Can Multiple Imputation Help Improve Children's Health Insurance Coverage Estimates from the Current Population Survey?

By

Rebekah Young

and

Luis A. Sanchez

Departments of Sociology and Demography

The Pennsylvania State University

University Park, PA 16801

August 2009

A working paper submitted to the Population Association of America (PAA) 2010 Annual Meeting in Dallas, TX, April 15-17, 2010.

Abstract

The Current Population Survey is known to produce larger state level estimates of the number of uninsured children than those observed from other sources. The hot deck methods used to correct item nonresponse for the questions comprising these estimates are biased and the need for practical alternative options for correcting this situation is widely recognized. While it is clear that the Census Bureau should consider more modern techniques for dealing with missing data, such as multiple imputation, this is not a clearly feasible option at present. In the meantime, we examine the contribution that multiple imputation methods make towards correcting nonresponse among health insurance coverage variables. We believe that multiple imputation can be a useful tool for (1) improving the construction of the current hot deck, (2) developing correction ratios, and (3) assisting in the identification of crucial state-level variables for weights designed to correct biased estimates already in use.

Extended Abstract

The Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) is the most commonly cited source of estimates of the number of uninsured children in the United States. However, the CPS is known to produce larger state level estimates of the number of uninsured children than is observed from other sources. These estimates are a key component of federal allocation formulas that distribute between $3-4 billion federal funds to the State Children's Health Insurance Program. Reliance on these potentially biased estimates has important consequences, specifically for vulnerable populations.

The hot deck (HD) methods used to correct item nonresponse for the questions that comprise ASEC's estimates have been shown to be biased. Possible explanations for its biased estimates include a small sample size for state-level estimates and lack of theoretically grounded variables HD uses to match donors. The need for practical, alternative options for correcting this situation is widely recognized. While it is clear that the Census Bureau should consider more modern techniques for dealing with missing data such as multiple imputation (MI), this is not a clearly feasible option at present. In the meantime, the estimates from MI can be used to help improve the current imputation procedures as well and to develop weights for correcting biased estimates already in use.

In this paper we examine the contribution that multiple imputation methods make towards correcting nonresponse among health insurance coverage variables. In our preliminary analyses we use the 2007 CPS file to construct a MI model to account for nonresponse on health insurance coverage. We then compare (state-level) results derived from the MI model to those

derived using HD procedures. Additionally, recognizing our units of analysis (states) are not independent due to spatial relationships, we project our results in ArcGIS (spatial software) to investigate spatial patterns which may assist in determining weights or adjustment ratios to correct imputation bias.

Our preliminary results compare the HD and MI models' estimates of uninsured children. We calculated differences by subtracting the MI model estimate from the HD model (HD – MI = difference in % children uninsured). In addition, we used a two-tailed t-test of difference in proportions to test the significance of model differences. We found significant differences ($p < .001$) for 33 states in the models' estimations of percent uninsured children. The MI estimates almost universally estimated smaller percentages of children being uninsured when compared to HD estimates. The differences ranged from -.73 % (Hawaii) to 2.7% (Utah). It is important to note that even small differences in percentage points have large implications for federal funding that uses these estimates in the allocation formulas.

We examined our results on GIS-produced maps and discovered evident regional patterning of method bias, particularly in the western states. This spatial patterning suggests reason to sort the hot deck based relationship between geographic location *and* magnitude of difference. Additionally, we set out to explore state-level variables that might explain variation in the magnitude of differences between HD and MI estimates. We discovered that states with larger populations of children and Hispanics were associated with a greater magnitude in the difference between HD and MI estimates. Surprisingly, the percentage of imputed data per state was only slightly related to the magnitude of difference. We believe that when creating weights to correct the estimates, weighting respondents to be representative of age and ethnic structure of the population will be a crucial factor in the performance of the weights.

Introduction

The Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS) is the most commonly cited source of estimates of the number of uninsured children in the United States (Fisher and Turner 2003). These estimates, however, are known to produce larger state level estimates of the number of uninsured children than estimates observed from other, possibly more accurate, sources (Lewis, Ellwood and Czajka 1998). When used to make policy decisions or allocate federal funds, reliance on these potentially biased estimates has important consequences, particularly for vulnerable populations.

For half of a century, hot deck allocation methods have been one of the common approaches used by the Census Bureau to deal with incomplete item-level data. These procedures have come under increasing criticism for yielding biased population and subpopulation estimates and for underestimating the amount of uncertainty in the imputed values. In the ASEC, roughly 11% of individuals do not answer the health insurance supplement. The hot deck methods used to correct this item nonresponse have been shown to be biased (Davern et al. 2004). Alternative ways of accounting for the effect of the missing data that could yield less biased estimates have been proposed but there is little consensus about the best approach. The need for practical alternative options for correcting this situation is widely recognized. In this paper we examine the contribution that multiple imputation methods make towards correcting nonresponse among health insurance coverage variables for the subpopulation of children.

