 0 and 250% of the Federal Poverty Level. Other subsets of the population are of less interest in this study. However, because we expect them to be correlated

[*] U.S. Department of the Treasury
[†] U.S. Census Bureau

with the sets of interest, we include them in the model, and obtain estimates for them.

Our approach is to model the number of people without health insurance by demographic subgroups that cover the entire population. In particular, we estimate the number uninsured, N_{UI}, for groups defined by county, sex, age group, and IPR category. We actually model the number insured, N_{IC}.

We use demographic estimates of the number of people in each of the county by age by sex groups, and treat them as known. We estimate the number insured by estimating two proportions: $p_{IPR,h,i,j,k}$, which is the proportion of those in county h, in age group i of sex j who are in IPR category k; and $p_{IC,h,i,j,k}$, the proportion with health insurance in county/age/sex/IPR category h, i, j, k.

We have several data sources. First, we have the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) direct estimates of proportions in the IPR categories, $\tilde{p}_{IPR,h,i,j,k}$, and of proportions insured by county/sex/age/IPR, $\tilde{p}_{IC,h,i,j,k}$. We have Food Stamp participation rates by county, fs_h. We have the Medicaid Eligibles File, which is a list of enrollees. From this file, we calculate proportions of the population that are enrolled in Medicaid by county, age, and sex. In the future, we will have American Community Survey (ACS) direct estimates of IPR membership by age and sex, but they are not available yet. The predictors for the models are described in the following sections.

Fisher and Riesz (2006) describe a model used for similar estimates at the state level. In that model, the data are all conditioned on the IPR categories and insurance coverage is conditioned on the numbers in the IPR categories, so that the number in the IPR category and the proportion insured in the IPR category are estimated jointly. For county-level estimates, this is no longer practical. Instead, we use a pair of models, one for IPR and one for IC, which seem to yield good results and are easier to implement. They are similar to the models presented in Ghosh *et al.* (1998), but differ in the way the survey data are modeled.

In contrast to the assumption in the model for state estimates, we assume here that the proportions in the IPR categories and the proportions insured within the IPR categories are independent, conditioned on the predictors and the various parameters. This assumption was made to facilitate model development for the feasibility study, and will be challenged in future work. We describe a method for challenging the conditional independence assumption in the Model-Checking section below.

The paper proceeds as follows. Section 2 describes the model and data for the proportions in the IPR categories. Section 3 describes the model and data for insurance coverage. Section 4 describes model-checking methods. Section 5 contains results from preliminary runs of the model. Section 6 describes future research.

2 Income to Poverty Ratio Model and Data

2.1 Model

The model for proportions in the IPR categories assumes that those proportions follow a multiple-category logistic regression model, and that the CPS direct estimates of those proportions are independent, normal, and unbiased. More precisely, we have the following.

$$\tilde{p}_{IPR,h,i,j,k} | p_{IPR,h,i,j,k} \sim N(p_{IPR,h,i,j,k}, V_{IPR,\epsilon,h,i,j,k})$$

where $\tilde{p}_{IPR,h,i,j,k}$ is the CPS direct estimate of the proportion of those in the h, i, j county/sex/age domain who are in the k^{th} IPR category, $V_{IPR,\epsilon,h,i,j,k}$ is sampling variance, and

$$p_{IPR,h,i,j,k} = \frac{\exp(\phi_{IPR,h,i,j,k})}{\sum_s \exp(\phi_{IPR,h,i,j,s})}$$

where

$$\phi_{IPR,h,i,j,k} \quad = \quad X_{IPR,h,i,j,k}\eta_{IPR} + u_{IPR,h,i,j,k} \; .$$

Here, $X_{IPR,h,i,j,k}$ is a vector of covariates for county h, sex i, age group j, and IPR category k. The $u_{IPR,h,i,j,k}$ are random effects, or model error, and are assumed to be independent with

$$u_{IPR,h,i,j,k} \sim N(0, V_{IPR,u}).$$

The matrix, X, includes indicators for sex/age categories, and county or higher level variables, including administrative records data and region. We can rewrite the expression for ϕ as

$$\begin{aligned} \phi_{IPR,h,i,j,k} \quad = \quad & \alpha_{IPR,k} + \beta_{ipr,i,k} + \gamma_{IPR,j,k} + (\beta\gamma)_{IPR,i,j,k} + Z_{IPR,h,i,j,k}\xi_{IPR} \\ & + u_{IPR,h,i,j,k}. \end{aligned}$$

Here, α, β, and γ represent main effects in the Generalized Linear Model (GLIM), $(\beta\gamma)_{IPR}$ is the sex/age interaction, and Z represents the additional covariate information. Note that the main effects and interactions are equal to zero when any of their indices are 1. These are the *corner-point restrictions* (*cf* Ghosh, *et al.* (1998)).

