Small Area Estimates by Zip Code For the Miami-Dade County Health Insurance Survey

Cynthia Wilson Garvan
Statistician

Weir Hou
Statistical Assistant

R. Paul Duncan
Principal Investigator

Colleen K. Porter
Project Coordinator

July 2003

Commissioned by: Public Health Trust of Miami-Dade County

Prepared by:
The Department of Health Services Administration
University of Florida

Contents

Introduction

Sample-based surveys can provide very sound estimates of numerous characteristics for the populations to which the samples refer. But the estimate and the resulting inference have limited application. An opinion poll based on a national sample may provide a very reliable estimate of the nation's views, but may not be a valid indicator of how Floridians (or some other subset of the nation) might feel.

Nevertheless, there is great interest in using sample-based results to make inferences about various sub-populations of the survey sample. Of particular interest are "small area estimates" derived from survey data – synthetic estimates that are calculated in order to describe certain segments of the survey population.

The primary focus of the Miami-Dade County Health Insurance Survey was estimating the percentage of residents who are uninsured, as well and the percentage of various groups within the community who share that circumstance, with clearly specified confidence intervals. Such estimates were provided in a companion report. (Duncan et al., *The Miami-Dade County Health Insurance Survey,* June 2003).

In addition, however, there is interest in comparable estimates for much smaller segments of the community, in this case various residential areas defined by zip codes. The assumption is that these areas differ from the experience of the larger community, but the sample design does not allow direct estimation for the smaller areas.

To be valuable, small area estimates require two attributes. First, they must be derived by a calculation that is reasonable, logical, and mathematically sound. For example, one might believe that Florida is an exact replication of the entire nation, and simply apply a national estimate to Florida. Alternatively one might believe that because Florida has an older population, people living in our state would be expected to have views that differ from other parts of the nation. In the latter case, a national estimate would be not be applicable to Florida unless it was "adjusted" in some way to reflect the belief that Florida differs from the nation. Any such "adjustment" would have to be based on reasonable, logical methodology that can be described at the level of precise detail necessary for replication.

Second, the derivative estimates must include a means of quantifying our confidence. One of the advantages of survey research is that properly designed and drawn samples provide a means of specifying the level of confidence that can be placed in those estimates, allowing us to state an estimate in terms that include a confidence interval or margin of error. As a result, estimates are commonly expressed in a form such as, "65% of the sample (plus or minus 3 percentage points) agree with the statement that…" and readers can reasonably infer that the actual percentage of all people in the population to which the sample refers who agree with the statement falls between 62% and 68%.

In this report, we provide estimates of the rate of uninsurance experienced by the residents of 76 zip codes in Miami-Dade County. The estimates are derived by a calculation that adjusts the community wide estimate using other characteristics known to be associated with health insurance status. The results are shown in the accompanying table, and technical specification of the methodology is provided in the subsequent pages.

Standard errors are provided for each estimate in the table. The standard error refers to error in estimates resulting from random fluctuations in samples. Smaller standard errors reflect greater reliability in the accuracy with which the sample statistic portrays the population parameter.

Zip Code Estimates from the 2003 Miami-Dade Health Insurance Survey

Area	Zip Code	Percent Uninsured	Approximate std error
A	33015	28.20%	0.0610
A	33054	29.66%	0.0640
A	33055	29.71%	0.0628
A	33056	29.29%	0.0644
A	33127	30.39%	0.0661
A	33128	31.58%	0.0680
A	33129	27.53%	0.0593
A	33130	31.16%	0.0650
A	33131	27.64%	0.0630
A	33132	29.21%	0.0699
A	33136	28.94%	0.0638
A	33137	28.33%	0.0636
A	33138	24.84%	0.0635
A	33142	31.11%	0.0669
A	33147	30.61%	0.0657
A	33150	28.98%	0.0632
A	33161	26.38%	0.0623
A	33162	26.21%	0.0638
A	33167	28.72%	0.0643
A	33168	27.97%	0.0628
A	33169	27.47%	0.0639
A	33179	24.42%	0.0624
A	33180	22.55%	0.0641
A	33181	25.04%	0.0635
B	33109	22.52%	0.1585
B	33139	25.99%	0.1176
B	33140	23.75%	0.1186
B	33141	25.78%	0.1079
B	33154	24.10%	0.1207
B	33160	24.53%	0.1224
C	33010	30.65%	0.0578
C	33011		
C	33012	29.97%	0.0582
C	33013	30.27%	0.0584
C	33014	28.34%	0.0660
C	33016	29.02%	0.0617
C	33018	29.25%	0.0620
C	33122		
C	33125	30.04%	0.0600
C	33126	30.26%	0.0591
C	33135	30.55%	0.0587
C	33144	30.12%	0.0590

