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

Robin Fisher*and Steven Riesz†
Small Area Methods Branch
Data Integration Division
U.S. Census Bureau

September 13, 2006

1 Introduction

The Centers for Disease Control and Prevention (CDC) provide screening services for breast and cervical cancer to low-income, uninsured, and underserved women through the National Breast and Cervical Cancer Early Detection Program (NBCCEDP). They are interested in calculating participation rates for states and counties by demographic subgroups. The U.S. Census Bureau's Small Area Health Insurance Estimates (SAHIE) program is engaged in research regarding the feasibility of producing the denominators for these rates. In this paper we discuss the estimation of uninsured rates for low-income women by age, race and whether they are Hispanic. For a general discussion of the problem, see U.S. Census Bureau (2006), hereafter referred to as "the report".

The problem is to estimate the number of uninsured by categories defined by age, race, sex, and Hispanic origin (ARSH) and by categories defined by the ratios of family income to Federal Poverty Levels (FPLs), which we call Income to Poverty Ratios (IPR). Here, these estimates are formed for every state; numbers of eligibles are formed by aggregating the numbers in the

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

categories. The smallest unit, defined by the state/ARSH/IPR crossclassification, will be called a *domain*.

That is, we want to estimate the number of insured, namely $N_{IC,h,i,j,m,k}$ for state h, sex i, age j, race/ethnicity category m, IPR category k. We decompose this into

$$N_{IC,h,i,j,m,k} = p_{IC,h,i,j,m,k} N_{h,i,j,m,k},$$

where $p_{IC,h,i,j,m,k}$ is the proportion insured in the same state/ARSH/IPR, and $N_{h,i,j,m,k}$ is the total number of people in the state/ARSH/IPR. We will similarly decompose this number.

$$N_{h,i,j,m,k} = p_{IPR,h,i,j,m,k} N_{h,i,j,m,+},$$

where $p_{IPR,h,i,j,m,}$ is the vector of proportions in the K IPR categories.

We will generally denote sub-matrices this way when we fix one or more indices. The free indices will be left blank or replaced with a range, so $p_{IPR,h,i,1:3,m,}$ denotes the sub-array formed from the p_{IPR} array by fixing the first index at h, the second at i, restricting the third index to the range $1, 2, 3$, fixing the fourth at m, and allowing the fifth subscript to be free over its range. We denote the sum over an index by replacing the index with a '+' sign, so $N_{h,i,j,m,+}$ is the total number of people in the state/ARSH category h, i, j, m. There are H states, I sexes, J age groups, M race/ethnicity groups, and K IPR groups.

We have several data sources, as described in the report. First, we have the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) direct estimates of proportions in the IPR categories, $\tilde{p}_{IPR,ASEC,h,i,j,m,k}$, and all h, i, j, m, k, proportions insured by IPR and ARSH from the CPS ASEC, $\tilde{p}_{IC,h,i,j,m,k}$. We will have all H, I, J, M, K American Community Survey (ACS) direct estimates of IPR membership by ARSH (though this is not yet available). The tax data are available in four categories for each state, defined as follows:

- $ftax_{h,1,1}$ is the ratio of the number of child exemptions in families with $IPR \leq 2.0$ to the number of childern,

  $$ftax_{h,1,1} = \frac{tax_{h,1,1}}{N_{h,+,1,+,+}},$$

  where $tax_{h,1,1}$ is the number of child exemptions in state h in families with $IPR \leq 2.0$;

- $ftax_{h,1,2}$ is the ratio of the number of child exemptions in families with $IPR > 2.0$ to the number of children,

$$ftax_{h,1,2} = \frac{tax_{h,1,2}}{N_{h,+,1,+,+}};$$

- $ftax_{h,2,1}$ is the ratio of the number of adult exemptions in families with $IPR \leq 2.0$ to the number of adults,

$$ftax_{h,2,1} = \frac{tax_{h,2,1}}{N_{h,+,2:5,+,+}};$$

and

- $ftax_{h,2,2}$ is the ratio of the number of adult exemptions in families with $IPR > 2.0$ to the number of adults,

$$ftax_{h,2,2} = \frac{tax_{h,2,2}}{N_{h,+,2:5,+,+}}.$$

Food stamp participation is available as a number of enrollees by state, which is converted to a proportion of the population, yielding fs_h. Finally, we have the Medicaid Eligibles File from the Centers for Medicare and Medicaid Services (CMS) Medicaid Statistical Information System, which is a list of enrollees, and these are converted to proportions of the population enrolled by state, age, and sex, $fmed_{h,i,j}$.

Note that the eligibility requirements are defined as multiples of a person's family's poverty threshold. This threshold is not defined for every person in the population, for example, foster children. The set of those for whom it is defined is the *poverty universe*. The demographic population estimates used to construct the numbers in the eligibles group represent the total population in the relevant ARSH groups, and is not restricted to the poverty universe. This is a minor deficiency, which will be corrected in later versions of these estimates.

