A COMPARISON OF IMPUTATION TECHNIQUES IN A PHYSICIAN SURVEY

Sara L. Thran and Kurt D. Gillis, American Medical Association
515 N. State Street, Chicago, IL 60610

KEY WORDS: Nonresponse Adjustment

The American Medical Association's (AMA) Socioeconomic Monitoring System (SMS) is an ongoing telephone survey of physicians which collects information about practice characteristics. Survey response rates in recent years have been approximately 70%. Much work has been done examining characteristics of survey nonrespondents, and a weighting strategy has been developed to correct for unit nonresponse to the survey. However, item nonresponse is a problem for some of the survey variables, most notably practice expenses and net income from medical practice. This paper will examine several imputation techniques to adjust for item nonresponse.

Virtually all public use tapes from government surveys have missing values replaced by imputed values. The AMA sells a small number of copies of the SMS public use file each year, mostly to researchers who plan to use the file for multivariate analyses. We are doing this research in the context of determining whether it is appropriate to begin to impute for missing values for certain variables on the SMS public use file, and, if so, to determine the "best" method to use.

Background

Within otherwise completed questionnaires, certain items may have missing responses due to refusal or insufficient knowledge on the respondent's part. If one believes that the answers for most data items would be distributed in a similar manner for respondents and nonrespondents, there is no need to be concerned about the occurrence of missing data. In this situation, the available data could be analyzed directly and reliable estimates obtained for most types of survey statistics. However, the distribution of respondent data is often very different from the nonrespondent distribution. Therefore, survey estimates obtained from respondent data will be biased with respect to describing characteristics of the survey population unless compensations are made for the missing data. For this reason, imputation procedures are often implemented to reduce the bias caused by missing data. Imputation ensures consistency between results of different analyses and enhances the ability to apply standard analysis techniques to data sets with complete data profiles without loss of sample size (Cox and Cohen, 1985).

For item nonrespondents, information often exists that can be used to predict the missing response. There are several methods commonly used to deal with item nonresponse including which are described below and will be applied to two key SMS survey variables, annual practice expenses and annual net income, in this paper.

The No Imputation Procedure

Nonrespondents are not included in the analysis and survey estimates are derived solely from respondent data. The bias in resulting survey estimates depends upon the extent of missing data and the degree to which nonrespondents as a group differ from respondents as a group.

The Mean Value Imputation Procedure

This procedure simply replaces missing data with the average value from the respondent data. The imputation may be done within imputation classes-groups of people judged to have similar responses to the data item subject to imputation. When imputation classes can be defined that are strongly related to the response being imputed, the mean estimated using the imputation-revised data should have much of the bias removed. Mean value imputation will result in overestimation of the percentage of responses falling into the middle of a distribution and will underestimate the percentage of high and low values. The effect of mean value imputation on distributional estimates severely limits its utility. This technique also is recognized as having limitations for multivariate analyses.

The Hot Deck Imputation Procedure

The hot deck imputation strategy involves partitioning the survey respondents into imputation classes. Often the records within classes are ordered using other data available that relate to response. Cross-classification and sorting are done to create a pool of donors and recipients with similar characteristics. A random donor is selected and its reported value is used to impute for the matched recipient. Variations of the hot deck procedure have been developed to reduce the use of multiple donors. The ease of use and flexibility of implementation of the hot deck technique have led to it becoming the most commonly used item non-response imputation procedure. It typically preserves the distribution of the parameter of interest as well as the covariation between variables.

The Regression or Model-Based Imputation Procedure

This approach is most appropriate for use with a quantitative variable when there are other quantitative variables that can be used to predict missing responses. It can be used when there are data items available for nonrespondents as well as respondents that can be used to model the variable for which data are missing. Regression imputation suffers from many of the same problems as mean value imputation, since the same value will be imputed for cases with missing values which have the same characteristics. Regression-based imputed data can be modified by including an estimated residual to address such problems.

Data Base Used for Imputation

The American Medical Association's (AMA) Socioeconomic Monitoring System (SMS) is a series of semi-annual telephone surveys of non-federal patient care physicians (excluding resident physicians). The spring survey collects data from approximately 4,000 physicians through an interview averaging 25 minutes in length. The autumn survey collects data from approximately 2,800 respondents through a 16-minute interview. The data used in this study are from the 1991 spring survey.