Background and Rationale

The quality of data on health insurance coverage has important implications for research, policy, and social welfare. For researchers, the ASEC is often used as the "gold standard" by which investigators measure the quality of their data and develop population estimates for applying weights to survey data (Groves 2004). For example, the National Survey of American Families uses the ASEC estimates as external validation for their sample distribution of earnings (NSAF). Researchers also use the ASEC data to produce state and county level estimates of various social and economic characteristics such as health insurance coverage to examine a wide range of substantive questions.

The ASEC data are the official source used to estimate the number of uninsured children and number of children living in poverty in each state. The state level estimates of the number of uninsured children is a key component of federal allocation formulas that distribute $3-4 billion of federal funds to the State Children's Health Insurance Program (SCHIP; Census Bureau 2005). The Centers for Medicare and Medicaid Services report that more than 6.6 million children were enrolled in the SCHIP programs at some point during 2006 (SCHIP). On February 4, 2009, President Obama signed into law the Children's Health Insurance Program Reauthorization Act of 2009 (CHIRPA), a new law that will allocate $32.8 billion to states over the next four and a half years to cover an additional 4 million uninsured children, illustrating the increasingly important role that estimates of the uninsured will play in allocation of federal dollars during the next few years.

In the 2007 ASEC, 13.2% of the items comprising the variables indicating whether or not children have health insurance contain missing data. Item non-response is a common cause of missing data in survey research and researchers have implemented a variety of imputation strategies to deal with this issue (Allison 2001; Schafer 1997; Schafer and Graham 2002).

Imputing missing values is a process of replacing a missing (unknown) value with a plausible estimate. This method of dealing with missing data is generally regarded as preferable to options such as complete case analysis, which limits analysis to the subset of cases with no missing information, or mean substitution, which assigns to each missing case the average value observed from complete cases (Allison 2001).

The ASEC employs three principal imputation methods, relational imputation, longitudinal edits, and hot deck (HD) allocation (CPS 2003). Relational imputation assigns values for blank or inconsistent responses on the basis of other characteristics on the person's record or information from other members of the household. Longitudinal edits (primarily used for labor force edits) look at a previous month's data to replace the missing value. Finally, the method used to replace the health insurance variables of primary interest for this paper, HD allocation assigns responses for missing data to sample persons with information from matched sample persons with similar demographic and economic information who answered the same questions. HD imputation techniques assign actually observed values from a non-missing record, called the donor, to a record with a missing value, called the recipient. Donors and recipients are matched on key demographic variables such as age, sex, employment status and other characteristics of the household. Replacing a missing value with a value that actually occurred in the dataset is generally considered an advantage of the method. All missing data within the health insurance variables are replaced with HD imputation.

There are many types of HD procedures, which are differentiated by how the donor case is chosen (Huisman 2000). The HD imputation used by the ASEC is conducted in a logical and deliberate sequence. Values are not imputed for inappropriate or illogical entries and out-of-range values are not permitted. Data are sorted by state and primary sampling unit (PSU) such

that missing values are typically allocated from geographically related areas. For example, missing values for records in Oregon are not likely to be matched to observed records for Pennsylvania. Geographic information, however, is not part of the HD, a point we shall return to later. This distinction is critical due to the geographic clustering of labor force and industry and occupation characteristics known to be related to health insurance coverage. (For a more detailed description of the ASEC HD procedures used by the Census Bureau for health insurance coverage see Davern et. al 2004.)

The HD procedure used by the Census Bureau has long been a subject of criticism. For example, Rubin (1983) shows that the CPS HD underestimates income by 7 percent and Lillard, Smith and Welch (1986) suggest that the CPS HD underestimates wages and salary by 73 percent. The ASEC health insurance coverage estimates have continued to be scrutinized through recent years, with many researchers finding that the ASEC estimates of people without health insurance coverage are higher than those found in other large surveys (Bennefield 1996; Fronstin 2000; Lewis, Elwood, and Czajka 1998). Davern et al (2004) demonstrate that the HD employed in the ASEC leads to bias in estimating health insurance coverage for subpopulations at the state level.

This bias produced by the HD is not particularly surprising. When missing values are replaced by imputed values a single imputation will generally underestimate variability (Rubin 1987). Additionally, a reasonably large sample size is required for a HD method to work properly. Small scale estimates from the ASEC HD imputation, such as those used to generate estimates for a state-level sub-population of uninsured children for example, may be especially problematic due to the small sample size of available donors. Although the ASEC is a large scale survey of more than 78,300 households nationwide, the state-specific sample sizes vary widely

from approximately 900 interviewed households in Arkansas to 5,600 in California. This makes the potentially available "donors" a relatively small group, which means that the number of variables to be included in the deck is restricted.

Little and Rubin (2002) have illustrated that multiple imputation produces unbiased estimates of the mean in the presence of missing data. Davern et al. (2004) suggest that MI may be a preferable method that the Census Bureau should explore. The Census Bureau recognizes that the state-level estimates are insufficient for many policy purposes and has recently made a number of revisions in an attempt to improve the quality of the estimates for insured and uninsured data. The 2007 March ASEC (with results for health insurance data for calendar year 2006) reflects these changes. The Census Bureau has called for external research proposals to provide insights on the best way to impute missing items for surveys to help improve their results. Specifically, they are interested in studies that discuss imputation techniques and how they can be applied to Census Bureau data and an assessment of MI (Census Bureau 2006).