2 Model

We describe the model with a hierarchical set of assumptions about the conditional distribution of various parameters, the "true" IPR and proportions

insured, and the observed data. First, consider the proportions in the IPR categories. This is modeled as a multiple-category logistic regression.

$$\begin{aligned} \ln(\phi_{IPR,h,i,j,m,k}) &= \alpha_{IPR,k} + \beta_{IPR,i,k} + \gamma_{IPR,j,k} + \delta_{IPR,m,k} + u_{IPR,h,i,j,m,k}, \\ p_{IPR,h,i,j,m,k} &= \frac{\phi_{IPR,h,i,j,m,k}}{\sum_k \phi_{IPR,h,i,j,m,k}}. \end{aligned}$$

Some of these parameters are set equal to zero for identifiability. In particular,

$$\alpha_{ipr,1} = \beta_{ipr,1,k} = \beta_{ipr,i,1} = \gamma_{ipr,1,k} = \gamma_{ipr,j,1} = \delta_{ipr,1,k} = \delta_{ipr,m,1} = 0.$$

These are the so-called *corner point restrictions* (see p.276, Ghosh, *et al.* (1998)). These restrictions also have the effect of making the lowest IPR category the reference category within each ARSH group. The priors for all of the non-zero regression coefficients are $N(0, 10^5)$. The random effect, $u_{IPR} \sim N(0, \frac{1}{\tau_{u,IPR}})$, and the prior for its precision is given by $\tau_{u,IPR} \sim \Gamma(0.001, 0.001)$, where $\Gamma(a, b)$ is the gamma distribution with mean a/b and variance a/b^2.

In this model these factors are defined as follows.

- $\alpha_{IPR,k}$ is the IPR main effect for IPR group k
- $\beta_{IPR,i,k}$ is the effect of Sex group i within IPR group k; The only value of i for which $\beta_{IPR,i,k} \neq 0$ is 2, so we may refer to this as the effect of being female within IPR group k.
- $\gamma_{IPR,j,k}$ is the effect of Age group j within IPR group k
- $\delta_{IPR,m,k}$ is the effect of Race/Hispanic group m within IPR group k.

The CPS ASEC data, conditioned on p_{IPR}, is unbiased.

$$\tilde{p}_{IPR,ASEC,h,i,j,m,k} \sim N(p_{IPR,h,i,j,m,k}, (\tau_{IPR,ASEC,h,i,j,m,k})^{-1}),$$

where

$$\tau_{ipr,h,i,j,m,k} = \frac{N_{h,i,j,m,+}}{(0.46)(1249)max((p_{ipr,h,i,j,m,k})(1 - p_{ipr,h,i,j,m,k}), 0.005)},$$

where 1249 is a factor that comes from estimating the variance of a survey estimator with a generalized variance function, and 0.46 is a factor that

reduces the variance of the survey estimator since our survey estimate is actually the average of 3 years of survey estimates that are not independent. (U.S. Census Bureau (2005))

The tax data are not unbiased for the proportions in the IPR categories. Also, their expectations depend on the population-weighted averages of the proportion in the corresponding IPR categories in the various ARSH groups. We have

$$
\begin{aligned}
ftax_{h,1,1} &\sim N\left(\frac{\sum_{i,m} p_{tax,1}p_{IPR,h,i,1,m,1}N_{h,i,1,m,+}}{N_{h,+,1,+,+}}, \tau_{tax}^{-1}\right), \\
ftax_{h,1,2} &\sim N\left(\frac{\sum_{i,m,k=2,3} p_{tax,k}p_{IPR,h,i,1,m,k}N_{h,i,1,m,+}}{N_{h,+,1,+,+}}, \tau_{tax}^{-1}\right), \\
ftax_{h,2,1} &\sim N\left(\frac{\sum_{i,m,j=2,\ldots,5} p_{tax,1}p_{IPR,h,i,j,m,1}N_{h,i,j,m,+}}{\sum_{j=2,\ldots,5} N_{h,+,j,+,+}}, \tau_{tax}^{-1}\right), \\
ftax_{h,2,2} &\sim N\left(\frac{\sum_{i,m,j=2,\ldots,5,k=2,3} p_{tax,k}p_{IPR,h,i,j,m,k}N_{h,i,j,m,+}}{\sum_{j=2,\ldots,5} N_{h,+,j,+,+}}, \tau_{tax}^{-1}\right).
\end{aligned}
$$

Thus, values of $p_{tax,k} = 1$ imply that the proportion of exemptions in domain h, i, j, m, k is the actual proportion of people in IPR group. In IPR groups where many individuals fail to file, this number should be smaller than 1; it may be higher, if there is a systematic tendency, for example, for people to over- or under-report their incomes, so they appear in a different category. Future versions of the model may incorporate correlated error terms to explicitly model this possibility.

We have modeled the proportion of exemptions in the state/ARSH/IPR categories as though they depend only on the IPR category through the parameters $p_{tax,k}$. There are reasons to be skeptical of this assumption; one example is that different states have different tax policies. Notably, Alaska has traditionally given a yearly stipend to residents, which gives them an incentive to file, so we might expect the filing rate to be higher there. We also note that the obvious interpretation of $p_{tax,k}$ is as a filing rate, and we do constrain it to the unit interval. The precision, τ_{tax} is constant across all

state/ARSH/IPR categories; this assumption will be examined in the model fitting and checking phase.

