A COMPARISON OF METHODS FOR IMPUTING MISSING MEDICAL EXPENDITURE DATA IN THE NATIONAL MEDICAL EXPENDITURE SURVEY

Jill J. Braden, John P. Sommers, Agency for Health Care Policy and Research
Jill J. Braden, AHCPR, 2101 E. Jefferson St. Suite 500, Rockville, MD 20852

KEY WORDS: Item Nonresponse, Hot Deck, Regression Models

Introduction

Imputed values are often supplied for missing item values in large surveys. Imputation makes a data set easier to analyze because many standard statistical techniques and packages require rectangular data sets. Imputation also ensures consistency between the results from different analyses. Finally, it can reduce the nonresponse bias resulting from item nonresponse.

There are a number of different strategies for dealing with item nonresponse. Mean imputation, cold-deck imputation, various hot-deck imputation procedures and regression imputation are some of the approaches used to replace missing data. (Kalton and Kasprzyk, 1986) This paper compares the weighted sequential hot-deck imputation method for imputing two categories of hospital-related expenditures, inpatient stays and emergency room visits, with alternative procedures based on regression modeling.

The data for this investigation are from the 1987 National Medical Expenditure Survey (NMES II). The Household Survey is representative of the civilian noninstitutionalized population of the United States in 1987. Each family was interviewed five times between February 1987 and July 1988. The stratified multi-stage area probability design yielded approximately 34,500 individuals in roughly 14,000 families who completed all rounds of data collection. The survey collected information on illnesses, use of and charges for health services, health insurance coverage, employment, income and other related characteristics. Details of the household sample design can be found in Cohen, et al. (1991)

Household respondents are unable to provide accurate expenditure amounts for many types of health care utilization. The Medical Provider Survey (MPS) is the component of the NMES II that obtained information from the physicians, hospitals, outpatient clinics, emergency rooms and home health agencies used by the household sample. All medical providers who provided care in a hospital or in a clinic setting, all providers of home health care, and all providers reported by households with at least one member eligible for Medicaid were to be included. All medical providers, including those associated with ambulatory office-based care, were added for a nationally representative 25 percent sample of households. (Tourangeau and Ward, 1992)

MPS data, when available, were used to construct the expenditures for an event. In the absence of MPS data, household-reported data were used. The combined response rate using these two sources was 67 percent for inpatient hospital stays. For emergency room visits the combined response rate was 57.5 percent.

This paper is part of a larger study to determine the effects of imputation on NMES II estimates and variances. Nonresponse to inpatient and emergency room expenditures provided an interesting situation for evaluating imputation methods. First, these expenditures represented about 45 percent of the total expenditures for health care and the levels of nonresponse, as presented above, are relatively high. Second, the data sets contained numerous sparse donor cells when the hot-deck imputation method was implemented for these expenditures in NMES II. Collapsing across cells may compromise the accuracy of the imputation due to the loss of detail from omission of some variables at the level where a donor is found.

Estimation of Variances

In order to estimate variances for the relative standard errors presented in the tables, adjusted balanced repeated replication (BRR) methods which accounted for the effects of imputation were used. We propose that standard unadjusted BRR and Taylor Series techniques may be significantly biased by the imputation associated with high levels of item nonresponse. Methods such as adjusted BRR break the variance of a sample estimate E into two components:

$$Var[E] = Var_S[E_I[E|S]] + E_s[Var_I[E|S]]$$

where I is ith imputation and S is the sampling process.

left term is the variance over all samples of the expected value of E over the imputation process given the sample selected. The right term is the expected value of the variance of the value of E for a fixed sample given the sample selected.

To estimate the Var[E], adjusted replication methods adjust the values of imputed records by the difference of the expected value of the imputation for the replicate versus full sample. Inclusion of this extra variability allows one to account for the right side of the equation. Examples of this technique can be seen in Rao, 1994; Shao, 1994; and Rao and Shao, 1992.

Weighted Sequential Hot-Deck Imputation of Missing Hospital-Related Expenditures

A weighted sequential hot-deck approach (Cox, 1980) was used to impute expenditures for non-pregnancy-related hospital inpatient stays and for non-pregnancy-related emergency room visits. The donor sets were comprised of both household and MPS data.