There are reasons to believe that MI will offer better estimates than a HD procedure. Research on the theoretical properties of HD methods is sparse, especially compared to MI which is strongly theoretically grounded (Little and Rubin 2002; Marker, Judkins, and Winglee 2002). Whereas a single imputation is likely to underestimate the error variances and provide biased significance tests, MI does not (Little and Rubin 2002; Schafer and Graham 2002).

The HD procedure is severely limited by the number of categories and variables that can be used to create the deck. A complete listing of the variables used in the ASEC HD are included in Table 1, though it is important to realize that only a subset of these variables are included for the imputation of each variable. For example, the imputation of whether or not a person has group health coverage uses two decks. First, people are divided into groups of workers and non-

workers. Workers are allocated based on Age1, Family Relation, Class of Worker, Earnings Level and Firm Size. Non-workers are allocated based on Age1, Government Health Coverage and Family Relation. The maximum number of variables included in any particular ASEC deck is six and the minimum is two.

Effective HD imputations should match the donor on as many characteristics as possible, but reliance on too many characteristics may result in too few matches and donors must be used who are less similar. The constraints imposed by sample size also lead to arbitrary categorization of variables and recoding of the informing variables so that much of the information and variance is lost. For example, the categorization of variables such as age is necessary for enough matches to occur in the deck, but this approach reduces age variance in the absence of any clear rationale. Further, if the matching categories used do not represent all the important correlates of the variable being imputed then the relationship between the imputed values and other variables in the data can be distorted. For instance, in the example above, marital status may be an important correlate of whether or not a person is covered by group health insurance but is not included in the deck. The Census does not state a logical reason as to why particular sets of variables are included in each allocation specification. Identification of variables related to item non-response to health insurance items as well as those related to health insurance coverage itself may be an important step towards improving the construction of the deck.

An important consideration in selecting a strategy to impute missing data is how consistently the method can yield plausible estimates that do not bias results gleaned from the data. Practicality is also of great consequence. While is clear that the Census Bureau should look into more modern techniques for dealing with missing data, MI is not a clearly feasible option at present. In the meantime, how can we use the estimates from MI to inform future changes and

research?

If the estimates from MI can be regarded as unbiased estimates of the true value of the sample mean, then exploring the source of discrepancies between the original HD estimates and those from MI could offer several valuable methodological insights. First, MI does not operate under the same constraints as a HD procedure because hundreds of variables can theoretically be taken into consideration in the imputation process. This allows us to explore the contribution of a wide variety of variables that could be incorporated into future deck construction. Second, if weights or adjustments ratios are used to correct the imputation bias (e.g. Ziegenfuss 2009), the estimates from MI offer us a standard for comparison. Third, the MI estimates themselves could easily be used to create an adjustment ratio for descriptive results. Finally, the variables identified as being related to the magnitude of misestimating by the HD could be used in the development of adjustment weights. This paper is a beginning exploration of these potential advantages.

Data and Methods

The CPS is a monthly survey of 50,000 or more households conducted by the Census Bureau for the Bureau of Labor Statistics mainly to estimate the unemployment rate (Census Bureau 2005). The ASEC is a supplement to the CPS that is conducted annually in the month of March. The March CPS supplement contains approximately 78,000 households and includes detailed income and health insurance questions asked of the household respondent for every household resident (Census Bureau 2005). Respondents are asked about health insurance coverage for the previous calendar year for themselves and for all other household members. We use the 2007 file which describes health insurance coverage for all or part of 2006.

The Census Bureau distinguishes between private and government health insurance. Private health insurance is provided by an employer or union or can be privately purchased and unrelated to employment. Government health insurance includes Medicare, Medicaid, military health insurance, health insurance from somebody outside the household and "other". Respondents are asked separate questions about each type of health insurance and asked to answer yes or no for each type. Those who answer "no" have their answers verified. People are considered insured if they were covered at any time during the year.

We construct a variable indicating whether or not children in the household were covered by health insurance by combining three questions asking whether or not children in the household were covered by health insurance of someone in the household, covered by health insurance of someone outside the household, or covered by any other type of health insurance. All imputed variables in the ASEC data have been "flagged" such that imputed values can be distinguished from reported values, but these flags are not without limitation. The documentation of the children's health insurance coverage is unclear about the imputation of coverage for each of the dependents in the family. We follow the logic used by Davern et al. (2004) to deal with this limitation and only treat those who were allocated to have a family policy as missing cases. We create a single flag indicating whether or not each case included an imputed value on health insurance and use the flag to set the values imputed by the HD to missing.

Overall, 13.2% of the responses (unweighted) indicating whether or not children were covered by health insurance contained imputed values. Percentage of data imputed by state is included in Table 3. Connecticut, Florida, New York, Vermont and Nebraska had the highest percent of imputed data, with each of these states having more than 17 percent missing responses. Montana, Oklahoma, Alabama, Idaho and Arkansas had the lowest percent of imputed

data, with each of these states having less than 9 percent missing responses.