For the CPS ASEC sampling error, $V_{IPR,\epsilon,h,i,j,k}$, we use the Generalized Variance Function (GVF) for the CPS ASEC. The GVF for a proportion, p, in a population of size N is given by

$$V_{gfv} = b\frac{p(1-p)}{N}$$

where b is the Generalized Variance Function parameter for a rate in the CPS ASEC (U.S. Census Bureau, 2005) multiplied by a factor to account for the fact that the CPS ASEC direct estimate in our data is a three-year average. For the year 2000, the parameter is 1249, and the correlation between adjacent single-year averages is 0.3. We can show it follows that, in our situation, $b = 574.54$. In the future, expect to include parameters to be estimated in the variance function. This could take into account possible weaknesses in the GVF, which was formulated for higher level estimates. Estimates at higher levels may have different properties. For example, the larger areas' estimates may be more affected by the population controls, where survey aggregates are ratio-adjusted to agree with other estimates derived from the decennial census and administrative records. We also put an upper bound of 0.25 on the variance, because that is the maximum variance for a variable that must be between 0 and 1. For numerical reasons, the variance is bounded below at 0.0001. Thus the sampling variance is

$$V_{IPR,\epsilon,h,i,j,k} = \left\{ \begin{array}{rcl} 0.25 & : & 0.25 < V^*_{IPR,\epsilon,h,i,j,k} \\ V^*_{IPR,\epsilon,h,i,j,k} & : & 0.0001 < V^*_{IPR,\epsilon,h,i,j,k} \leq 0.25 \\ 0.0001 & : & o.w. \end{array} \right.$$

where

$$V^*_{IPR,\epsilon,h,i,j,k} = v_{IPR,\epsilon} * 574.54\ \frac{p_{IPR,h,i,j,k}(1-p_{IPR,h,i,j,k})}{N_{h,i,j,k}}.$$

The normality assumption is clearly not strictly met, since the proportions in the IPR categories must sum to 1.0 and must all be in the interval $[0,1]$ if they are formed by dividing the direct estimate of the number in the (h,i,j,k) cell by the directly-estimated number in the $(h,i,j,)$ cell. It is not clear that this is the best choice, since the latter has a high variance itself.

The IPR categories and age categories are defined in Tables 1 and 2, respectively.

Table 1: IPR Group Definitions

IPR Group	IPR Range
1	0 to 200% FPL
2	200 to 250% FPL
3	more than 250% FPL

Table 2: Age Group Definitions

Age Group	Ages
1	0 to 17
2	18 to 39
3	40 to 49
4	50 to 64
5	65 or more

We use flat prior distributions for all of the non-zero regression parameters. For the variance of $u_{IPR,h,i,j,k}$, we use an inverted gamma prior, $v_{IPR,u} \sim I\Gamma(0.1, 1.0)$.

2.2 Data

The predictors for the IPR section of the model, besides the effects α, β, γ, and their interactions, are the elements of the matrix Z, which are described here.

Log Food Stamp Participation by IPR category Two variables, $m = 1, 2$,

$$LOGFS_{h,i,j,k,m} = \log(FS_i/pop_i)I(k = m),$$

where $I(.) = 1$ if its argument is true, 0 otherwise.

Logit Exemptions Tax data are available in four categories for each county, defined as follows:

- $FTAX_{h,1,1}$ is the proportion of child exemptions in families with $IPR \leq 2.0$;
- $FTAX_{h,1,2}$ is the proportion of child exemptions in families with $IPR > 2.0$;
- $FTAX_{h,2,1}$ is the proportion of non-child exemptions in families with $IPR \leq 2.0$; and
- $FTAX_{h,2,2}$ is the proportion of non-child exemptions in families with $IPR > 2.0$.

Tax data are represented in our model as

$$LTAX_{h,i,j,k,m} = \log\left(\frac{FTAX_{h,m,2}}{FTAX_{h,m,1}}\right) I(k=m),$$

where $m = 1$ if $j = 1$, and $m = 2$ otherwise.

