Imputation of Low-Income Status in the National Survey of Children with Special Health Care Needs, 2001

Matthew D. Bramlett, Ph.D.

These tables report the results of analyses conducted by staff of the Special Population Surveys Branch of the Division of Health Interview Statistics, National Center for Health Statistics. All estimates are subject to sampling variability, as well as survey design flaws, respondent classification and reporting errors, and data processing mistakes.

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics
Division of Health Interview Statistics
Special Population Surveys Branch

October 28, 2003

SAFER • HEALTHIER • PEOPLE™

Introduction

The National Survey of Children with Special Health Care Needs (NS-CSHCN) was conducted by the National Center for Health Statistics in 2000-2002, using the National Immunization Survey (NIS) sampling frame. The NIS is a large random-digit-dialing telephone survey that monitors vaccination of children in the United States (1, 2). The NS-CSHCN module of the State and Local Area Integrated Telephone Survey was sponsored by the Maternal and Child Health Bureau of the Health Resources and Services Administration. The NS-CSHCN obtained interviews with the parents of 38,866 children with special health care needs and 176,296 children without special health care needs from 196,888 households with children in the United States.

Children with special health care needs are children who are limited in their ability to do things most children can do; who need or use prescription medications or get special therapies such as physical, occupational or speech therapy; or who have emotional or developmental problems requiring treatment or counseling for at least 12 months. Data from the survey include a screener file for assessing the prevalence of special health care needs in the population, a household file containing characteristics of the household, an interview file containing interviews with the parents of one special needs child within each household, a file containing health insurance coverage data for children with and without special needs, and a supplemental file comprised of the program participation data for uninsured children in low-income families in the survey.

The interview file contains data on the health and functional status of the child with special needs, that child's access to health care and unmet health care needs, the parent's satisfaction with health care, adequacy of health insurance, and the impact of the child's condition on the family. The Low Income Uninsured Supplement consisted of survey questions collecting much of the same information for children without special health care needs who live in low-income households and are not insured. A separate questionnaire was administered for children without special health care needs in order to determine their insurance status and eligibility for the supplemental sample. This also enabled estimation of health care coverage for all children and comparisons between children with and without special health care needs.

Uninsured children living in households with total household income levels less than 200% of federal poverty guidelines[1] were included in the Low Income Uninsured Supplement. All uninsured children in households for which income was not ascertained (because of don't know or refusal responses) were also included, in order to ensure that all low-income households with uninsured children received the supplement. A total of 29,463 households (15% of the sample) were missing data on income. An imputation algorithm was derived to predict which households with missing income were most likely to have income below 200% of the federal poverty level. For uninsured households with missing income, an indicator variable was coded to identify imputed income status below 200% of the federal poverty level or not. This brief report describes the imputation algorithm in detail.

[1] The coding of poverty status according to federal poverty guidelines is described in the methodology report for NS-CSHCN (3)

Specification of the Imputation Model

The imputation of low-income status was accomplished by fitting a logistic regression model to predict income status (below 200% of poverty or not, coded as 1 or 0) for all households for which income data were collected, and then applying that model to predict income status for households where income was not ascertained. The model includes two classes of variables: characteristics of the household (such as presence of a fax line, more than one phone line, number of adults, number of special needs children, and age of the oldest child); and characteristics of the community, where "community" is defined as households sharing a common telephone exchange (the first three digits of a telephone number following the area code).

The telephone exchange is not always an accurate indicator of the physical boundaries of a neighborhood. However, data measuring the characteristics of the physical community were not available. Telephone exchanges are a proxy for conglomerations of neighbors because exchanges are typically assigned on a geographic basis. Households that share a common telephone exchange are generally more likely to be located near one another geographically than are households that do not share a common telephone exchange. There are 37,317 telephone exchanges represented in the NS-CSHCN data. Each exchange is represented by an average of 5.3 households, ranging from 1 household to 131 households within the exchange.

If households clustered by a telephone exchange closely resemble each other and do not closely resemble households with other exchanges, simple regression models that include only household characteristics but do not control for the clustered nature of the community context can overstate the effects of household variables in the model. Methods of dealing with such clustering include: 1) the inclusion of indicators of the characteristics of communities defined by common telephone exchange in the model, or 2) multilevel modeling, which explicitly models the clustering itself by removing correlated errors due to clustering from the residual error term (4).