The MI model we construct was designed to maximize the amount of information included in the estimation within the limits imposed by available software. The model included all household members and every question the ASEC survey asked regarding details of health insurance coverage. In addition to health insurance coverage variables, we included 124 auxiliary variables. This auxiliary information included all variables that the Census Bureau used in their HD matrices, a set of variables that were chosen because they were highly correlated with children's health insurance coverage, and additional demographic variables we expected to be related to health insurance coverage at individual levels. The comprehensive list of auxiliary variables is included in Table 2. We used Stata ICE (Royston 2005) to perform the imputations. The model was constructed under the fully normal assumption except for the three variables regarding children's health insurance, which were imputed using a logistic model. We ran 200 burn-in iterations, 100 between-dataset iterations and used 5 datasets.

We acknowledge that our units of analysis (states) are not independent due to spatial relationships. and may have similar characteristics with nearby states. To account for spatial relationships among state-level data, we use geographic information systems (GIS) to better visualize the data and further investigate the geographic distribution of our variables of interest. One of the primary reasons the HD is known to be biased is due to state-level associations with health insurance coverage that are not accounted for by variables in the HD (Davern et al. 2007). Unfortunately, without a much larger sample size, the variable "state" cannot be included in the HD. Uncovering spatial patterns of state's demographic characteristics may assist in determining weights or adjustment ratios to correct imputation bias.

We performed the analysis in ArcGIS, a spatial software, which allows for the ASEC data to be projected across the United States in order to uncover spatial patterns and better understand the significance of geographical differences. This procedure is performed by joining the state data to the corresponding state shapefile (map) in ArcGIS. After constructing this spatial database, we are able to project data such as percentages, counts, and categorical information. An advantage of GIS is the ability to spatially analyze our data *and* display our results in a meaningful way that will be understood by a wide audience. This may be beneficial given the possible policy implications of this research.

Results

The percentage point differences between the HD and MI imputations are shown in Table 3. The column labeled "% imputed" shows the percentage of the health insurance data for children that was missing and therefore imputed. The column labeled "% difference" shows the MI estimate of the percent of children uninsured subtracted from the HD estimate of the percent of children uninsured (HD – MI = difference in % children uninsured). For example, in Alabama, if the HD estimated that 7.7% of children were uninsured and MI estimated that 6.01% of children were uninsured, the difference would be 1.686. The column showing the number of children is a rough estimate of the difference actual number of children that MI estimates to have insurance but the HD does not. These estimates come from the number of children per state reported in the 2000 U.S. Census (Meyer 2001) and should be regarded as very rough estimates only.

We used a two-tailed t-test of difference in proportions, testing the difference between the values imputed using the HD compared to those imputed with MI. MI produced a statistically

significant difference ($p < .001$) compared to the HD estimates for 33 states. The difference was not significant for 18 states. The MI estimates almost universally estimated a smaller percentage of children being uninsured, with the exception of Hawaii, Tennessee, Arkansas, South Carolina, Montana, Michigan and Maryland. Of these states, only the difference between Montana and Hawaii was statistically significant. The largest observed differences occurred for Virginia, Nebraska, Mississippi, Florida, Utah and Arizona, where MI produced estimates between 2 and 3 percentage points lower than HD estimates of uninsured children. Although it is tempting to only pay attention to the most extreme cases, even small differences in percentage points may have large dollar implications for federal funding that uses these estimates in the allocation formulas. Further, the number of children that each percentage point comprises is important when thinking about health care services, and state level distribution of funds.

Figure 1 provides a graphical representation of the information in Table 3. Looking at the magnitude of how the differences between HD and MI estimates vary by state underscores an important point. Geography clearly matters when adjusting for estimation errors in the ASEC estimates. A “blanket” correction or national-level adjustment ratio would not improve the estimates. Two states, Montana and Hawaii, may not need to have their estimates adjusted down at all.

If state-level characteristics are impossible to incorporate into the HD, intentional regional sorting is likely to be helpful. Some regional patterns of the magnitude of difference (or bias) are evident, particularly in the western states. It may be useful to sort the deck based on magnitude of bias rather than arbitrary geographical location. For example, Washington, Oregon, Idaho, Nevada, and California have similar levels of bias and it is probably reasonable for donors from any of these states to contribute to recipients in another. Although Nebraska is

geographically close to South Dakota and Iowa, it may be more meaningful to sort it closer to Utah and Arizona.

Finally, we set out to explore state-level variables that might help explain the variation in the magnitude of differences between HD and MI estimates. Insight into the variables that help account for this difference may help improve a future version of the deck used by the Census Bureau as well as contribute information to the development of state-level adjustment ratios or national weights. We expect this to be true even if the variables we identify are only proxy information for a more direct cause or due to a spurious relationship.

Variables that are highly correlated with difference between the HD and MI estimates are listed in Table 4. The percentage of the Hispanic population and the percentage of the population under age 18 by are shown by state in Figure 2. We discovered two important differences. First, the age structure of the population is related to the magnitude of difference between the HD and MI estimates; the larger the proportion of the population is children, the larger the level of bias produced by the HD. Second, the greater the proportion of the population is Hispanic, the greater the magnitude of difference between the HD and MI. The mere correlation between these state-level variables and the magnitude of bias again suggest that a more meaningful geographic sort of the allocation deck could help overcome some of the state-level reasons known to bias the current deck. Additionally, when weights are constructed to correct the estimates, it is likely to be important to match the respondents to state-level proportions on these particular variables.

