Anomalies under Jackknife Variance Estimation Incorporating Rao-Shao Adjustment in the Medical Expenditure Panel Survey - Insurance Component[1]

Robert M. Baskin[1], Matthew S. Thompson[2]
[1] Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850
[2]U.S. Census Bureau, 4600 Silver Hill Rd, Suitland, MD 20746

Abstract

The Medical Expenditure Panel Survey – Insurance Component (MEPS-IC) is an annual, establishment survey that collects data on health insurance options made available to employees. The MEPS-IC imputes data to account for item nonresponse but currently published estimates of variance do not account for the variance due to imputation. Previous research on the impact of imputation on the variance has been conducted but found some unexpected results. In the previous research three different variance estimation methods were used: standard jackknife estimation, jackknife estimation with re-imputed replicates, and jackknife estimation with adjustments made to imputed values. For continuous type variables the previous research found that the baseline estimates were within 10 percent of the adjusted estimates but for a proportion in large establishments the adjusted standard error was twice as large as the re-imputed standard error. It was also noted that in this case the adjusted replicate estimates of proportions could be negative or greater than one. The goal of this research is to attempt to determine by simulation the source of this unexpected result. Preliminary results indicate that for binomial events the adjusted jackknife estimator produces estimates twice as large as the re-imputed jackknife estimator about ten percent of the time but on average across repeated samples the adjusted jackknife estimator is unbiased.

Key Words: Establishment Survey, Variance Estimation, Jackknife, Missing Data

1. Introduction

The Medical Expenditure Panel Survey – Insurance Component (MEPS-IC) is an annual survey of business establishments and governments sponsored by the Agency for Healthcare Research and Quality (AHRQ) and conducted by the U.S Census Bureau. MEPS-IC employs a complex survey design which is a stratified single stage design with unequal probability selection within strata. The data is collected through a mail survey with telephone follow-up.

The purpose of the MEPS-IC is to collect data on employer sponsored health insurance and information about firms that offer employer sponsored health insurance. Information

[1] **Disclaimer:** This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau, the Agency for Healthcare Research and Quality, or the Department of Health and Human Services.

collected includes firm characteristics; whether the firm offers employer sponsored health insurance; if offered the types of plans, coverage and costs; and if offered the uptake of insurance by employees.

As with any survey, missing data, whether from unit nonresponse or item nonresponse, is a concern. In order to correct for unit non-response, the MEPS-IC employs a reweighting procedure that adjusts the weights of the responding establishments to account for any non-responding units. This paper, however, focuses specifically on item nonresponse and its impact on variance estimation for the estimates produced by the MEPS-IC.

Item nonresponse occurs when an establishment returns the survey form with enough questionnaire items answered to be deemed a respondent but still leaves one or more items blank. This can obviously cause problems when an item left blank is required for estimation. When this occurs, the survey uses a nearest neighbor hot-deck imputation procedure to obtain values for these blank items.

To begin the imputation process, donors and recipients are identified for each item needed for estimation. Donors are establishments that have reported valid responses for the item in question, whereas recipients are those establishments that did not have valid responses to the questionnaire item, whether due to nonresponse or an erroneous response. The imputation procedure then selects for each recipient a donor establishment with similar characteristics. Once selected, the data from the donor establishment can either be directly substituted into the blank item or used in combination with reported data from the recipient to fill in the blank item.

For example, survey respondents often report offering both individual and family health insurance coverage. If, in this instance, an individual plan deductible is reported but no family deductible is provided, the family deductible will not be imputed by a direct substitution from the donor. Instead, the donor will provide a ratio of family deductible to individual deductible which will then be multiplied by the recipient's individual deductible in order to obtain a value for family deductible. In this way, the missing item can be filled in using a response from the recipient as a starting point.

1.1 Previous Research

As explained in Rao and Shao (1992) the process of imputation causes variance estimators that treat the imputed values as they are observed to underestimate the variance. Thompson and Kearney (2010) conducted research on estimating the variance of variables from MEPS-IC both ignoring the imputation and accounting for the imputation. Jackknife variance estimation was used in three ways. A *naive* estimate that ignored the imputation was used as a baseline. A method of reimputing at the replicate level due to Burns (1990), called here the *Burns method*, was employed. And the method of *Rao and Shao* that adjusts the replicates for imputed values was also calculated. These three methods were applied to two continuous type measures, Average Total Family Premium and Average Employee Contribution to an Individual Coverage Plan, and to two proportions, Proportion of Employees Enrolled in Individual Coverage and Proportion of Employees Enrolled in Insurance.

The theoretical work in Rao and Shao (1992) showed that the naïve estimate underestimates the variance in the presence of imputation, the Burns method overestimates the variance, at least asymptotically, and the Rao-Shao method produces a consistent estimator of variance.

In the work by Thompson and Kearney (2010) the continuous type measures reflected the results in Rao-Shao (1992) exactly as expected. The naïve estimates of standard error for average total family premium and for average employee contribution to an individual coverage plan were 5% to 10% lower than the Rao-Shao adjusted standard errors while the Burns method reimputed standard errors were 70% larger than the Rao-Shao adjusted standard errors. When subset by single versus multi unit the Rao-Shao adjusted standard error estimates were 9%-12% larger than the naïve estimates while the Burns method standard errors could be slightly smaller than the Rao-Sao estimates to almost double in size.