Iterative analyses using multilevel models with too many explanatory variables that are related to each other will often not converge on a single solution. It is preferable to use theory and preliminary analysis to keep the number of explanatory variables as low as possible (5, 6). This approach is useful when the researcher wants to investigate the effect of a particular explanatory variable on the dependent variable of interest and needs to control for clustering. However, the goal here is to explain, or predict, low-income status of households as accurately as possible, and it is therefore preferable in this case to include many explanatory variables to derive the best prediction of the value of the dependent variable; the particular effects of the explanatory variables are of secondary importance to the goal of accurately predicting low-income status.

In addition, preliminary analysis suggested that the data requirements for multilevel modeling to account for clustering of data in telephone exchanges could not be met. Multilevel modeling is not robust, and coefficients are likely to be unstable or biased when there are very few cases in too many of the clusters (5). In the NS-CSHCN data, 26% of the telephone exchanges are represented by only one household and 67% are represented by fewer than 5 households. As a result, multilevel modeling is not feasible in this case. Because the amount of clustering is small, it is more appropriate to simply include characteristics of the telephone exchange community in the model at the household level.

The model used to impute low-income status is a simple logistic regression model that includes 21 indicators of household characteristics and 13 indicators of characteristics of communities defined by common telephone exchange. All households with valid income data were included in fitting this model, not just those with uninsured children. The model was fitted using SUDAAN.

Model Results

The goal of the model is to predict low-income status, not to investigate the effects of particular variables on low-income status, but the model results are presented here for interested researchers. The imputation model contains 34 variables representing aspects of the household and of communities defined by common telephone exchange. The variables are listed in Table 1, which shows the parameter estimate, the standard error, and odds ratio associated with each variable. The interpretation of the odds ratios in this table is slightly different depending on whether the variable is a continuous or dichotomous measurement. For dichotomous variables that indicate the presence or absence of a particular characteristic, the odds ratio describes the odds that households with the indicated characteristic have income below 200% of poverty, compared with households without the indicated characteristic. For example, Table 1 shows that the odds of being below 200% of poverty were 311 times as high for households receiving welfare as for households not receiving welfare. For continuous variables, the odds ratio describes the incremental change in the odds of being below 200% of poverty for each incremental increase in the value of the continuous variable. For example, Table 1 shows that each additional adult in the household increases the odds of being below 200% of poverty by 2%.

Predictive Accuracy

The main goal of the analysis was to produce a model that, as a whole, accurately predicts low-income status. There are various measures of "goodness of fit" (i.e., predictive accuracy) that are typically used to compare competing models to determine which model performs better at fitting the data. One very basic measurement of model fit is the percent of cases in which the actual value of the dependent variable (below 200% of poverty or not) matched the predicted value generated by the model. Predicted probabilities range from 0 to 1, and values between those extremes are rounded down to 0 or up to 1 in order to make the comparison with the actual values, where 0 = not below 200% of poverty and 1 = below 200% of poverty.

Table 2 shows the comparison of actual and predicted low-income status for the unweighted distribution of households with valid income data. Summing the percent of households that were accurately predicted to be above or equal to 200% of poverty (61.7%) and the percent that were accurately predicted to be below 200% of poverty (18.8%) results in 80.5% of households accurately predicted. Of the remaining 19.5% that were incorrectly predicted, the majority (16%) were predicted to be above or equal to 200% of poverty when they were actually below 200% of poverty, and only 3.5% were actually above or equal to 200% of poverty and predicted to be below 200% of poverty. This suggests that using the model to impute missing low-income status in order to exclude households above 200% of poverty from analysis will result in retaining very few households with income above or equal to 200% of poverty in the sample.

The imputation model performs better at identifying high-income (i.e., not low-income) status than low-income status. The specificity, or percent of households with income above or equal to 200% of poverty correctly identified as such, is 95% (table 3). The sensitivity, or percent of low-income households correctly identified as low-income, is 54%. The positive predictive value is the proportion of low-income status predictions that are correct, while the negative predictive value is the proportion of high-income status predictions that are correct. When the model predicts low-income status, it is correct 84% of the time, and when it predicts high-income status, it is correct 79% of the time (table 3).

Table 3 shows that two-thirds of the cases with missing income data are predicted to be above or equal to 200% of poverty, and only one-third are predicted to be low-income households. A weighted average of the positive and negative predictive values (weighted by the number of cases with missing income data that are predicted to be above or equal to and below 200% of poverty) suggests the best estimate of how well the model predicts income status for cases missing on income: (9,524)(0.84) + (19,939)(0.79) / (9,524 + 19,939) = 0.81. Assuming that the positive and negative predictive values hold true for the cases with missing income data suggests that 81% of the missing income cases are correctly predicted. While not perfect, this is the best that could be done with the data available at the time.