Surprisingly, the percentage of imputed data was only slightly related to the magnitude of the difference between HD and MI. Knowing that the HD allocation is biased, we might expect to find more error in the estimates for states that had a higher proportion of missing data. We did

not find this to be the case, as percentage of imputed data explained virtually none of the variation in magnitude of difference (results not shown).

Discussion

Our primary research goal was to explore the contributions that MI could make towards improving and correcting the ASEC estimations of health insurance coverage. We compared the difference in children's health care insurance estimates from the ASEC based on the publicly released HD imputations and on our own MI estimates. We found that, as expected, the MI approach found lower percentages of uninsured children by state than the ASEC estimates. Our findings suggest that sorting the HD based on the geographic magnitude of bias in the HD estimates may be a useful method of improvement. Incorporating individual level variables about Hispanic ethnicity might help improve the matching of donor and recipient allocations. There is potential for the difference between the HD and MI estimates to help serve as a state-level correction ratio. Finally, we believe that when creating weights to correct the estimates, weighting respondents to be representative of age and ethnic structure of the population will be a crucial factor in the performance of the weights.

There are some important limitations to our approach. Statisticians such as Little and Rubin (2002) have shown that MI is more likely to yield estimates that more accurately take into account the uncertainty introduced by the imputation than those from a single imputation method. Because the true value in the population is not known, however, it is not possible to say that the MI estimates are in fact correct. Further, both missing data techniques compared here handle only item level missingness. Nonresponse to the whole survey, which is a high as 17 percent of people in the CPS, could be a greater problem. Survey nonresponse is adjusted for by

weighting the data and no comparisons have been made of this approach compared to an MI or HD approach.

The set of variables we include in the MI model are not ideal. First, we were unable to include state of residence in the model. The amount of computer time it takes to run an imputation model increases exponentially as the number of variables increases. We estimated that including 50 additional parameters to incorporate the states would have caused the imputation to take approximately six weeks to run with no guarantee that the estimates would converge. Although Davern et al. (2004) argue that part of the reason the Census HD is biased is due to its inability to account for geographic location, this could not be completely overcome with the MI approach. This may be a limitation because we would expect that state policy would impact the proportion of children who have health insurance. A compromise was to include region of the country to help capture some of the geographic location correlates. Region could easily be included because only four parameters needed to be added to the model.

The second major limitation is that although all the individual variables from the HD are included in the MI model, MI software cannot support the range of interactions between these variables that is an inherent part of the HD. These two limitations mean that the MI model is more informed than the HD in some respects but not in others. On the other hand, the MI model does not require that continuous variables such as age, education, income, etc. are represented by a small number of categories as is required in an HD approach.

Another possible limitation relates to the way we handled missing data in variables other than those that made up the health insurance measures. Many of these other variables had their missing values imputed with the HD approach. Because some of the variables we used in the MI imputation had imputed HD values, this may have lead to different estimates than might have

been obtained had we removed HD imputed values for all variables. Because the MI needs to be at least as informed as the HD imputation, we would have had to use all the variables included in the hot deck models for each of the variables which would likely have taxed the capacity of the MI software.

We asked the question, can multiple imputation help improve children's health insurance coverage estimates from the Current Population Survey? We believe there are some promising ways that MI can contribute to improving future hot deck allocations, creating state-level adjustment ratios, and identifying crucial variables in the development of correction weights. Clearly the next step in our analysis is to test the actual performance of our suggested contributions.

References:

Allison, Paul D. 2001. *Missing Data*, vol. 07-136. Thousand Oaks, CA: Sage.

Bennefield, R. L. 1996. "A Comparative Analysis of Health Insurance: Data from CPS and SIPP." in *Joint Statistical Meetings of the American Statistical Association*. Chicago: U.S. Census Bureau.

Census Bureau, U.S. 2005. "Health Insurance Overview." vol. 2007. Washington D.C. : U.S. Census Bureau.

—. 2006. "Potential Research Data Center Methodological Topics." Washington, D.C. : U.S. Census Bureau.

CPS. 2003. "Current Population Survey Technical Paper 63: Design and Methodology." edited by U. S. C. B. f. B. o. L. Statistics. Washington: U.S. Census Bureau.

Davern, Michael , Lynn A. Blewett, Boris Bershadsky, and Noreen Arnold. 2004. "Missing the Mark? Imputation Bias in the Current Population Survey's State Income and Health Insurance Coverage Estimates." *Journal of Official Statistics* 20:519-549.

Davern, Michael, Holly Rodin, Lynn A. Blewett, and Kathleen Thiede Call. 2007. "Are the CPS Uninsurance Estimates Too High? An Examination of Imputation." *Health Services Research* 45:2038-2055.

Fisher, Robin and Joanna Turner. 2003. "Health Insurance Estimates for Counties." in *2003 Joint Statistical meetings - Section on Survey Research Methods*: U.S. Census Bureau.