Practice expense information is requested from respondents who are full or part owners of their practices. The practice expense section typically is the most difficult part of the interview to complete, although sampled physicians are sent a list of the expense questions with the advance letter.

Item nonresponse does not appear to be problematic for many survey variables, but the income and expense questions have high nonresponse rates. For the purposes of this preliminary analysis, we will examine imputation techniques for only the annual net income from medical practice and annual

total practice expense questions. Item nonresponse rates for these questions were 23.0% and 34.7% respectively.

Implementing the Cell Mean and Hot Deck Methods

Both the cell mean and hot deck methods of imputation rely on the construction of strata to identify item respondents that have characteristics that most closely match those of item nonrespondents. Increasing the number of such stratification variables will more closely match nonrespondents to respondents, but will reduce cell sizes, thus reducing the precision of the cell mean and increasing the likelihood of encountering empty cells. The choice of stratification variables thus represents an important step in making these types of imputations.

In selecting stratification variables, we consider both the set of variables that affect the likelihood of nonresponse, and the set of variables that affect the item itself (i.e., income and expenses). The intersection of these two sets constitutes the set of variables to be used for classification.

Multivariate analyses were performed to identify significant predictors of item nonresponse. Estimated coefficients from logistic regressions of item response (dependent variable = 1 for response, 0 for nonresponse) against key physician characteristics are displayed in Table 1. In these regressions, experience and its square take the place of categories for this variable, and an estimate of annual hours worked (YRHOURS) is added. All other variables are 0-1 dummies.

Many characteristics were significant predictors of response to income, including specialty, sex, location, AMA membership, employment status, and interview type. Physicians with busier practices (as measured by YRHOURS) were somewhat less likely to report income.

Significant covariates in the logistic regression for response to total expenses were specialty, type of practice and AMA membership. Annual hours worked did not have a significant effect on response to this question.

Table 2 presents coefficient estimates from least square regressions of income and expenses (both estimated in logs). All potential explanatory variables except interview type (initial, reinterview) and major professional activity (office- or hospital-based) were significant and of the expected signs for predicting income. Similarly, the only explanatory variable not related to expenses was interview type.

The results summarized in Tables 1 and 2 were considered in selecting the stratification variables to be used in the cell mean and hot deck imputations and the sorting variables to be used in the hot deck procedure. These results led us to select different stratification and sorting variables for income and expenses. We constructed cells and calculated cell sizes and response rates within cells using different combinations of the significant predictors of item response as well as variables that affect the value of the item itself.

For income there were many significant predictors of item response that were also related to the value of the item itself. We tried various combinations of these as stratification and sorting variables and determined that using significance level of the logistic regression coefficients as the decision criterion worked well; that is the variables which were significant at the .001 level were chosen as stratification variables and the variable significant at the .01 level was used as the sorting variable.

For expenses, the choice of stratification variables was relatively straightforward since only specialty, type of practice, and AMA membership status were significant predictors of response to expenses and all were also significantly related to the value of expenses. Given the results from Table 2 showing lower item response rates for those in large group practices, we selected practice size rather than type of practice (solo vs. group) as a stratification variable.

The final choice of imputation classes for income was specialty and employment status (10 specialties, 2 types of employment status); the smallest cell with any item nonrespondents had 48 cases, and item response rates in the cells ranged from 63.7% to 88.9%. The imputation classes selected for expenses were specialty and practice size (10 specialties, 3 categories of practice size); the smallest cell with any item nonrespondents had 12 cases in it (3 with missing data for expenses) and item response rates within the cells ranged from 44.0% to 100%.

The hot deck imputation procedure substitutes data of responding individuals for the missing responses. The hot deck method utilizes two data files: a data file of item respondents (donors) and a data file of nonrespondents (recipients). These two files are merged and then sorted by the stratification variables. For each recipient within a cell, a donor was then chosen from immediately above or below the recipient in the file. With the exception of those cases where a recipient was either the first or last observation in a cell, there were two potential donors for each recipient. One donor was chosen among the two at random and the value of the response for this donor was imputed for the recipient.