There is no ARSH information in the food stamp participation data; all we have are the number of participants by state, and we convert that to a rate by dividing by the population. This is modeled as a population-weighted average of proportions of the population in the IPR categories.

$$fs_h \sim N(\frac{\sum_{i,j,m,k} p_{fs,k} p_{IPR,h,i,j,m,k} N_{h,i,j,m,+}}{N_{h,+,+,+,+}}, \tau_{fs}^{-1}).$$

The parameter $p_{fs,k}$ may be interpreted as a food stamp participation rate among those eligible. The set of eligible people should in fact be a subset of the people with family incomes $\leq 200\%$ of the FPL. We set $p_{fs,k} \sim U(0,1)$, where $U(a,b)$ denotes the uniform distribution in the interval (a,b), and anticipate that the posterior distribution of $p_{fs,2}$ and $p_{fs,3}$ have most of their masses close to 0. The precision has the following prior distribution: $tau_{fs} \sim \Gamma(0.001, 0.001)$.

Within a given state/ARSH/IPR category, we model the proportion insured (we use the subscript IC for insurance coverage) as

$$\begin{aligned} logit(p_{IC,h,i,j,m,k}) &\sim N(\mu_{IC,h,i,j,m,k}, v_{IC}), \\ \mu_{IC,h,i,j,m,k} &= a + \alpha_{IC,k} + \beta_{IC,i} + \gamma_{IC,j} + \delta_{IC,m} + u_{IC,h,i,j,m,k}. \end{aligned}$$

Here, the logistic regression effects are defined as follows:

- $\alpha_{IC,k}$ is the main effect of IPR category k
- $\beta_{IC,i}$ is the main effect of sex category i
- $\gamma_{IC,j}$ is the main effect of age category j
- $\delta_{IC,m}$ is the main effect of race/Hispanic category m.

The random effect is $u_{IC,h,i,j,m,k} \sim N(0, \tau_{u,IC}^{-1})$.

The CPS ASEC direct measurement of the proportion insured is modeled as

$$\tilde{p}_{IC,h,i,j,m,k} \sim N(p_{IC,h,i,j,m,k}, v_{IC,ASEC,h,i,j,m,k}),$$

where $v_{IC,ASEC,h,i,j,m,k} = 1/(\tau_{IC,ASEC,h,i,j,m,k})$,

$$\tau_{IC,h,i,j,m,k} = \frac{max(\tilde{N}_{h,i,j,m,k}, 1)}{((0.46)(2652)max((p_{IC,h,i,j,m,k})(1 - p_{IC,h,i,j,m,k}), 0.005))},$$

$\tilde{N}_{h,i,j,m,k}$ is the CPS ASEC estimate of the number of people in the state/ARSH/IPR, 2652 is a factor that comes from estimating the variance of a survey estimator with a generalized variance function, and 0.46 is a factor that reduces the variance of the survey estimator since our survey estimate is actually the average of 3 years of survey estimates that are not independent. (U.S. Census Bureau, 2005). The prior distributions for the regression coefficients are *iid* $N(0, 10^5)$ and $\tau_{u,IC} \sim \Gamma(0.001, 0.001)$.

The only administrative records covariate available for insurance coverage is the Medicaid data, which we model similarly to the tax data.

$$\begin{aligned} \mu_{med,h,i,j} &= E(fmed_{h,i,j}|pa(fmed_{h,i,j})) \\ &= \frac{\sum_{m,k} p_{med,k} p_{IC,h,i,j,m,k} p_{IPR,h,i,j,m,k} N_{h,i,j,m,+}}{N_{h,i,j,+,+}}, \end{aligned}$$

and

$$fmed_{h,i,j} \sim N(\mu_{med,h,i,j}, \tau^{-1}_{fmed}).$$

The prior distributions are given by $p_{med,k} \sim Exp(0.5)I(0, 2)$, the exponential distribution with mean 0.5, truncated at 2. This allows for the over-reporting of medicaid while keeping the majority of the probability mass less than one. Finally, $\tau_{fmed} \sim \Gamma(0.001, 0.001)$.

3 Model-Checking

We rely heavily on Bayesian model-checking methods to examine the fit of these models, primarily the posterior predictive p-values (PPP-values). See, for example, Gelfand (1995). 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

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

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

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

Posterior predictive p-values represent the probability, under the posterior distribution, that the value of the T function is larger than that actually observed, so values around 0.5 would be expected. Posterior predictive p-values close to 0 or 1 indicate 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

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

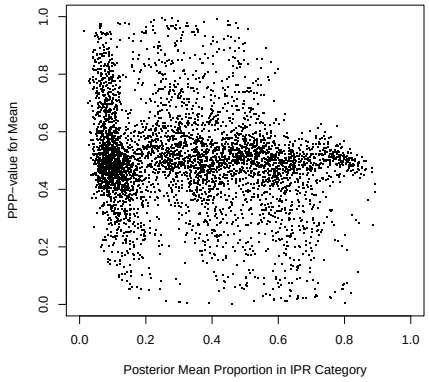
Thus, if the PPP-value for T_1 were close to 1, that would be an indication that the model tends to overestimate means, so that replicated observations are larger than that observed with high probability. Similarly, large values for T_2 indicate that the squared difference from the mean is larger than that expected with high probability.