Fronstin, P. 2000. *Counting the Uninsured: A comparison of National Surveys. Employee Benefit Research Institute Brief No. 225*. Washington, D.C.: The Employee Benefit Research Institute.

Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology.* Hoboken, NJ: Wiley.

Huisman, Mark. 2000. "Imputation of Missing Item Responses: Some simple techniques." *Quality and Quantity* 34:355-351.

Lewis, Kimball, Marilyn R. Ellwood, and John L. Czajka. 1998. *Counting the Uninsured: A Review of the Literature*. Washington D.C.: The Urban Institute.

Marker, David A., David R. Judkins, and Marianne Winglee. 2002. "Large scale imputation for complex surveys." in *Survey Non-Response*. New York: Wiley.

Meyer, Julie. 2001. *Age: 2000* Washington, D.C. : U.S. Census Bureau, Economics and Statistics Administration.

NSAF. 1997 and 1999. "National Survey of America's Families Methodology Reports." The Urban Institute Web site.

Royston, Patrick. 2005. "Multiple imputation of missing values." *Stata Journal* 4:227-241.

Rubin, Donald B. 1983. "Imputing income in the CPS: Comments on "Measures of aggregate labor cost in the United States." Pp. 333-344 in *The measurement of labor cost*, vol. 48, *Studies in income and wealth*, edited by J. E. Triplett. Chicago: University of Chicago Press.

—.1987. *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate data*. New York: Chapman and Hall.

Schafer, Joseph L and John W. Graham. 2002. "Missing Data: Our view of the state of the art." *Psychological Methods* 7:147-177.

SCHIP. 2007. *FY 2007 Annual Enrollment Report*. Baltimore, MD: Centers for Medicare and Medicaid Services.

Ziegenfus, Jeanette. 2009. "A Correctin for the Full-Supplement Imputation Bias in the Current Population Survey's Annual and Social Economic Supplement." in *AAPOR Proceedings*. Hollywood, FL.

Table 1. Range of Variables Included in the CPS ASEC Health Insurance Edit and Imputation Specifications

Age1	
	15-24
	25-34
	35-44
	45-64
	65+
Age2	
	Less than 15
	15-24
	25-44
	45-64
	65+
Children	
	One or more own children under 18
	All others
Class of Worker	
	Self-employed
	All others
Earnings Level	
	Under $2000
	$2000-14999
	$15000-29999
	$30,000 or more
Family Relation	
	Reference person (w/relatives) or spouse
	Child or other relative
	Unrelated individual
Government Health Coverage	
	Covered by Medicare, Medicaid, or CHAMPUS
	All others
Group Health Coverage	
	Covered by employers-provided health plan
	All others
Marital Status	
	Married
	Never Married
	Divorced or Separated
	Widowed
Poverty Status	
	Received public assistance or SSI
	All others
Privately Purchased Health Coverage	
	Covered by privately purchased health plan
	All others

Table 1. (continued)

Size of Employer
Firm
Under 25 employees
25-499 employees
500-999 employees
1000 or more employees
Social Security
Income
Received Social Security
All others
Spouse Employment Status
NIU (non married)
Spouse worked last year
Spouse did not work last year
Veteran Status
Veteran
Non-Veteran
Current Armed Forces, or longest job last year was AF
Work/Disability
Status
Worked last year
Did not work last year - disabled
All others

Table 2. Comprehensive List of Variables Included in MI Procedure

Age in years
Armed Forces, ever served (yes/no)
Children's health insurance (yes/no), # children covered by
    Insurance of someone not in household
    Medicare
    Other health insurance
Discouraged worker (yes/no)
Educational Attainment
    Children, Less than 1st Grade, 1st through 4th grade, 5th or 6th grade, 7th or 8th grade, 9th grade, 10th grade, 11th grade, 12th grade (no diploma), High school graduate, Some college but no degree, Associates degree in college, Bachelor's degree, Master's degree, Professional school degree, Doctorate degree
Ethnicity
    Hispanic, Spanish or Latino, yes or no
Family Income
Family size (number of persons)
Family Type
    Primary family, nonfamily householder, related subfamily, unrelated subfamily, secondary individual
Food; number of children who ate hot lunch at school
Full/Part-time Status
    Children or Armed Forces, Full-time schedules, Part-time for economic reasons usually FT, Part-time for non-economic reasons usually FT, Part-time for economic reasons usually PT, Unemployed, Not in labor force
Health care coverage (was anyone in the household covered by)
    Employment or union based health insurance coverage (yes/no)
    Private health insurance (yes/no)
    Coverage from outside the household (yes/no)
    Medicare (yes/no)
    Medicaid (yes/no)
    State sponsored health insurance plan (yes/no)
    Other health insurance including CHAMPUS, CHAMPVA, VA or military health care (yes/no)
Health Status
    Self-rated, 1-5
Hours worked last week
Immigrant nativity (in years)
Immigrant Status (yes/no)
Labor Force Status
    Children, Armed Forces, Working, With job not at work, Unemployed looking for work, Unemployed on layoff, Not in labor force
Marital Status
    Married, Never married, Divorced or separated, Widowed
Metropolitan Status
Number of people in the family
Number of own children
    Less than 6 years of age
    Less than 18 years of age
Number of people employeed by employer
Poverty Level (ratio of family income to poverty)