Area	Zip Code	Percent Uninsured	Approximate std error
C	33145	29.34%	0.0626
C	33155	27.93%	0.0687
C	33165	28.68%	0.0639
C	33166	26.80%	0.0773
C	33172	29.54%	0.0621
C	33174	29.79%	0.0585
C	33175	28.44%	0.0636
C	33178	26.99%	0.0755
C	33182	29.39%	0.0671
C	33184	29.01%	0.0611
C	33185	27.59%	0.0674
C	33192		
C	33194		
D	33133	19.62%	0.0823
D	33134	20.66%	0.0702
D	33143	19.21%	0.0800
D	33146	18.23%	0.1031
D	33149	18.71%	0.0775
D	33156	17.95%	0.0852
D	33157	18.84%	0.0808
D	33158	17.02%	0.0914
D	33173	20.02%	0.0691
D	33176	18.79%	0.0791
D	33177	21.48%	0.0645
D	33183	20.71%	0.0643
D	33186	20.09%	0.0709
D	33187	20.36%	0.0710
D	33193	21.77%	0.0637
D	33196	21.10%	0.0674
E	33030	35.10%	0.1181
E	33031	29.25%	0.1585
E	33032	34.78%	0.1142
E	33033	36.11%	0.1094
E	33034	35.73%	0.1181
E	33035	29.22%	0.1478
E	33039		
E	33090		
E	33170	32.74%	0.1327
E	33171		
E	33189	32.88%	0.1209
E	33190	33.47%	0.1174

NOTE: Zip codes without estimates are those for which Census 2000 data were not available.

Technical Summary of Methods—
Zip code level estimates

The MDHIS survey yielded person level data that included whether or not a person had health insurance (the outcome variable of primary interest) along with a number of variables such as age, gender, race/ethnicity, education, household size, family income, and geographic location within Miami-Dade county (i.e., the "region" A, B, C, D, E) that are known to be associated with the probability of having insurance. Additionally, information is available from the 2000 U.S. Census about characteristics of Miami-Dade residents living in each geographic area defined by zip code. Our challenge, then, is to produce zip code level estimates that synthesize information available from both the MDHIS survey and 2000 U.S. Census data, using methods that have been validated by other researchers. Toward this end we combined the approaches of Popoff, Judson, and Fadali [Measuring the Number of People Without Health Insurance: A Test of Synthetic Estimates Approach for Small Areas Using SIPP Microdata, Fall 2001] and Ghosh, Kim, and Sinha [Hierarchical Bayesian Models For Small Domain Estimation, in preparation].

Popoff et al. devised a small area estimation approach using synthetic estimation techniques. Using 1996 SIPP (Survey of Income and Program Participation) data for 80,923 individuals, they demonstrated that the characteristics of age, race, gender, and Hispanic origin predicted the proportion of uninsured quite well. They proposed that the proportion of uninsured in a small geographic area could be estimated as follows:

1) Obtain survey data that represents the population as a whole. Estimate the effects of age, gender, race and Hispanic origin on the probability of uninsurance for the population based on the survey data.

2) Divide a small geographic area into domains based on age, gender, race, and Hispanic origin and obtain Census estimates of the numbers of residents in each domain.

3) A synthetic estimate of the proportion of uninsured in each small geographic area is then found by calculating the number of uninsured within each domain defined by age, gender, race, and Hispanic origin (by overlaying estimates derived from population survey); summing the number of uninsured in each domain; and dividing the estimated number of uninsured by the total number of residents living in a small area. Table 1 illustrates the Popoff et al. approach for estimating the number of uninsured individuals in a domain defined as "White non-Hispanic females less than 18 yrs old." A complete illustration of the method would require extending Table 1 for all other domains (e.g., White non-Hispanic males less than 18 yrs old, Hispanic females less than 18 yrs old, Hispanic males less than 18 yrs old, etc.).

Table 1: Illustration of Synthetic Estimation

Small geographic area	**# White non-Hispanic females less than 18 yrs old (from Census data)**	**Estimated proportion of White non-Hispanic females less than 18 yrs old who are uninsured (from survey data)**	**Estimated # of uninsured White non-Hispanic females less than 18 yrs old**
1	$n_{1,W,F,<18}$	$p_{1,W,F,<18}$	$n_{1,W,F,<18} \cdot p_{1,W,F,<18}$
2	$n_{2,W,F,<18}$	$p_{2,W,F,<18}$	$n_{2,W,F,<18} \cdot p_{2,W,F,<18}$
3	$n_{3,W,F,<18}$	$p_{3,W,F,<18}$	$n_{3,W,F,<18} \cdot p_{3,W,F,<18}$
…	…	…	…