4 Results

4.1 Proportions in The Income Categories.

The PPP-values associated with T_1 and T_2 were examined. Figures 1 and 2 show plots of the PPP-values for T_1 and T_2, respectively. Figures 3 through 6 show boxplots of the PPP-values associated with T_1 for the sexes, age groups, race/ethnicity groups, and IPR categories. In none of those cases do we see an indication that the model fits poorly in terms of the central tendency of distribution, overall. The overall PPP-value overall for T_1 is 0.50, which is not an indication that the proportions in the IPR categories are over- or under-predicted on average. The overall PPP-value associated with T_2 is approximately 0.55. This value is moderate, so is not an indication that variances are over- or under-predicted on average. Figures 7 through 10 show boxplots of the PPP-values for T_2 for the sexes, age groups, race/ethnicity

Figure 1: Plot of the Posterior Predictive P-value for the Mean against the Posterior Mean Proportion for IPR for all Domains. There is no trend visible in this plot, so there is no evidence in it that the model fits poorly with respect to the means of the proportions in the Income to Poverty Ratio (IPR) categories.



groups, and IPR categories. They show the PPP-value is consistent across the groupings, so there is no evidence that the variance is badly over- or under-estimated for any of those groups.

The mean coefficient of variation (CV) for all proportions in the IPR categories is 0.16. Table 1 shows it for the individual IPR categories.

The mean CV for the 200% to 250% category is higher because the category itself is smaller than the other two, so the denominator of the CV is somewhat smaller. The estimands of interest involve either the first or the combination of the first two IPR categories, so the relevant mean CVs are closer to 0.11.

The mean of the ratio of the posterior variance to the sampling error variance for the IPR ratios is 0.44. That is, the variance is reduced by more than half.

Figure 2: Plot of the Posterior Predictive P-value for the Variance against the Posterior Mean Proportion for IPR for all Domains. There is no trend visible in this plot, so there is no evidence in it that the model fits poorly with respect to the variances for proportions in the Income to Poverty Ratio (IPR) categories.

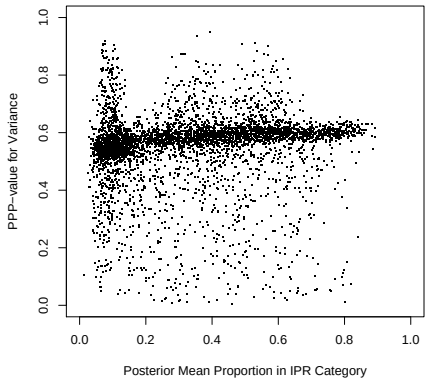


Figure 3: Box Plots of PPP-values for Mean (T_1), for the Proportion in the IPR Group, by Sex Group

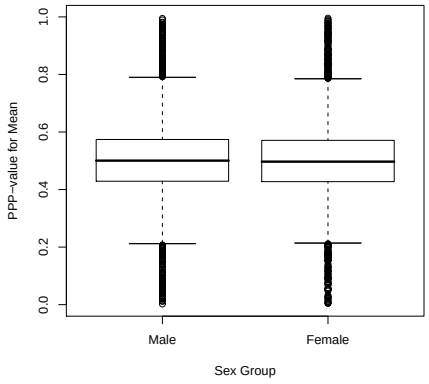


Figure 4: Box Plots of PPP-values for Mean (T_1), for the Proportion in the IPR Group, by Age Group

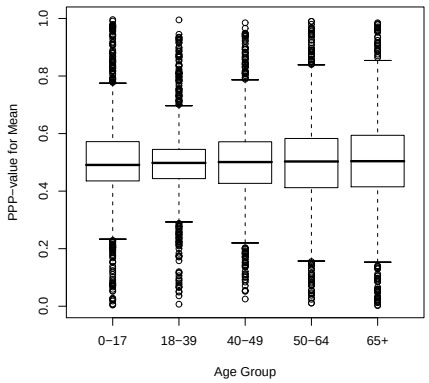


Figure 5: Box Plots of PPP-values for Mean (T_1), for the Proportion in the IPR Group, by Race/Hispanic Group

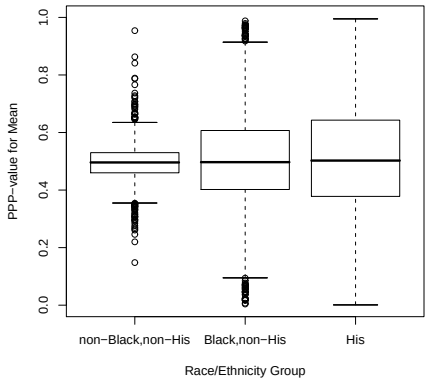


Figure 6: Box Plots of PPP-values for Mean (T_1), for the Proportion in the IPR Group, by IPR Group