Uninsured Households

The supplemental survey questions were administered in low-income and missing-income households with uninsured children. Subsetting the analysis of model fit and model predictions to those households with uninsured children suggests that the imputation model does not perform as well for this group as for the full sample. Table 4 shows that 70% of uninsured households were correctly predicted to be either above or equal to or below 200% of poverty. The remaining 30% are almost equally split between those incorrectly predicted to be low-income households and those incorrectly predicted to be high-income (i.e., not low-income) households.

Table 5 shows that 16% of the uninsured households with missing income data are predicted to be above or equal to 200% of poverty, and 84% are predicted to be low-income households. The sensitivity is 75% and the specificity is 59%. The positive and negative predictive values suggest that 54% of the high-income predictions and 79% of the low-income predictions are correct. A weighted average of the positive and negative predictive values suggests the best estimate of how well the model predicts income status for uninsured households missing on income: (332)(0.54) + (1,734)(0.79) / (332 + 1,734) = 0.75. Assuming that the positive and negative predictive values hold true for the cases with missing income data suggests that 75% of the missing income cases among uninsured households are correctly predicted.

Conclusion

Although the model does not perform quite as well for the subsetted sample of uninsured households as it does for the full sample, attempts to develop an imputation model restricted to uninsured households did not reveal a better model. The model that appears in Table 1, which was generated using the full sample of all households with valid income data, was better at predicting low-income status among uninsured households than alternative models restricted to uninsured households. Therefore, the full model remains the best imputation algorithm for this purpose.

References

1. Zell ER, Ezzati-Rice TM, *et. al*. National Immunization Survey: The methodology of a vaccination surveillance system. Public Health Reports, 115: 65-77. 2000.

2. Smith PJ, Battaglia MP, *et al*. Overview of the sampling design and the statistical methods used in the National Immunization Survey. American Journal of Preventive Medecine 20(4S): 17-24. 2001.

3. Blumberg SJ, Olson L, *et al*. Design and operation of the National Survey of Children with Special Health Care Needs, 2001. National Center for Health Statistics. Vital Health Stat 1(41). 2003.

4. Mosher WD, Deang LP, Bramlett MD. Community environment and women's health outcomes: Contextual Data. National Center for Health Statistics. Vital Health Stat 23(23). 2003.

5. Snijders TAB, Bosker RJ. Multilevel analysis: An introduction to basic and advanced multilevel modeling. London: Sage Publications, 1999.

6. Kreft I, DeLeeuw J. Introducing multilevel modeling. Thousand Oaks, CA: Sage Publications, 1998.

Table 1: SUDAAN Logistic regression model predicting low-income status, National Survey of Children with Special Health Care Needs, 2001

Parameter	*Estimate (std err)*	*Odds Ratio*
Intercept	-0.88 (0.21)	
Household (HH) Characteristics		
Presence of phone line dedicated to fax/modem (0/1)	-0.83 (0.05)	0.44*
Presence of more than one voice line (0/1)	-0.49 (0.04)	0.61*
Whether HH ever had phone interruption (0/1)	1.07 (0.05)	2.93*
Phone is directory listed (0/1)	-0.17 (0.03)	0.84*
HH refused the open-ended income question (0/1)	-0.15 (0.04)	0.86*
HH member received assistance from welfare (0/1)	5.74 (0.38)	310.97*
An interviewed child in HH was insured (0/1)	0.72 (0.05)	2.04*
An interviewed child was insured privately (0/1)	-2.25 (0.03)	0.11*
Interview conducted in Spanish (0/1)	1.85 (0.08)	6.35*
Presence of Hispanic child in HH (0/1)	0.58 (0.05)	1.78*
Presence of non-Hispanic black child in HH (0/1)	0.80 (0.04)	2.23*
Presence of college-educated mother in HH (0/1)	-1.13 (0.02)	0.32*
Number of adults in HH	0.02 (0.02)	1.02
Number of children in HH	0.42 (0.02)	1.52*
Age (years) of oldest child in HH	-0.01 (0.01)	0.99*
Age (years) of youngest child in HH	-0.00 (0.01)	1.00
Number of special needs children, category 1[1]	-0.07 (0.03)	0.93*
Number of special needs children, category 2[2]	0.05 (0.05)	1.05
Number of special needs children, category 3[3]	0.26 (0.06)	1.30*
Number of special needs children, category 4[4]	0.07 (0.06)	1.07
Number of special needs children, category 5[5]	0.09 (0.05)	1.09
Telephone Exchange (TEX) Community Characteristics		
TEX households are located in rural area (0/1)	0.21 (0.03)	1.23*
Percent renting in TEX households	0.01 (0.00)	1.01*
Percent college-educated in TEX households	-0.02 (0.00)	0.98*
Percent black in TEX households	-0.00 (0.00)	1.00
Percent Hispanic in TEX households	-0.00 (0.00)	1.00*
Total number of persons in TEX households	-0.00 (0.00)	1.00*
Percent ages 18-24 in TEX households	0.01 (0.01)	1.01
Percent ages 25-34 in TEX households	-0.03 (0.01)	0.98*
Percent ages 35-44 in TEX households	-0.00 (0.01)	1.00
Percent of TEX households with income $10,000 - $14,999	0.03 (0.00)	1.03*
Percent of TEX households with income $15,000 - $24,999	0.05 (0.01)	1.05*
Percent of TEX households with income $25,000 - $34,999	0.00 (0.01)	1.00
Percent of TEX households with income $35,000 - $49,999	-0.02 (0.00)	0.98*
Sample size (unweighted)	167,425	