The same variables used in the cell mean approach were also used to stratify the sample for the hot deck approach: specialty and employment status for income; specialty and practice size for expenses. We further sorted by sex for the income imputation and AMA membership status for the expense imputation within each cell. These additional variables were used as sorting rather than stratification variables in order to ensure adequate cell sizes. Most donors were used to impute a value for a nonrespondent only once; 4.5% of donors for income were used twice and 6.7% of expense donors were used twice -- no donors were used more than twice.

Implementing the Model-Based Method

This technique uses a regression model for item respondents to predict values for item nonrespondents. We used two variations of this technique (Model-Based I and Model-Based II). Virtually the same model as that displayed in Table 2 was used in the Model-Based I method. The only differences were that the initial interview variable was dropped and the dependent variable was not in log form. Thus, the regressors were typically dummy variables except for experience and experience squared. When negative values of income or expenses were predicted, they were set to zero.

For the second model based method (Model-Based II), two adjustments were made to the Model-Based I approach. First, to take into account the skewed nature of the distribution of physician net income and expenses, the prediction model was estimated with the dependent variable in log form. Second, imputed values were constructed as the sum of the predicted values from the model and an error term. As with the mean-based approaches, the model based approach without an error term will tend to load too many imputed values into the middle of the distribution, thus underestimating the sample variance and distorting the tails. Values for the error term were drawn from a normal random number generator with mean zero and standard deviation equal to the estimated standard deviation of the error term from the prediction model.

Results

Table 3 and 4 present resulting sample statistics and distributional characteristics from each of the imputation methods for annual net income and practice expenses, respectively.

As can be seen in Tables 3 and 4, all of the imputation approaches except the overall mean technique result in slightly higher estimates of mean net income and expenses than are obtained without imputation. This seems intuitively correct, in that we expect that nonrespondents to these items typically have higher values than respondents. For income, all approaches except for the hot deck approach and Model-Based II method lead to lower standard deviations than are obtained without imputation; for expenses, all approaches except the Model-Based II method lead to downward biased estimates of standard deviation. The different techniques lead to widely varying estimates of median income and expenses.

Both the overall mean and cell mean approaches have a leveling effect on the percentile distributions. The distributions of responses to net income following hot deck imputation and using the Model-Based II approach are generally close to the distribution without imputation, with the Model-Based II approach performing better at the extreme upper tail of the income distribution. However, for the Model-Based II approach to imputation of expenses, the 90th, 95th, and 99th percentiles appear to be biased upward.

A Simulation for Responses to Net Income

In the next stage of the analysis, a simulation was performed to assess the possible impacts that each of the imputation methods may have on several key summary statistics. The analysis was limited to assessing the impact of imputation for only the net income variable. In this simulation, we started with the sample consisting of respondents to net income. The value for net income for some respondents was then coded as missing in order to generate mock samples. Each of the imputation strategies described previously (as well as the no imputation method) was then applied to these mock samples. Sample statistics for the full set of respondents were then compared with the same sample statistics for the mock samples in order to assess the impact of each imputation method.

Experimental Design

There were 3,126 respondents to net income on the 1991 SMS survey out of a total of 4,057 potential respondents for an item response rate of just over 77%. We attempted to match synthetic item nonrespondents as closely as possible to actual nonrespondents on the SMS survey, and to construct the samples in such a way as to make the item response rate in the mock samples as close as possible to the actual response rate of 77%.

One approach to creating such samples involves choosing synthetic nonrespondents with the same measured characteristics as actual nonrespondents (Johnson and Cohen, 1990). It may be difficult to find exact matches for all nonrespondents, however, particularly when the item nonresponse rate is high. Another approach to creating mock samples is to group the full set of respondents into cells and then randomly select synthetic nonrespondents from each cell (Kalton, 1983). The probability of selection within each cell can be given by the cell item nonresponse rate for the full sample. In a sample the size of SMS, however, the number of stratification variables used to construct these cells would not need to get very large before very small, or even empty cells would be encountered.

Instead, we chose synthetic nonrespondents based on predicted probabilities from a logistic regression of item response. The steps involved were:

1) Estimate a model of response to net income for the 1991 SMS survey. The model specification is the same as that presented for income in Table 2 except the variable YRHOURS was excluded.
2) Calculate the predicted probability of response for each observation in the sample.
3) Generate a uniform (0,1) random number for each observation (U) and compare it to the probability of response (P). If $U > c*P$ then recode the response to net income as missing (synthetic nonresponse). The variable c is a constant, defined as the actual response rate in the survey divided by the mean predicted probability of response among the actual respondents to net income.