In order to estimate the proportion of uninsured in each domain, we used the approach developed by Ghosh et al. They used a hierarchical Bayes modeling procedure to estimate the proportion of individuals without health insurance for domains cross-classified by age, gender, and race/ethnicity. The model is built in stages, hence the name *hierarchical*. As part of the estimation scheme, available covariates at the individual level are incorporated in the model specification to improve the predictive capacity for estimation at the domain level. Ghosh et al. used data provided by the National Center for Health Statistics (NCHS) to formulate a hierarchical Bayes model that provided estimates for proportion of uninsured in cross-classified domains and standard errors for those estimates. The NCHS data set included individual level data for more than 100,000 people and included over 800 covariates. In a covariate selection procedure the variables retained for the final model were family size, education level, and family income. The Markov chain Monte Carlo (MCMC) numerical integration technique employing the Gibbs sampler was used to compute estimates and corresponding standard errors for the NCHS study.

For the MDHIS zip code level estimates we combined the approaches of Popoff et al. and Ghosh et al. The MDHIS zip code level estimates were produced as follows:

1) For each of the MDHIS survey regions (i.e., A, B, C, D, E), domains were defined based on age group (0-18, 19-24, 25- 44, 45-64), gender (M, F), and race/ethnicity (non-Hispanic White, Hispanic, Black, and Other). A total of 160 domains were thus defined (5 regions × 4 age groups × 2 genders × 4 race/ethnicity categories).

2) For each of the 160 domains, cross-classified by age, gender, and race/ethnicity, we applied a hierarchical Bayes modeling procedure to estimate the proportion of individuals without health insurance. We used the MDHIS survey variables of family size, education level (of respondent), and family income as covariates (following the result from the NCHS study). Professor Dalho Kim (Kyungpook National University) wrote specialized FORTRAN software to apply MCMC numerical integration that employed the Gibbs sampler to produce the MDHIS hierarchical Bayes (HB) domain estimates and their corresponding standard errors.

3) A data set was prepared using 2000 U.S. Census zip code level data that included the number of residents in each of the domains cross-classified by age, gender, and race/ethnicity. The definition of each domain is given in Table 2.

4) Within each zip code, the proportion of uninsured was estimated by calculating the number of uninsured in each domain (multiplying the HB domain estimates of proportion of uninsured by the number of individuals in each domain), summing across domains to find the estimated number of uninsured in each zip code, and dividing the number of uninsured in each zip code by the total number of people living in the zip code area.

5) The MDHIS zip code level estimates of proportion of uninsured were then calibrated to match the regional estimate (i.e., A, B, C, D, E) by multiplying each zip code level estimate by a constant coefficient that ensured parity between MDHIS survey region estimates and MDHIS zip code level estimates.

6) An approximate standard error for each zip code level estimate was calculated by using the standard errors for each of the 160 domains. Assuming that proportions are approximately normally distributed, the variance of a zip code level estimate is a function of the standard errors of 32 domain proportions and their covariances. Maximal covariances were assumed in order to produce conservative (i.e., overestimated rather than underestimated) approximate standard errors. The approximated standard errors do not take into consideration error associated with estimation of the 2000 U.S. Census zip code level data.

Table 2: Domain definitions for each zip code

Domain	Definition
1	Non-Hispanic White females 0-18 years of age
2	Non-Hispanic White males 0-18 years of age
3	Hispanic females 0-18 years of age
4	Hispanic males 0-18 years of age
5	Black females 0-18 years of age
6	Black males 0-18 years of age
7	Other females 0-18 years of age
8	Other males 0-18 years of age
9	Non-Hispanic White females 19-24 years of age
10	Non-Hispanic White males 19-24 years of age
11	Hispanic females 19-24 years of age
12	Hispanic males 19-24 years of age
13	Black females 19-24 years of age
14	Black males 19-24 years of age
15	Other females 19-24 years of age
16	Other males 19-24 years of age
17	Non-Hispanic White females 25- 44 years of age
18	Non-Hispanic White males 25- 44 years of age
19	Hispanic females 25- 44 years of age
20	Hispanic males 25- 44 years of age
21	Black females 25- 44 years of age
22	Black males 25- 44 years of age
23	Other females 25- 44 years of age
24	Other males 25- 44 years of age
25	Non-Hispanic White females 45-64 years of age
26	Non-Hispanic White males 45-64 years of age
27	Hispanic females 45-64 years of age
28	Hispanic males 45-64 years of age
29	Black females 45-64 years of age
30	Black males 45-64 years of age
31	Other females 45-64 years of age
32	Other males 45-64 years of age

References

Popoff C, Judson DH and Fadali B. *Measuring the Number of People Without Health Insurance: A Test of a Synthetic Estimates Approach for Small Areas Using SIPP Microdata*, paper presented at the 2001 Federal Committee on Statististical Methodology Conference.

Ghosh M, Kim D and Sinha K. *Heirarchal Bayesian Models for Small Domain Estimation.*