National Survey of Children with Special Health Care Needs, 2000-2002

* $p<0.05$
[1] Child uses more medical care, mental health/educational services than most children of same age
[2] Child currently needs or uses prescription medications
[3] Child is limited or prevented in ability to do things most children of same age can do
[4] Child needs or gets specialized therapy such as physical, occupational or speech therapy
[5] Child has emotional/developmental/behavioral problem requiring treatment/counseling
For all categories 1-5, problem has persisted or is expected to persist at least 12 months

Table 2: Model fit: Comparison of actual and predicted low-income status for all households with valid income data

	Predicted to be above/equal to 200% of poverty	*Predicted to be below 200% of poverty*
Actually above/equal to 200% of poverty	103,295 (61.7%)	5,872 (3.5%)
Actually below 200% of poverty	26,735 (16.0%)	31,523 (18.8%)

National Survey of Children with Special Health Care Needs, 2000-2002

Table 3: Model predictions for all households

Statistic	*Cases with valid income data*	*Cases without valid income data*
Total unweighted sample size (N)	167,425	29,463
N (percent) below 200% of poverty	58,258 (34.8%)	N/A
N (percent) above/equal to 200% of poverty	109,167 (65.2%)	N/A
N (percent) accurately predicted	134,818 (80.5%)	N/A
N (percent) inaccurately predicted to be below 200% of poverty	5,872 (3.5%)	N/A
N (percent) inaccurately predicted to be above/equal to 200% of poverty	26,735 (16.0%)	N/A
Sensitivity	0.54	N/A
Specificity	0.95	N/A
Positive predictive value	0.84	N/A
Negative predictive value	0.79	N/A
N (percent) predicted below 200% poverty	37,395 (22.3%)	9,524 (32.3%)
N (percent) predicted above/equal to 200% poverty	130,030 (77.7%)	19,939 (67.7%)

National Survey of Children with Special Health Care Needs, 2000-2002

Table 4: Model fit: Comparison of actual and predicted low-income status for households with uninsured children and valid income data

	Predicted to be above/equal to 200% of poverty	*Predicted to be below 200% of poverty*
Actually above/equal to 200% of poverty	2,113 (19.6%)	1,476 (13.6%)
Actually below 200% of poverty	1,830 (16.9%)	5,425 (49.9%)

National Survey of Children with Special Health Care Needs, 2000-2002

Table 5: Model predictions for households with uninsured children

Statistic	*Cases with valid income data*	*Cases without valid income data*
Total unweighted sample size (N)	10,862	2,066
N (percent) below 200% of poverty	7,255 (66.8%)	N/A
N (percent) above/equal to 200% of poverty	3,607 (33.2%)	N/A
N (percent) accurately predicted	7,556 (69.6%)	N/A
N (percent) inaccurately predicted to be below 200% of poverty	1,476 (13.6%)	N/A
N (percent) inaccurately predicted to be above/equal to 200% of poverty	1,830 (16.9%)	N/A
Sensitivity	0.75	N/A
Specificity	0.59	N/A
Positive predictive value	0.79	N/A
Negative predictive value	0.54	N/A
N (percent) predicted below 200% poverty	6,901 (63.5%)	1,734 (83.9%)
N (percent) predicted above/equal to 200% poverty	3,961 (36.5%)	332 (16.1%)

National Survey of Children with Special Health Care Needs, 2000-2002