The likelihood that an observation is selected as a synthetic nonrespondent is thus proportional to the predicted probability of nonresponse. In this way, respondents to net income with measured characteristics associated with higher nonresponse are more likely to be selected as synthetic nonrespondents. The constant c is chosen to assure that the expected nonresponse rate in the mock sample is equal to the actual nonresponse rate in the full survey.

After a mock sample was constructed, each of the imputation methods described previously was applied to impute net income values for the synthetic nonrespondents. Several statistics on net income were then computed for the mock samples post-imputation, including the sample mean, median, standard deviation and selected percentiles. Due to interest in the impact that imputation might have on relationships between variables, we also performed a standard physician earnings regression with the log of hourly earnings regressed against specialty, experience, and other physician characteristics. The estimated coefficients, t-statistics and R-squares from these regressions were of particular interest.

Results--Descriptive Statistics

The process of creating a mock sample, applying the imputation methods, and computing summary statistics from the imputation-adjusted samples was repeated 250 times to generate a sample of 250 observations on each of the summary statistics of interest. The response rate and mean net income over these 250 mock samples are compared with the respective values from the actual 1991 SMS survey. (Table not presented here.)

Mean net income in the mock samples (before imputation) was nearly $4,000 lower than the actual mean of $164,300. Furthermore, the standard error of mean net income over the 250 samples was only $87, indicating that chance could account for, at most, only a small portion of this difference. The lower mean net income in the mock samples is consistent with the notion that physicians with higher than average incomes are less likely to respond to net income. The mean response rate in the mock samples, was in fact very close to the actual response rate of 77.05%.

Summary statistics for the mock samples with the imputation methods are displayed in Table 5. The means for each of these statistics are compared with their actual values from the 1991 SMS survey. Considering the sample mean first, it is not surprising that the no imputation approach and overall mean approach fail to account for the pattern of selectivity in the sample and, as a result, underestimate the actual mean. The remaining four methods provide more accurate estimates of the sample mean; for all approaches (but the Model-Based II) the mean from the actual SMS sample is within a 95% confidence interval of mean income over the mock samples.

All methods (except no imputation) appear to result in upward biased estimates of median income, but the magnitude of the bias differs greatly by method. The no imputation method resulted in a downward biased estimate of median income. The estimated bias ranged from only 0.4% for the hot deck method,

to almost 22% with the overall mean approach. Furthermore, only the hot deck method and the model-based II method resulted in standard deviations that were close to the actual value--the standard deviation of net income was biased downward for all other methods by as much as 15%. There are techniques that can be used to correct variance estimates but we assume that many users of our data set would not be familiar with them (Little and Rubin, 1987).

Examining the distribution of net income with the application of each imputation method, it is clear that the overall and cell mean approaches tend to level the distribution. The model based approach without an error term preserves the lower tail of the income distribution relatively well but apparently fails to predict enough large values for net income. The result is downward biased estimates of the higher percentiles with the magnitude of the bias increasing when moving further out into the tail. The model based approach with an error term and the hot deck method perform very well relative to the other imputation strategies, generating samples with distributions very close to the distribution of actual responses.

Results--Regression Estimates

Beyond simple descriptive statistics, the simulation was also geared to finding the effects of each of the imputation techniques on more complex relationships between variables. Given the importance of linear regression as an analytical technique to users of the SMS public use file, a simple log of hourly earnings regression model was estimated both with and without imputation.

Table 6 shows the coefficient estimates from the actual 1991 SMS survey, along with the means (over the 250 mock samples) of regression coefficients after each of the imputation methods was employed. As with the descriptive statistics displayed in Table 5, the coefficients from the actual 1991 survey are considered the "true" or ideal values.

With the no imputation method, means of all coefficients are very close to those from the actual sample. Note that with the overall mean imputation method, the means of all coefficients except the intercept are smaller in absolute value than those from the actual sample. The overall mean approach would appear to result in coefficient estimates that are biased toward zero.