Logit Nonfilers The nonfiler rate is the proportion of people not among tax exemptions. We use the logit of this rate.

$$LNF_{h,i,j,k,m} = \log\left(\frac{\sum_{j=2}^{K} pop_{h,+,j} - FTAX_{h,m,2}}{pop_{h,+,1} - FTAX_{h,m,1}}\right) I(k=m),$$

Mean Log IPR This is the county mean of persons' log IPR from the IRS data;

$$LIPR_{h,i,j,k,m} = \left(\sum_{exemption\ r \in D_h} \log\left(\frac{FI_r}{FPL_r}\right)\right) I(k=m)$$

FI is family income, FPL is the Federal Poverty Level, and D_h is the set of indices corresponding to county h.

Variance Log IPR This is the variance of persons' log IPR.

South This is an indicator of the South Census region, multiplied by the indicator for IPR category.

Medicaid The logit of the Medicaid rate for the county/sex/age category.

Census IPR The Census 2000 proportions in the IPR categories within the county/sex/age domains was tabulated and transformed to the logistic scale.

Note that data from the American Community Survey are not incorporated into this model yet. Future versions of the model will include them.

3 Insurance Coverage Model and Data

3.1 Model

The model for proportion with insurance coverage within a county/sex/age/IPR category is similar to the model for IPR proportions. The model assumes that the proportion insured follows a logistic regression model and that the CPS ASEC direct estimates are independent, with

$$\tilde{p}_{IC,h,i,j,k} | p_{IC,h,i,j,k} \sim N(p_{IC,h,i,j,k}, V_{IC,\epsilon,h,i,j,k})$$

Here, $\tilde{p}_{IC,h,i,j,k}$ is the CPS direct estimate of the proportion of those in the h, i, j, k county/sex/age/IPR domain who are insured, $V_{IC,\epsilon,h,i,j,k}$ is sampling variance. Then

$$p_{IC,h,i,j,k} = \frac{\exp(\phi_{IC,h,i,j,k})}{1 + exp(\phi_{IC,h,i,j,k})}$$

where

$$\phi_{IC,h,i,j,k} \quad = \quad X_{IC,h,i,j,k}\eta_{IC} + u_{IC,h,i,j,k} \ .$$

$X_{IC,h,i,j,k}$ is a vector of covariates for county h, sex i, age group j, and IPR category k. The random effects or model errors follow $u_{IC} \overset{iid}{\sim} N(0, v_{IC,u})$. As before, the model can be equivalently written as a model with fixed effects and additional covariate information

$$\begin{aligned}\phi_{IC,h,i,j,k} \quad = \quad & \alpha_{IC,k} + \beta_{IC,i,k} + \gamma_{IC,j,k} + (\beta\gamma)_{IC,i,j,k} + Z_{IC,h,i,j,k}\xi_{IC} \\ & + u_{IC,h,i,j,k}.\end{aligned}$$

The basic function for the CPS ASEC sampling error is again taken from the Generalized Variance Function (GVF) for the CPS ASEC, so that

$$V_{IC,\epsilon,h,i,j,k} = \begin{cases} 0.25 & : \quad 0.25 < V^*_{IC,\epsilon,h,i,j,k} \\ V^*_{IC,\epsilon,h,i,j,k} & : \quad 0.0001 < V^*_{IC,\epsilon,h,i,j,k} \leq 0.25 \\ 0.0001 & : \quad o.w. \end{cases}$$

where

$$V^*_{IC,\epsilon,h,i,j,k} = 1219.2 \, \frac{p_{IC,h,i,j,k}(1 - p_{IC,h,i,j,k})}{N_{h,i,j,k}}.$$

The factor 1219.2 is the GVF parameter 2652 for health insurance estimates times a factor to account for the fact that the CPS ASEC direct estimate is a three-year average.

We use flat prior distributions for all of the non-zero regression parameters. For priors for the variance parameters, we use $v_{IC,u} \sim I\Gamma(0.5, 1.0)$.

3.2 Data

The direct CPS ASEC estimate for the proportion insured, $\tilde{p}_{h,i,j,k}$ is the ratio of the estimate of the number insured to the estimate of the population in the corresponding county/age/sex/IPR category.