Inpatient Hospital Stays

The classification scheme for inpatient stays initially included indicators of:

- probability of the event being a "true" hospital stay
- surgery being performed during the stay
- length of stay (0-1, 2-5, 6-20, 21+ nights)
- number of doctors seen (0-2, 3+)
- MSA status (large, small or non-MSA)

Cells with meager donor-recipient ratios were collapsed.

Regression of these variables on reported inpatient hospital expenditures yielded an R^2 of .175. The results of this hot-deck imputation strategy are presented in Table 1.

Emergency Room Visits

The classification scheme for emergency room visits initially included indicators of:

- performance of X-rays
- performance of scans or imaging
- performance of surgery
- MSA status
- region

Some collapsing of cells to achieve acceptable donor-recipient ratios was performed.

Regression of these variables on reported hospital emergency room expenditures yielded an R^2 of .036 when both MPS and Household data were used. The hot-deck results are presented in Table 2.

Two Regression Models for Imputing Missing Hospital-Related Expenditures

One advantage of a modeling procedure for imputation is that many main effects can be included simultaneously, compared with the limitation and loss of relevant variables with constructing and collapsing of cells for hot-deck imputation. Predictions should be more accurate if relevant variables have been excluded from the hot-deck. The assumption that the nonresponse mechanism is ignorable is also more feasible if all relevant effects are included in the model.

Inpatient Hospital Stays

The models for inpatient hospital expenditures were fitted first on 2,894 cases, consisting of MPS and household reported expenditure data. The models include the following variables as predictors of the expenditure.

- probability of the event being a "true" hospital stay
- surgery being performed during the stay
- length of stay (0-1, 2-5, 6-20, 21+ nights)
- number of doctors seen (0-2, 3+)
- MSA status (large, small or non-MSA)
- family as a source of payment
- government as a source of payment
- age (0-6, 7-18, 19+)
- condition categories
- region
- sex
- link to an emergency room visit
- death preceding discharge

The variables were represented in the regression by dummy variables with one category omitted from each set to avoid collinearities. The first model attempted to predict the per diem expense. The R^2 for this regression model was .198.

The second model used the logarithm of the per diem expenditure as the dependent variable. The R^2 for this approach was .280. This improvement may be offset by the fact that the log model is, by default, biased.

The resulting inpatient per diem expenditure equations were used to impute the missing expenditure values for the length of the hospital stay. The results are shown in Table 1.

Emergency Room Visits

The models for emergency room expenditures were fitted on 4,591 cases, consisting of MPS and household reported expenditure data. The models include the following variables as predictors of the expenditure.

- performance of X-rays
- performance of scans or imaging

- performance of surgery
- MSA status
- region
- sources of payment
- age (0-64, 65+)
- ADL status (0-2, 3+)
- any coverage through Medicare in 1987
- any coverage through Medicaid in 1987
- link to an ambulatory physician visit
- health status (poor, fair-excellent)

The first model attempted to predict the emergency room expenditure. The R^2 for this regression was .052. The second model used the logarithm of the expenditure as the dependent variable. The R^2 for this approach was .192.

The resulting emergency room expenditure equations were used to impute the missing expenditure values. The results are shown in Table 2.

Addition of Residuals to the Regression Models

The models discussed above impute, for the missing expenditure amounts, a mean of the predictive distribution, conditional on the included predictors. As a result, the distribution of the imputed values has a smaller variance than the distribution of the true values. This characteristic affects the quality of subsequent analyses. One strategy to adjust for this attenuation is the addition of random errors to the predicted means. For our work, we used random gamma and normal deviates as well as randomly selected residuals from the model. Although we know of no cases where gammas have been used, the nature of the data suggested this distribution as one logical choice.

Inpatient Hospital Stays -- the per diem model

For the model which directly estimates the per diem expenditures for inpatient hospital stays, the pattern of the distribution of the standardized residuals was skewed to the right. This distribution appeared more adequately modeled by a gamma-type function than the normal. The following strategy was used to add the residuals to this model.