At first glance, the cell mean approach appears to result in estimates that are not systematically different from their actual values. However, with only two exceptions, the mean estimated coefficients for the stratification variables (SELFEMPL and specialty) are found to be larger than the actual coefficients, while the mean estimated coefficients for all non-stratification variables are lower than the actual values. Thus, in this case, cell mean imputation resulted in upward biased (in absolute value) coefficient estimates for stratification variables, and downward biased estimates for all other variables. Evidence of a similar pattern of bias also exists for the hot deck method but the pattern is not as strong.

We found no obvious patterns of bias for the model based approaches. The model-based II approach generally resulted in coefficients that were closer to the actual values. However, essentially the same model was used to impute values using the model based approaches as was used to test the impact of imputation on regression results. It will be important to estimate other types of models in assessing the impact of imputation before the model based approach can be endorsed as not distorting regression results.

In addition to the magnitude of the regression coefficients, qualitative results--the signs and significance of the estimated coefficients--are also of interest. Of particular interest are cases where a variable is insignificant in the true sample, but is significant in a mock sample after imputation (Type I error); and where a variable is significant in the true sample, but is insignificant in a mock sample after imputation (Type II error).

Most variables in the log hourly earnings regression are highly significant and there are only a few variables where Type I or II errors occur. There was never a significant reversal in sign, e.g., a variable being negative and significant in the true sample and positive and significant in the mock sample after imputation. The hot deck method performed worst of all methods, as far as the number of both Type I and Type II errors. The model-based approaches were slightly better than the cell mean approach.

Finally, we considered the impact of imputation on goodness of fit of the estimated log earnings model. The R^2 for the model estimated on the actual SMS sample was 0.3098. Over all 250 mock samples, the mean value of the R^2 was lower than this when using all approaches except the model-based I approach. In fact, the minimum R^2 obtained from the model-based I approach was larger than the actual R^2. The average R^2 was particularly low for the overall mean and hot deck approaches.

Conclusion

The results presented here indicate that imputation may be appropriate in the SMS survey, at least for variables with high item nonresponse. The choice of the best technique, however, is not clear-cut. Either the hot-deck approach or the model-based approach with an error term appear to generate reasonable sample statistics and distributional estimates, but do not work as well as the "no-imputation" approach when complex relationships between variables are examined. Perhaps the best solution will be to flag imputed values on the data set and warn the users that imputed values should only be used in univariate analyses.

Further analysis is necessary to determine the appropriate imputation strategy for other variables on the data file. The set of variables will probably be limited to the few with extremely low response rates so that the optimum imputation technique can be used for each variable, rather than following the usual approach of using one imputation strategy for all variables.

The authors gratefully acknowledge the comments made by Marc Berk, Director, Project HOPE Center for Health Affairs.

REFERENCES

Cox, Brenda and Steven Cohen. Methodological Issues for Health Care Surveys. Marcel Dekker, Inc. New York. 1985.

Kalton, Graham. Compensating for Missing Survey Data. Institute for Social Research, Ann Arbor, Michigan. 1983.

Johnson, Ayah and Steven Cohen. "Assessing Quality of Imputation Techniques for the National Medical Expenditure Survey." Paper presented at the winter meeting of the American Statistical Association, January 1990.

Little, Roderick and Donald Rubin. Statistical Analysis with Missing Data. John Wiley and Sons, New York, 1987.

TABLE 1

Logistic Regression on the Probability of Responding to Selected Items

Coefficient	Income	Total Expenses
INTERCEPT	2.6695	0.1988
IM	-0.1491	-0.1151
SUR	-0.4416**	-0.0149
PED	0.3875+	0.2465
OB/GYN	-0.2186	-0.3701*
RAD	-0.8621***	-0.1983
PSYCH	0.6034**	0.9114***
ANES	-0.1611	0.1918
PATH	0.0607	0.5859+
OTHER	0.3556+	0.7218***
FEMALE	-0.3851**	-0.2352
SOLO	-0.1544	0.7714***
HOSPBASE	0.1114	0.2080
EXPERIENCE	-0.0180	-0.0086
EXPERIENCE^2	0.0002	0.0000
RURAL	0.2567*	0.0454
SMALLMET	-0.1354	-0.1100
INITIAL	-0.4717***	-0.1175
AMA	0.1721*	0.1939*
CERT	-0.0813	0.0250
YRHOURS	-0.0001+	0.0000
SELFEMPL	-0.4672***	NA
Sample Size	3823	2541
Log-likelihood	-1839	-1615

\+ p < 0.10
* p < 0.05
** p < 0.01
*** p < 0.001

SOURCE: American Medical Association Socioeconomic Monitoring System, 1991 Spring Survey.