The following variables were used as predictors in the Z_{IC} matrix.

South Indicator for the event that the county is in the South Census region.

$$ISOUTH_{h,i,j,k} = I(h \in S).$$

Here S is the set of indices for counties in the South Census region.

West Indicator for the event that the county is in the West Census region.

$$IWEST_{h,i,j,k} = I(h \in W).$$

Here W is the set of indices for counties in the West Census region.

Medicaid by IPR Medicaid participation rates, transformed to the logistic scale. We obtain files from the Centers for Medicare and Medicaid Services in the Department of Health and Human Services. From these we tabulate the numbers of participants for each county, age group, and sex. Income information is not available on the Medicaid data.

See U.S. Census Bureau(2006) for a detailed description of the data. In this application, we transform the Medicaid participation rates to the logistic scale as follows.

$$MED_{h,i,j} = ln\left(\frac{N_{med,h,i,j}+1}{N_{h,i,j}-N_{med,h,i,j}+1}\right),$$

whenever $N_{h,i,j} \geq N_{med,h,i,j}$. This prevents missing values and preserves the domain for model estimation, though this method should be examined more closely in the future. One possibility is to model it as a response, as in the state model. In a small number of county/sex/age combinations, the inequality did not hold. In these cases,

$$MED_{h,i,j} = ln(N_{med,h,i,j} - 0.5).$$

We allow the effect of Medicaid on the conditional distribution of the insurance coverage rate to change by IPR category, so the model includes interactions between $MED_{h,i,j}$ and the IPR categories,

$$MEDIPR_{h,i,j,k,m} = MED_{h,i,j} * I(k = m).$$

Food Stamps Food Stamp Participation, transformed to the logistic scale. We obtain files for Food Stamp participation from the Food and Nutrition Service at the United States Department of Agriculture. From this we tabulate the number of participants for each county. No demographic information is available from these data. The transformation is similar to that for the Medicaid variables.

$$LFS_{h,i,j,k} = \begin{cases} \log\left(\frac{N_{FS,h}+1}{N_h-N_{FS,h}+1}\right) & : \quad N_h \geq N_{FS,h} \\ \log(N_h - 0.5) & : \quad o.w. \end{cases}$$

For the Medicaid and Food Stamps variables, we add 1 to the numerator and denominator so that the variable is defined even when the reported Medicaid or Food Stamp participation is zero. The motivation is that the resulting value is defined, yet still smaller than if only one person had participated. When the reported participation is larger than or equal to the population, we treat the variable as if only one-half of one person did not participate. These are imperfect solutions to certain problems in the Medicaid data, and in future versions we expect this to be unnecessary.

4 Model-Checking

We rely heavily on Bayesian model-checking methods to examine the fit of these models. Primary among our methods is the use of posterior predictive p-values (PPP-values). For some discrepancy function, designed to examine some aspect of the model fit, $T(Y, \theta)$, where Y is the data and θ is the set of parameters, the posterior predictive p-value is defined as

$$p = P\left[T\left(Y^{(rep)}, \theta^{(rep)}\right) > T\left(Y^{(obs)}, \theta^{(rep)}\right)\right],$$

where the (rep) superscript indicates the variable is drawn from the posterior predictive distribution:

$$\left(Y^{(rep)}, \theta^{(rep)}\right) \sim P(y|\theta)P(\theta|data).$$

A high proportion of posterior predictive p-values close to 0 or 1 indicates that some aspect of the model fits poorly. The function T can be chosen to check some particular piece of the model. Useful choices here, for a data point y and generic parameters θ, are

$$T_1(y, \theta) = y,$$

and

$$T_2(y, \theta) = (y - E(y|\theta))^2,$$

A PPP-value close to 1 for T_1, for example, indicates that replications from the posterior distribution yield values larger than the observed most of the time, which suggests that means might be large or, put another way, that the estimates are biased upward. Similarly, PPP-values for T_2 close to 1 indicate that a variance estimate is biased upward.

One assumption of our model is that the proportions in the IPR categories and the proportion insured are conditionally independent, so we can estimate these as two separate models. There are reasons to expect that this assumption does not hold. For example, it may be the case that a person in the 200% to 250% IPR category in a county with 80% of its population in the lowest IPR category has a different probability of having insurance than a person with similar predictors in a county with more people in the upper IPR categories. A method to check this assumption, at least as a correlation, is to form PPP-values for the discrepancy function

$$T_4(y, \theta) = (y_{IPR} - E(Y_{IPR}|\theta))\,(y_{IC} - E(Y_{IC}|\theta)).$$

Table 3: Average measures of variability for estimates of uninsured rates for females 0 to 200% of FPL.