Using the size of the predicted value for each respondent, the respondent set was sorted divided into five sets of approximately equal size, such that the ith set (i = 1,2,3,4,5) contains quintiles of the data. For each of the quintiles, we calculated std_i and μ_i the standard deviation and mean of the residuals and then calculated $a_i = \mu_i - k_i$ where k_i is the smallest residual in the ith quintile set. For each quintile we let $\beta_i = (std_i)^2/a_i$ and $\alpha_i = (a_i/std_i)^2$. For each nonrespondent, we generated a random gamma variable γ (Hogg and Craig, 1970), where the predicted value in the ith quintile. The imputed expenditure for nonrespondents is $p_0 + (\gamma + k_i)*\beta_i$.

The movement by the lowest value, k_i was required since the mean of the residuals is zero and thus the errors must be estimated by a translated gamma to achieve this characteristic. The division of the residuals into quintiles was performed because the average size of the error increased with the size of the prediction.

The results of this strategy are shown on Table 1. Although the gamma distribution was more appropriate for the residuals than a normal distribution, it still produced some large negative and large positive estimates of per diem expenditures. We attempted, unsuccessfully, to correct these problems by using truncated gammas. This truncation and the other truncation needed to force charges to be non-negative, produced shifts in the estimates of totals and/or adverse effects on the variance of the estimates.

A second approach was developed to deal with the failure to mimic the underlying distribution of the residuals. We allowed the actual empirical distribution of the error to dictate the distribution. Residuals were added to the predicted values by classifying the respondents and nonrespondents into quintiles based on the predicted per diem expense. Within quintiles, residuals were assigned to nonrespondents from respondents in the same quintile. Residuals were left in random order and donor residuals were randomly selected for recipients by systematic sampling of these residuals. The results as shown in Table 1 are closer to the hot-deck estimated total expenditures than results from the first approach.

Inpatient Hospital Stays -- the log model

The increase in the average size of the error as the size of the prediction increased indicated that a logarithmic model, being multiplicative in nature, might perform better. When modeling the logarithm of the per diem expenditure, the first approach taken to add residuals to the model was as follows.

We again constructed quintiles and calculate the mean and standard deviation (std_i) for each quintile from the respondents. For each nonrespondent, we estimated the logarithm of the expenditure from the regression on the logs. The estimated per diem then equaled $e^{(\ln(est) + r)}$ where r is a random number from the normal distribution with mean equal to zero and standard deviation equal to std_i.

Results are displayed in Table 1. This method performed somewhat poorly, mainly because too much noise was added to the large values of expenditures by the assumption of normally distributed residuals,

producing what we felt was an unacceptably high estimate of total expenditures. This method also produced the highest estimate of variance.

Because the normality assumption in the first strategy appears unjustified, we again used an empirical residual method. This second method to add residuals to the predicted values involved classifying respondents and nonrespondents into quintiles based on the predicted logarithm of the per diem amounts. Then residuals were assigned to nonrespondents from respondents in the same quintile. Again, donors were selected for recipients by systematic sampling of randomly ordered residuals.

See Table 1 for the results from this method which yielded an estimate somewhat more in line with the hot-deck and linear prediction models and also had less added variance from the addition of the residuals.

Emergency Room Visits -- the linear model

The same strategies described above for the inpatient hospital per diem model were used for this application; first, a gamma-based error was introduced and second, a respondent-based residual was added to the predicted value. The results are presented in Table 2.

The weakness which were hinted at with the imputation for inpatient hospital expenditures are more obvious with the imputations for emergency room visits. The R^2 value was essentially zero for both the hot-deck and the linear model.

The approach which introduced a random gamma error increased the estimated total expenditures for emergency room visits by about 25 percent over the hot-deck estimate, which is both statistically and realistically significant.

The method which added residuals from respondents to the predicted values of the nonrespondents produced an estimate of the total which is closer to the hot-deck estimate, but given the inadequacy of prediction for the models, it is difficult to ascribe meaning to this outcome.

Emergency Room Visits -- the log model

The same approaches taken for log modeling of inpatient hospital stays were used for this application; first, a random normal error was introduced and second, a respondent-based residual was added to the predicted value. The results are presented in Table 2.