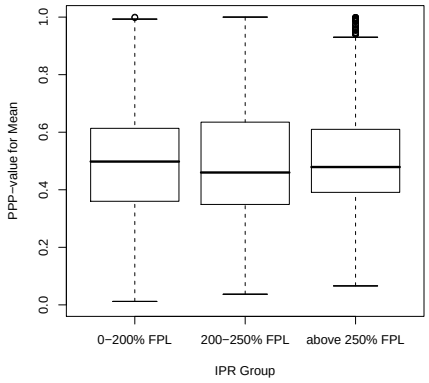


Figure 7: Box Plots of PPP-values for Variance (T_2), for the Proportion in the IPR Group, by Sex Group

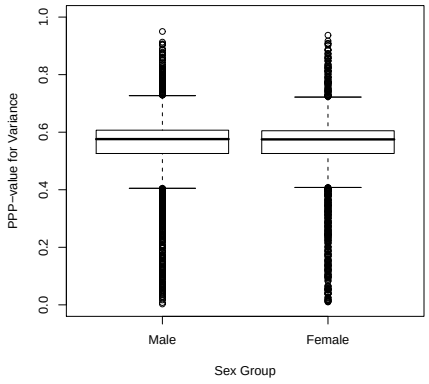


Figure 8: Box Plots of PPP-values for Variance (T_2), for the Proportion in the IPR Group, by Age Group

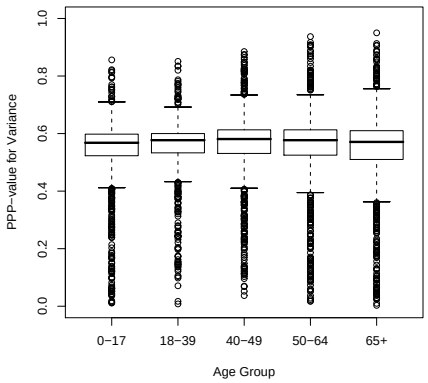


Figure 9: Box Plots of PPP-values for Variance (T_2), for the Proportion in the IPR Group, by Race/Hispanic Group

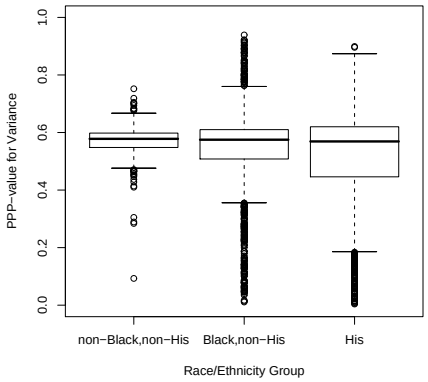


Figure 10: Box Plots of PPP-values for Variance (T_2), for the Proportion in the IPR Group, by IPR Group

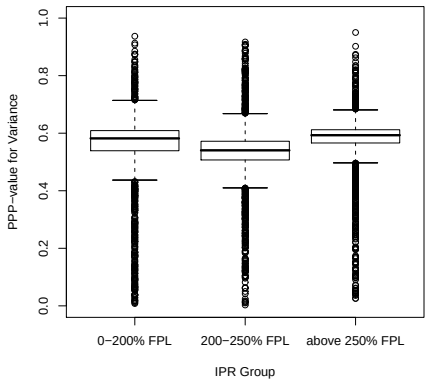


Table 1: Mean Coefficients of Variation (CVs) for Proportions by IPR Categories

Subgroups	Mean CV of p_{IPR}
IPR: 0 to 200% FPL	0.11
IPR: 200 to 250% FPL	0.27
IPR: above 250% FPL	0.11

4.1.1 Parameter Estimates

The parameter estimates related to IPR are presented in Table 2. These estimates are easier to interpret than in the GLIM of Fisher (2006), for example. We will not exhaustively interpret the parameters, but note that the tax parameters and food stamps parameters are interesting. The tax bias parameter is close to 1.0 for the lowest income category, decreasing for the higher ones. The food stamp participation rate is low for the lowest category (recall that not even all of the people in this category are eligible, so we shouldn't expect this parameter to be close to 1.0), but they get close to zero for the higher categories, as they should, since essentially no one in these categories should be eligible.

4.2 Insurance Coverage

Again, the PPP-values associated with T_1 and T_2 were examined. Overall, the PPP-value for T_1 is 0.50. The PPP-value for T_2 is 0.57. Figures 11 and 12 show plots of the PPP-values for T_1 and T_2 against the posterior mean of the proportion insured. Figures 13 through 16 show boxplots of the PPP-values for T_1 for the sexes, age groups, race/ethnicity groups, and IPR categories. They all seem moderate and well behaved, so there is no evidence for biases. Figures 17 through 20 show boxplots of the PPP-values for the sexes, age groups, race/ethnicity groups, and IPR categories for T_2. We see no evidence of failures in the model for variances, except for the 65+ age group; see Figure 18. Here the variance seems to be somewhat overestimated.

The mean coefficient of variation for all proportions insured is 0.07; Table 3 shows it for the individual IPR categories.