TABLE 2

Regression Results

Coefficient	Log (Income)	Log (Total Expenses)
INTERCEPT	10.6271	11.0644
IM	0.2225***	-0.0558
SUR	0.5730***	0.2772***
PED	0.0219	0.0361
OB/GYN	0.5693***	0.3263***
RAD	0.6184***	-0.5649***
PSYCH	0.1720***	-1.0425***
ANES	0.6839***	-0.6656***
PATH	0.4047***	-0.5410***
OTHER	0.3252***	-0.4894***
FEMALE	-0.3381***	-0.2380***
SOLO	-0.2221***	-0.1391***
HOSPBASE	-0.0161	-0.3312***
EXPERIENCE	0.0584***	0.0506***
$EXPERIENCE^2$	-0.0013***	-0.0010***
SMALLMET	0.0787**	0.0685
LARGEMET	0.0718**	0.1076*
INITIAL	-0.0162	-0.0081
AMA	0.1301***	0.1240**
CERT	0.1859***	0.1110*
SELFEMPL	0.3128***	NA
Sample size	3123	1750
Adjusted R-Square	0.39	0.26

* p < 0.05
** p < 0.01
*** p < 0.001

SOURCE: American Medical Association Socioeconomic Monitoring System, 1991 Spring Survey.

TABLE 3

Comparison of Imputation Methods

Annual Net Income

	No Imputation	Overall Mean	Cell Mean	Hot Deck	Model-Based I [a]	Model-Based II[b]
Percentile						
1st	20,000	23,000	23,000	19,000	20,000	21,000
5th	45,000	50,000	50,000	45,000	46,000	44,800
10th	60,000	68,000	68,000	60,000	65,000	60,000
90th	300,000	277,000	282,400	304,000	285,000	313,000
95th	380,000	350,000	350,000	400,000	350,000	400,000
99th	614,000	560,000	560,000	693,000	560,000	630,000
N	3126	4057	4057	4057	4057	4057
Mean	164,300	164,300	168,000	169,300	167,800	168,400
Standard Deviation	120,340	105,630	109,525	125,943	110,679	122,067
Median	132,000	164,300	148,000	139,000	149,000	137,600

All statistics are weighted to adjust for survey nonresponse.
[a]With no added residual.
[b]With added residual.

SOURCE: American Medical Association Socioeconomic Monitoring System, 1991 Spring Survey.

TABLE 4

Comparison of Imputation Methods

Annual Practice Expenses

	No Imputation	Overall Mean	Cell Mean	Hot Deck	Model-Based I[a]	Model-Based II[b]
Percentile						
1st	2,000	5,000	5,000	3,000	5,000	5,000
5th	17,000	25,000	25,000	18,000	24,400	21,000
10th	31,000	44,000	42,900	32,000	43,000	32,100
90th	293,000	241,000	241,000	293,000	250,000	303,000
95th	400,000	325,000	325,000	400,000	325,000	420,000
99th	662,000	607,000	607,000	657,000	607,000	704,200
N	1763	2698	2695	2695	2698	2698
Mean	150,000	150,000	153,800	150,500	153,100	155,800
Standard Deviation	134,073	108,368	112,030	129,442	112,259	146,924
Median	119,000	150,000	132,000	120,000	137,000	118,300

All statistics are weighted to adjust for survey nonresponse.
[a]With no added residual.
[b]With added residual.

SOURCE: American Medical Association Socioeconomic Monitoring System, 1991 Spring Survey.