Table 2. (continued)

- Poverty Status
- Public assistance
  - Education assistance (yes/no)
  - Transportation (yes/no)
  - Child care services (yes/no)
  - Public housing (yes/no)
  - Job assistance (yes/no)
  - Food stamps
    - Recipient (yes/no)
    - Value in dollars
    - Number of children covered
    - Number of months covered
  - Food programs
    - Free lunch (yes/no)
    - Reduced lunch (yes/no)
  - WIC program benefits (yes/no)
  - Energy assistance (yes/no)
- Race
  - White only, Black only, American Indian or Alaskan Native only, Asian only Hawaiian/Pacific Islander only, mixed races
- Reason not working
  - Not in labor force, Ill or disabled, Taking care of home or family, Going to school, Could not find work, Other
- Region
  - Northeast, Midwest, South, West
- Residential mobility; moved since last year (yes/no)
- Sex
  - Male, Female
- Source of Income
  - Unemployment compensation (yes/no)
  - Worker's compensation (yes/no)
  - Social Security Income (yes/no)
  - Supplemental Security Income (yes/no)
  - Public Assistence or Welfare (yes/no)
  - Veterans' Administration benefits (yes/no)
  - Disability income (yes/no)
- Total person income in $2,500 increments

Table 3. Percentage Difference in HD and MI Estimates by State

State	*% imputed*	*difference in %*	*# of children*	*State*	*% imputed*	*difference in %*	*# of children*
Alabama	7.4	**1.686***	9,165	Montana	5.4	**-0.208**	(227)
Alaska	14.6	**1.224**	1,898	Nebraska	17.3	**2.526**	5,509
Arizona	13.6	**2.770**	19,362	Nevada	9.7	**1.408**	4,124
Arkansas	8.5	-0.455	(1,393)	New Hampshire	13.2	0.497	803
California	12.0	**1.256**	70,998	New Jersey	15.1	**1.906**	18,518
Colorado	13.5	0.085	585	New Mexico	11.3	**1.845**	6,795
Connecticut	20.9	0.313	1,163	New York	19.2	**1.126**	25,249
Deleware	14.2	2.125	1,974	North Carolina	12.1	**1.871**	18,613
District of Columbia	15.2	**1.755**	791	North Dakota	9.2	0.684	454
Florida	20.2	**2.578**	21,625	Ohio	13.1	**1.561**	21,557
Georgia	12.4	**1.762**	24,381	Oklahoma	7.1	**0.614**	2,681
Hawaii	14.3	**-0.723**	(977)	Oregon	11.4	**1.132**	4,624
Idaho	7.9	**1.518**	3,387	Pennsylvania	15.0	**0.757**	7,598
Illinois	14.9	**2.003**	34,965	Rhode Island	15.6	0.645	616
Indiana	11.8	**1.493**	12,262	South Carolina	12.3	**-0.291**	(1,524)
Iowa	11.9	**1.419**	4,221	South Dakota	9.2	1.421	1,343
Kansas	9.1	0.095	337	Tennessee	16.3	-0.472	(3,282)
Kentucky	13.7	**1.449**	6,955	Texas	10.9	**0.689**	26,283
Louisiana	15.8	**1.298**	9,122	Utah	14.6	**2.692**	14,226
Maine	12.6	0.589	695	Vermont	19.0	**1.098**	769
Maryland	10.0	-0.014	(109)	Virginia	14.0	**2.349**	22,223
Massachusetts	12.9	0.563	3,601	Washington	12.8	**1.353**	11,525
Michigan	11.5	-0.123	(1,695)	West Virginia	12.0	0.451	566
Minnesota	10.2	**0.806**	5,584	Wisconsin	12.9	0.591	3,936
Mississippi	15.9	**2.576**	11,120	Wyoming	10.8	0.207	148
Missouri	15.1	0.338	2,275				