The parameter estimates for insurance coverage are presented in Table 4. Here again we will concentrate on the bias parameters. The Medicaid variables behave as one might expect; on average, something less than half of the people in the lowest IPR group who are insured participate in Medicaid. In the upper two groups, there should be very few Medicaid eligibles, and that is as we estimate participation, with $p_{med,2} = 2.3\%$ and $p_{med,3} = 0.27\%$. The mean CV of the uninsured rate is 0.35. The ratio of the posterior variance to the sampling error variance under the model of the uninsured rate is 0.33. This represents a substantial reduction in the variance under the model.

Table 2: Parameter Estimates for Parameters Related to IPR

Parameter	Relevant Quantity	Posterior Mean(SD)
$\alpha_{IPR,2}$	IPR Group 2	$-1.05(5.1 \times 10^{-2})$
$\alpha_{IPR,3}$	IPR Group 3	$0.74(3.3 \times 10^{-2})$
$\beta_{IPR,2,2}$	Female, IPR Group 2	$-0.17(4.1 \times 10^{-2})$
$\beta_{IPR,2,3}$	Female, IPR Group 3	$-0.29(2.5 \times 10^{-2})$
$\gamma_{IPR,2,2}$	Age Group 2, IPR Group 2	$0.27(5.0 \times 10^{-2})$
$\gamma_{IPR,3,2}$	Age Group 3, IPR Group 2	$0.30(5.5 \times 10^{-2})$
$\gamma_{IPR,4,2}$	Age Group 4, IPR Group 2	$0.21(5.7 \times 10^{-2})$
$\gamma_{IPR,5,2}$	Age Group 5, IPR Group 2	$0.038(6.0 \times 10^{-2})$
$\gamma_{IPR,2,3}$	Age Group 2, IPR Group 3	$0.48(3.1 \times 10^{-2})$
$\gamma_{IPR,3,3}$	Age Group 3, IPR Group 3	$1.08(5.1 \times 10^{-2})$
$\gamma_{IPR,4,3}$	Age Group 4, IPR Group 3	$1.00(5.1 \times 10^{-2})$
$\gamma_{IPR,5,3}$	Age Group 5, IPR Group 3	$0.049(5.3 \times 10^{-2})$
$\delta_{IPR,2,2}$	Race/Hispanic Group 2, IPR Group 2	$-0.57(3.9 \times 10^{-2})$
$\delta_{IPR,3,2}$	Race/Hispanic Group 3, IPR Group 2	$-0.48(3.9 \times 10^{-2})$
$\delta_{IPR,2,3}$	Race/Hispanic Group 2, IPR Group 3	$-1.07(3.9 \times 10^{-2})$
$\delta_{IPR,3,3}$	Race/Hispanic Group 3, IPR Group 3	$-1.28(3.3 \times 10^{-2})$
$p_{tax,1}$	Tax Bias for IPR Group 1	$0.99(6.3 \times 10^{-3})$
$p_{tax,2}$	Tax Bias for IPR Group 2	$0.92(8.0 \times 10^{-2})$
$p_{tax,3}$	Tax Bias for IPR Group 3	$0.82(1.9 \times 10^{-2})$
$p_{fs,1}$	Food Stamps Bias for IPR Group 1	$0.18(2.9 \times 10^{-2})$
$p_{fs,2}$	Food Stamps Bias for IPR Group 2	$0.086(7.9 \times 10^{-2})$
$p_{fs,3}$	Food Stamps Bias for IPR Group 3	$0.0063(6.2 \times 10^{-3})$

Table 3: Mean Coefficients of Variation (CVs) for Proportion Insured by IPR Categories

Subgroups	Mean CV of p_{IC}
IPR: 0 to 200% FPL	0.09
IPR: 200 to 250% FPL	0.08
IPR: above 250% FPL	0.04

Figure 11: Plot of the Posterior Predictive P-value for the Mean against the Posterior Mean Proportion for all Domains. There is no trend visible in this plot, so there is no evidence in it that the model fits poorly with respect to the means of the proportions insured.

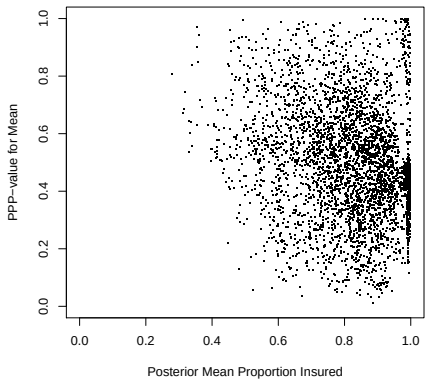


Figure 12: Plot of the Posterior Predictive P-value for the Variance against the Posterior Mean Proportion for all Domains. There is no trend visible in this plot, so there is no evidence in it that the model fits poorly with respect to the variances for proportions insured.