The R^2 for the log model was .2, an improvement over the R^2 values for both the hot-deck and the linear model. The approach which introduced a random normal error increased the estimated total expenditures for emergency room visits much less drastically than the linear model with the assumption of a gamma distribution underlying the residuals.

The method which added residuals from respondents to the predicted values of the nonrespondents produced an estimate of the total which is closer to the hot-deck estimate. The inadequacy of prediction for the model for the hot-deck, makes it difficult to come to a conclusion about this result. The improved R^2 for the log model yielded smaller RSEs in all three instances where it was used, without and with the addition of residuals compared to analogous runs with the linear models.

Observations

Because we are in the early stages with this evaluation, it seems that observations are more appropriate than conclusions.

- Erratic results are produced if the models do not have good predictive power. Better models produce more stable results, regardless of the imputation strategy used.
- Some models are known to be biased, like the log model, but the collapsing of hot-deck cells can also create some empirical bias. Collapsing cells assumes the same mean prediction for two cells which were determined by the model as having significantly different values.
- Error terms are tricky. Reliance on the assumption of random normal deviates, for example, may add far too much noise to the predicted values. We need to work on controlling this tendency.
- More structured analysis needs to be performed. Because the results from a single data set produce essentially a single data point for a given variable, it may be informative to construct a number of simulated data sets with differing characteristics in order to adequately investigate these imputation issues.

It is difficult to make a firm recommendation for the imputation method to be used. If the linear model used to create the hot-deck cells is adequate and the percent of the sample to be imputed in collapsed strata is small, hot-deck imputation appears to make the best use of residual empirical means and distributional structure.

Absent such a situation, as occurred with our emergency room data, use of regression imputation adding error from the empirical data could be considered. If equally

good, the linear model is to be preferred over the log model.

Finally, if the level of item nonresponse and, therefore, of imputation is high, imputation might be performed by more than one technique, with totals and variances estimated for each method. The distribution of the data should also be examined. These results should, then, provide the basis for a final, judicious choice.

Table 1 Comparison of imputation techniques for missing inpatient hospital expenditures.

Method	Estimated total (in billions)	Relative standard error percent
Hot-deck	137.35	3.66
Regression model prediction	137.67	3.57
Log Regression model prediction	131.85	3.51
Regression with gamma error	138.86	3.66
Regression with respondent residuals assigned to nonrespondents within quintiles	138.19	3.86
Log Regression with normal error	140.04	4.42
Log Regression with respondent residuals assigned to nonrespondents within quintiles	139.66	3.91

Table 2 Comparison of imputation techniques for missing emergency room expenditures

Method	Estimated total (in billions)	Relative standard error percent
Hot-deck	8.12	3.78
Regression model prediction	8.49	4.10
Log Regression model prediction	8.28	3.04
Regression with gamma error	10.17	3.89
Regression with respondent residuals assigned to nonrespondents within quintiles	8.90	4.83
Log Regression with normal error	8.93	3.04
Log Regression with respondent residuals assigned to nonrespondents within quintiles	8.58	3.95

References

Cohen, S.B., DiGaetano, R. and Waksberg, J. (1991) *National Medical Expenditure Survey: Sample Design of the 1987 Household Survey.* AHCPR Pub. No. 91-0037, U.S. Department of Health and Human Services, Rockville, MD.

Cox, B.G. (1990) The weighted sequential hot deck imputation procedure. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp 721-726.

Hogg, R.V. and Craig, A.T. (1970) Introduction to Mathematical Statistics, Third Edition, Macmillan Co.. New York, NY. p.101.

Kalton, G. and Kasprzyk, D. (1986) The Treatment of Missing Survey Data. *Survey Methodology*, 12, pp 1-16.

Rao, J.N.K. (1994) Resampling Methods for Complex Surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Rao, J.N.K. and Shao, J. (1992) Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. *Biometrika*, 79:4, pp 811-822.

Shao, J. (1994) Balanced Repeated Replication. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Tourangeau, K. and Ward, E. (1992) *Questionnaires and Data Collection Methods for the Medical Provider Survey.* AHCPR Pub. No. 92-0042, U.S. Department of Health and Human Services, Rockville, MD