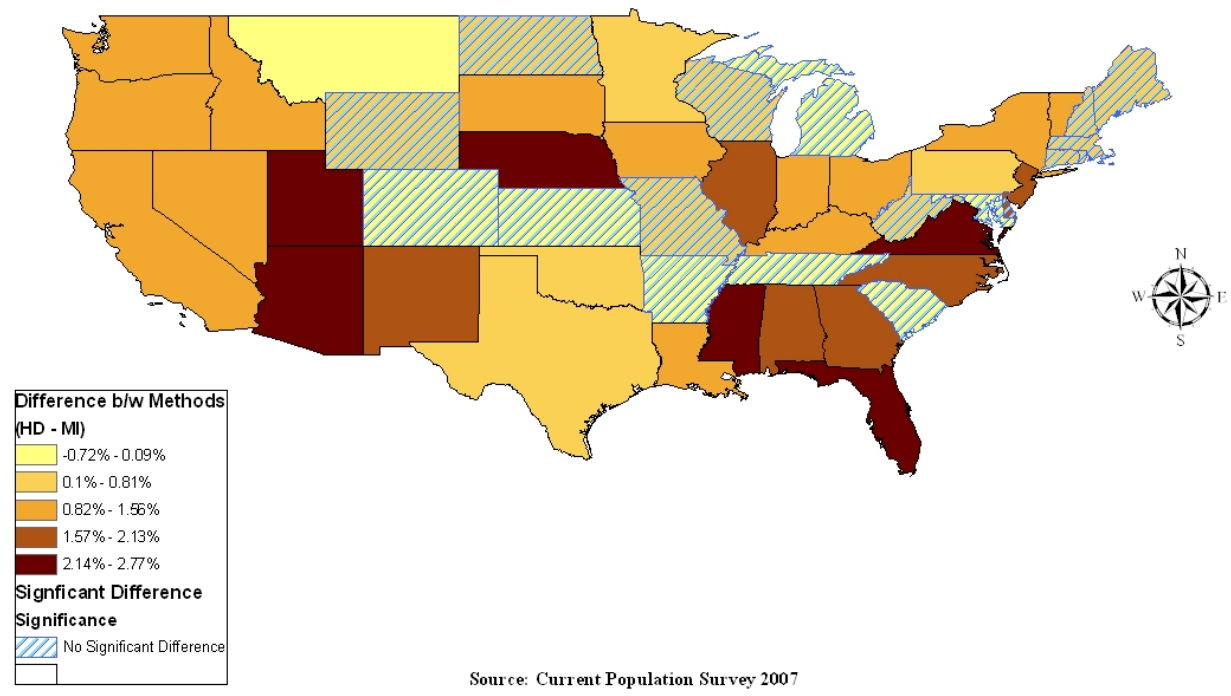
Notes: * **Bold** numbers indicate that the difference between the HD and MI estimates was significant at the .001 level.

Table 4. Correlation of Variables with difference between HD and MI

Variable	*Correlation*
Percentage of the population < 18 years of age	0.3206
Percentage of the child population US citizens	-0.3165
Percentage of the Population more than two races	-0.2756
Percentage of the Population age 65 +	-0.2578
Percentage of the Population age 75 +	-0.2557
Percentage of the population Hispanic	0.2554

Figure - 1

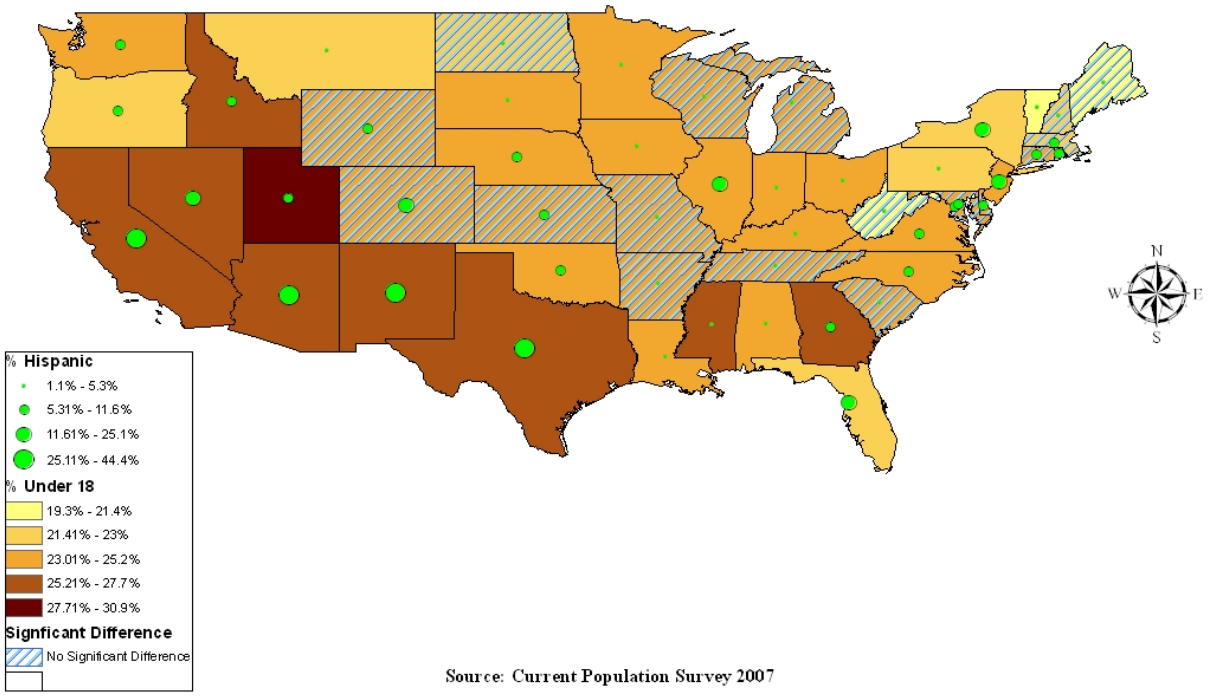
Magnitude of Difference in Estimating Uninsured Children: Hot Deck vs. Multiple Imputation



Source: Current Population Survey 2007

Figure - 2

% Hispanic and % Under 18 by State



Source: Current Population Survey 2007