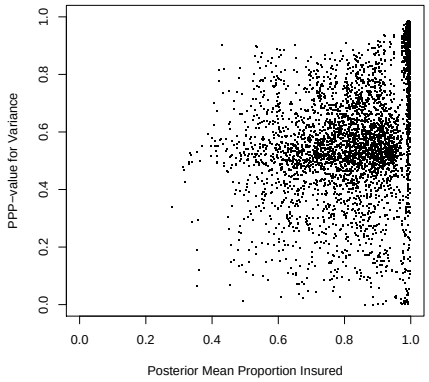


Figure 13: Box Plots of PPP-values for Mean (T_1), for the Proportion Insured, by Sex Group

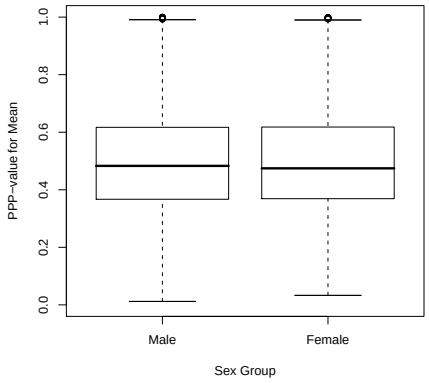


Figure 14: Box Plots of PPP-values for Mean (T_1), for the Proportion Insured, by Age Group

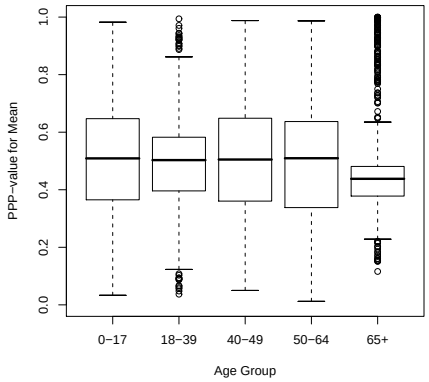


Figure 15: Box Plots of PPP-values for Mean (T_1), for the Proportion Insured, by Race/Hispanic Group

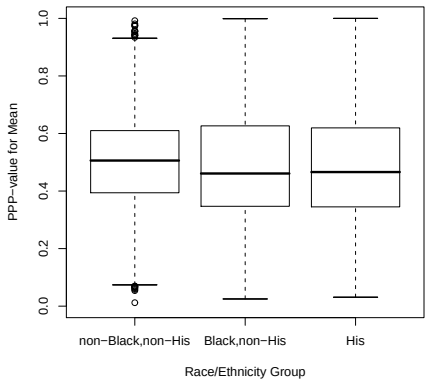


Figure 16: Box Plots of PPP-values for Mean (T_1), for the Proportion Insured, by IPR Group

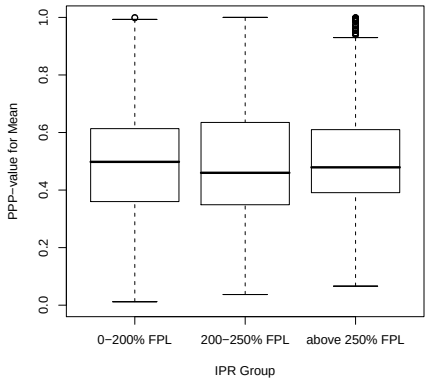


Figure 17: Box Plots of PPP-values for Variance (T_2), for the Proportion Insured, by Sex Group

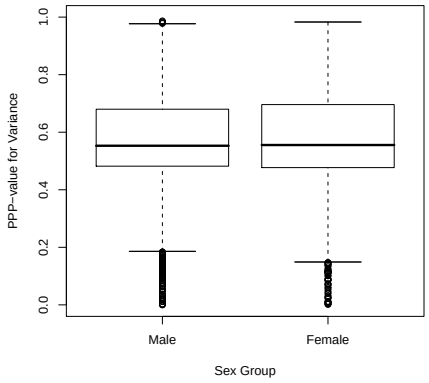


Figure 18: Box Plots of PPP-values for Variance (T_2), for the Proportion Insured, by Age Group

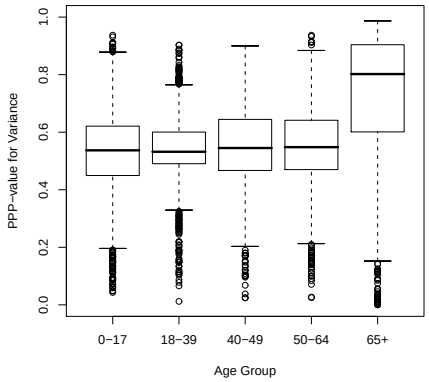


Figure 19: Box Plots of PPP-values for Variance (T_2), for the Proportion Insured, by Race/Hispanic Group

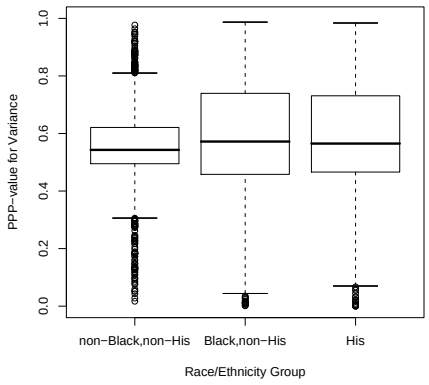


Figure 20: Box Plots of PPP-values for Variance (T_2), for the Proportion Insured, by IPR Group