Age	Mean Standard Deviation	Mean Coefficient of Variation
18-64	0.084	0.30
40-64	0.081	0.36

Table 4: Average measures of variability for estimates of uninsured rates for females 0 to 250% of FPL.

Age	Mean Standard Deviation	Mean Coefficient of Variation
18-64	0.070	0.28
40-64	0.067	0.33

We have not performed this check yet because of the difficulty of implementation. We have plans to do this in future work.

5 Results

Tables 3 and 4 contain average posterior standard deviations and coefficients of variation (CVs) for estimates of uninsured rates for some groups of interest. The CVs are about 0.3, which is reasonable, but with further model development, we expect to get smaller CVs.

The diagnostics for model fit do not appear to show any drastic failures of model fit. For the IPR part of the model, the mean of T_1 above, which assesses model fit with respect to the mean of the direct estimate is 0.50; and the mean for T_2, which assesses fit with respect to variance is 0.59. For the IC part of the model, these means are 0.53 and 0.69 respectively. These are close to the ideal of 0.5, except that it appears that our model tends to overestimate the sampling variance for the direct estimates of both IPR membership and insurance coverage.

6 Future Research

Preliminary results for the IPR part of the model are promising. Results from the IC part of the model have not been as good, perhaps because of the small samples within county/age/sex/IPR groups. We expect to improve on these results with further model development.

One area of research is to investigate alternatives to the variance function or, more generally, the whole sampling error distribution. Domains with direct estimates of 0 or 1 should be given special attention. One approach is to treat them as censored, as in Fisher and Gee(2004), where direct estimates of 0 in a poverty model are treated as if they are censored at some small number and not observed otherwise. Alternatively, we could use a mixture of a continuous model such as the normal or beta distribution and a Bernoulli distribution. The mixing probability would then be modeled on the basis of covariates.

We have other data sources to explore. One data set with potential for IPR proportions is the American Community Survey, which has a large sample size and is collected regularly. Perhaps the most significant data set is the Minimum Data Elements from the CDC. These data have the numbers of people screened under the NBCCEDP. This will allow for another layer of modeling for screening rates. It will be possible to model the screening rates within the county/sex/age/IPR/UI categories, using methods that are directly analogous to those for insurance coverage within county/sex/age/IPR categories. Thus, if screening coverage is correlated with other available variables, we will be able to use that information to improve the small area estimates, detect characteristics of domains which have low screening rates, and improve the estimates of proportions in the IPR categories.

We need to work to establish consistency with other estimates at higher levels, such as the county-level estimates (Fisher and Turner, 2004) or the state-level estimates in the companion paper. We need to establish whether the aggregation of the small domains to the higher levels yields the better estimates, or if they should be controlled somehow to the higher level estimates. Large inconsistencies between the two should be investigated.

7 References

Fisher, R., and Gee, G. (2004), "Errors-in-Variables County Poverty and Income Models", SAIPE Technical Paper, available from
http://www.census.gov/hhes/www/saipe/asapaper/FisherGee2004asa.pdf

Fisher, R., and Turner, J.,(2004) "Small Area Estimation of Health Insurance Coverage From the Current Population Survey's Annual Social and Economic Supplement and the Survey of Income and Program Participation", SAIPE Technical Paper, available from
http://www.census.gov/hhes/www/saipe/asapaper/FisherGee2004asa.pdf

Fisher, R., and Riesz, S.(2006), "A Model for the State-Level Estimation of Insurance Coverage by Demographic Groups."

Ghosh, M.,Natarajan, K., Stroud, T.W.F., and Carlin, B. (1998), "Generalized Linear Models for Small-Area Estimation", *Journal of the American Statistical Association,* 93, 273-282.

U.S. Census Bureau (2005), "Source and Accuracy of Estimates for Income,Poverty,and Health Insurance Coverage in the United States:2004", available from
http://www.census.gov/hhes/www/income/p60_229sa.pdf

U.S. Census Bureau (2006), "Initial Assessment of Small Area Estimation of the Number of Eligible Women for the CDC's NBCCEDP."