TABLE 5

Simulation Results

Means of Descriptive Statistics*

Sample Statistic	Actual Sample	Imputation Method: No Imputation	Overall Mean	Cell Mean	Hot Deck	Model Based I[a]	Model Based II[b]
Mean	$164,300	$160,739	$160,739	$164,396	$164,499	$164,279	$164,027
		(87)	(87)	(81)	(113)	(82)	(97)
Median	132,000	129,976	160,684	140,434	132,564	142,606	132,566
		(20)	(86)	(73)	(142)	(124)	(123)
Standard Deviation	120,340	116,560	102,322	106,261	120,141	107,128	118,619
		(172)	(153)	(154)	(250)	(152)	(194)
Percentiles							
1st	20,000	19,896	23,676	23,676	19,816	20,487	22,178
		(58)	(89)	(89)	(105)	(86)	(91)
5th	45,000	44,352	49,684	49,684	44,396	46,344	44,678
		(46)	(44)	(44)	(61)	(79)	(38)
10th	60,000	60,000	67,560	67,560	60,036	64,285	59,991
		(0)	(71)	(71)	(16)	(66)	(11)
90th	300,000	298,840	266,752	276,214	300,548	280,439	301,015
		(202)	(345)	(292)	(135)	(139)	(151)
95th	380,000	375,102	338,424	338,424	384,184	338,424	384,464
		(395)	(494)	(494)	(567)	(494)	(483)
99th	614,000	598,612	548,472	548,472	614,812	548,472	605,604
		(1321)	(1113)	(1133)	(2435)	(1133)	(1342)

*Based on 250 repetitions. Standard errors are in parentheses.
[a]With no added residual.
[b]With added residual.

SOURCE: American Medical Association Socioeconomic Monitoring System, 1991 Spring survey.

TABLE 6

Simulation Results
Means of Regression Coefficients*

Variable	Label	Actual	Imputation Method: No Imputation	Overall Mean	Cell Mean	Hot Deck	Model Based I[a]	Model Based II[b]
INTERCEP	Intercept	2.93472	2.94205	3.20346	3.06513	3.06540	2.92027	2.94402
SELFEMPL	SELF-EMPLOYED	0.19297	0.19089	0.13552	0.20801	0.18866	0.20542	0.19230
SPDUM2	INT MED	0.19181	0.19134	0.15185	0.23586	0.21368	0.22013	0.19296
SPDUM3	SURGERY	0.67492	0.67394	0.52833	0.72814	0.70374	0.70569	0.67253
SPDUM4	PEDIATRICS	0.07493	0.08018	0.04272	0.07379	0.09483	0.06496	0.07175
SPDUM5	OB/GYNE	0.51407	0.51334	0.38641	0.55386	0.53200	0.54823	0.52167
SPDUM6	RADIOLOGY	0.67864	0.67371	0.53270	0.73739	0.73347	0.70742	0.68012
SPDUM7	PSYCHIATRY	0.37353	0.37318	0.30343	0.35658	0.37414	0.36343	0.37298
SPDUM8	ANESTHESIOLOGY	0.62083	0.61781	0.47405	0.63628	0.64898	0.62235	0.61877
SPDUM9	PATHOLOGY	0.59273	0.59309	0.50232	0.62791	0.64179	0.60022	0.58855
SPDUM10	OTHER	0.38601	0.38911	0.30068	0.39719	0.40979	0.38890	0.38801
SMALLMET	METRO, <1M	0.07488	0.07122	0.06629	0.06411	0.05546	0.09124	0.07053
LARGEMET	METRO, 1M OR MORE	0.09359	0.09158	0.08243	0.07901	0.07009	0.10072	0.08724
AMA	AMA MEMBER	0.09718	0.09331	0.06289	0.06220	0.06424	0.09910	0.09496
EXPER	YRS SINCE MD GRAD	0.04218	0.04176	0.03034	0.03040	0.03099	0.04271	0.04185
EXPER2	YRS SINCE MD GRAD**2	-0.00087	-0.00085	-0.00060	-0.00060	-0.00062	-0.00088	-0.00086
SOLO	SOLO OR OTHER TYPE OF PRACTICE	-0.28268	-0.27775	-0.21150	-0.21634	-0.23098	-0.25379	-0.27454
CERT	BOARD CERTIFIED	0.15524	0.15205	0.11469	0.11299	0.10917	0.15836	0.15470
FEMALE	FEMALE PHYSICIAN	-0.14382	-0.13977	-0.06421	-0.06978	-0.15075	-0.15046	-0.14968
HOSPBASE	HOSPITAL BASED MD	-0.07742	-0.07490	-0.07289	-0.06922	-0.06864	-0.08143	-0.07593

*Based on 250 repetitions.
[a]With no added residual.
[b]With added residual.