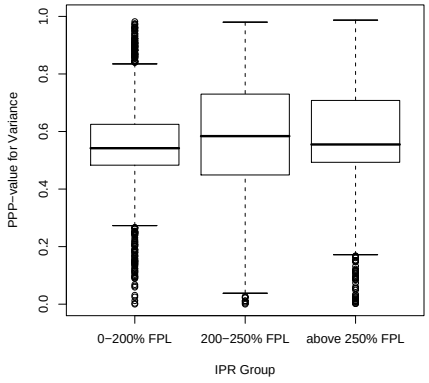


Table 4: . Parameter Estimates

Parameter	Posterior Mean(SD)
intercept	$1.75(3.7 \times 10^{-2})$
$\alpha_{IC,2}$	$0.47(3.6 \times 10^{-2})$
$\alpha_{IC,3}$	$0.74(3.3 \times 10^{-2})$
$\beta_{IC,2}$	$0.20(2.7 \times 10^{-2})$
$\beta_{IC,2}$	$0.20(2.7 \times 10^{-2})$
$\gamma_{IC,2}$	$-1.04(3.8 \times 10^{-2})$
$\gamma_{IC,3}$	$-0.61(4.4 \times 10^{-2})$
$\gamma_{IC,4}$	$-0.48(4.4 \times 10^{-2})$
$\gamma_{IC,5}$	$2.88(6.6 \times 10^{-2})$
$\delta_{IC,2}$	$-0.31(3.4 \times 10^{-2})$
$\delta_{IC,3}$	$-1.0(3.5 \times 10^{-2})$
$p_{med,1}$	$0.44(1.4 \times 10^{-3})$
$p_{med,2}$	$0.023(2.2 \times 10^{-2})$
$p_{med,3}$	$0.0027(2.5 \times 10^{-3})$

4.3 Eligibles

Table 5 presents the CVs for the numbers of low-income eligibles for the NBCCEDP. The CVs reported in this table are a composite of the CVs of the IPR and IC estimates and represent the final CVs. The CVs for the uninsured are in general higher than those for the IPR measures, largely because of domains with very low proportions uninsured. Table 6 presents them for their inverses, which are the CVs of the NBCCEDP coverage rates, assuming the NBCCEDP numbers covered have zero variance.

5 Conclusion

We have produced estimates of the proportions in the IPR categories for the various ARSH groups within states and the proportion uninsured in each ARSH/IPR group. This followed from a model of the relationship between the data sources, in particular, where we model the behavior of each data source, conditioned on the parameters of interest. The form of the model is general and intuitive, and the parameters have straightforward interpreta-

Table 5: Mean Coefficients of Variation (CVs) of the Number of Uninsured

Small Areas	Mean CV
All Domains, i.e. State x Sex x Age x Race/Ethnicity x IPR	0.40
Female, 50 to 64 years old, and IPR $\leq$ 200%, By State x Race/Ethnicity	0.32
Female, 40 to 64 years old, and IPR $\leq$ 200%, By State x Race/Ethnicity	0.22
Female, 18 to 64 years old, and IPR $\leq$ 200%, By State x Race/Ethnicity	0.16

Table 6: Mean Coefficients of Variation (CVs) of the Inverse of the Number of Uninsured

Small Areas	Mean CV
All Domains, i.e. State x Sex x Age x Race/Ethnicity x IPR	0.43
Female, 50 to 64 years old, and IPR $\leq$ 200%, By State x Race/Ethnicity	0.35
Female, 40 to 64 years old, and IPR $\leq$ 200%, By State x Race/Ethnicity	0.23
Female, 18 to 64 years old, and IPR $\leq$ 200%, By State x Race/Ethnicity	0.17

tions. In this case, the models seem to fit well overall. The resulting estimates have substantial reductions in variance relative to the direct estimates.

There is more work to do to improve the estimation methods. Interactions in the predictors may be useful; Fisher (2006) found they have an impact in the Generalized Linear Model for counties. Second, the model for the distributions of the responses should be examined more closely. The ASEC, especially, would likely be modeled better with some distribution besides the normal distribution.

Some of the other simple assumptions in this model should also be examined. The assumption that the proportion of tax exemptions in an IPR category depends only on the proportion of people in the IPR category, through the parameter $p_{tax,k}$, implying conditional independence of, for example, ARSH composition should be examined. Food stamp participation may also depend on ARSH composition or the policies of the particular state. Similar investigations can be made for Medicaid. Finally, we have not considered evidence that citizenship or tenure as a U.S. resident is important; this should be considered in future models.

While these investigations may yield improvements, the results here seem to indicate that the state-level estimates are feasible. The data are available, and substantial reductions in the variance over the direct estimates are possible. Further, estimates of the inverses of the numbers of eligibles, necessary to compute the NBCCEDP coverage rates, are available without a large extra investment or approximations.

6 References

Fisher, R. (2006), "A Model for the County-Level Estimation of Insurance Coverage by Demographic Groups," attached

Gelfand, A. E. (1995), "Model Determination Using Sampling-Based Methods", in *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), 145-161, London: Chapman and Hall

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. (1998) "Generalized Linear Models for Small-Area Estimation", *Jounal 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", attached