The results for proportions however, produced some anomalies. For the proportion of employees enrolled in an individual coverage plan the Rao-Shao adjusted standard error is four times as big as the naïve estimate and twice as large as the Burns method standard error. For the proportion of employees enrolled in any plan the Rao-Shao adjusted standard error is 70% larger than the naïve estimate but the Burns method standard error is more than twice as large as the Rao-Shao adjusted standard error. These two results for proportions seem inconsistent between the Rao-Shao- method and the Burns method. There were other cases involving estimates by class variables in which the Rao-Shao method estimate exceeded the Burns method estimates, sometimes by almost double. For Multi-unit establishments, the Rao-Shao adjusted standard was 65% and 88% larger than the Burns method standard error for proportion of employees enrolled in an individual plan and proportion of employees enrolled in any plan, respectively. The sometimes huge departures of the Rao-Shao method estimates above both the naïve estimates and the Burns method estimates was unexpected and some investigation of the construction of the estimates was conducted. It was noticed that in cases for proportions in which the Rao-Shao method estimates were extremely large relative to the other estimates, that the adjustment at the replicate level could produce replicate level proportion estimates that were negative.

Since proportions cannot be negative this raised a question as to whether the adjustment was being done correctly or if it was done correctly should the adjustment for proportions be truncated at zero to prevent impossible proportions at the replicate level. Truncating the proportions reduced the adjusted standard errors but could possibly introduce a bias.

1.2 Purpose of Current Research

The unexpected issues with proportions raised questions as to whether the adjustments should be truncated and to the possible cause of the large adjusted standard errors. The current research will attempt to address these issues through a simulation in which a gold standard is generated so that the true effect of the negative replicates can be judged.

2. Analysis Plan

2.1 Simulation

Since the issues in the previous research arose with proportions then only count data were needed in the simulation. Several small pseudo-populations were created with an increasing number of strata running from 10 to 40. In each stratum there were fifty establishments. Each establishment had two counts for employees generated, one count representing employees taking insurance and the other count representing employees not

taking insurance, so the total number of employees would be the sum of those two numbers. The count of employees taking insurance was produced as a random Poisson count multiplied by a constant that depended on the stratum. This provided some counts of zero but also large variability between the strata. Missingness was assumed to be at random so a missingness indicator was generated at random for number of employees enrolled. Thus all variable values were known in the pseudo-population but when the units with a positive missingness indicator were selected in the sample they were treated to be unknown. Also imputation classes were assigned based on size, strata, and a random assignment.

The sampling from the population was stratified probability proportionate to size with the size measure equal to the number of employees. The sampling was repeated 40,000 times in ten blocks of 4,000. For each repeat a twenty percent sample was drawn pps within each stratum. Then the full sample imputation was performed at random within each imputation class for observations that had the missingness indicator set to missing using donors from the observed values in the imputation class. Since the samples were fairly small the replicates were not grouped. For each observation in each stratum a replicate was created by deleting the observation. Three variance estimates, described in the next three sections, were calculated based on these replicates.

For each repeat of the sample, information was recorded on the three variance estimates and whether the adjustment involved an out of range replicate estimate.

The simulation is not designed to exactly replicate the MEPS-IC data but the goal is to have a stratified pps sample, replicate the problems observed in proportions from the MEPS-IC, and to be able to calculate a gold standard for the variance.

2.2 Naive Variance Estimate

The methods used here to adjust variance estimates for missing data all involve jackknife variance estimation. So the naïve estimate will be the standard stratified jackknife variance estimation method. For this method the missing values were imputed at the full sample level and treated as if they were observed data. No adjustments or reimputations were used. The li^{th} replicate is formed by deleting the i^{th} observation in the l^{th} stratum and the formula for variance is:

$$v(Y) = \sum_{l=1}^{L} \frac{(K_l - 1)}{K_l} \sum_{i=1}^{K_l} (Y_{li} - Y_{l0})^2 ,$$

where Y_{li} is the sample estimate from the i^{th} replicate in the l^{th} stratum and Y_{l0} is the sample estimate from the full sample in the l^{th} stratum.

2.3 Burns Method Jackknife Variance Estimate

In this method, replicates are created in the same way as they were with the naive variance estimates. The difference here is that before calculating the replicate estimates (Y_{li}), imputation and reweighting are run separately on each of the replicates. In this situation one observation is being dropped from each replicate so if a recipient is dropped no reimputation is necessary but if a donor is dropped a reimputation is performed for all of the missing values in the replicate.

Once imputation and reweighting have been run on each replicate, an estimate of the variance can be calculated using the same formula as in Section 2.2, where now (Y_{li}), is the sample estimate from the li^{th} replicate after imputation and reweighting have been rerun. This procedure, according to Rao and Shao (1992), is upwardly biased and will

thus on average serve as an upper bound of the true variance while the naive estimate will serve as a lower bound.

2.4 Jackknife Variance Using a Rao-Shao Adjustment for Imputed Values

The Rao and Shao (1992) adjustment was introduced as a way to capture the variability in an estimate due to imputation. The adjustment they proposed gives an asymptotically unbiased estimate of the variance by perturbing the imputed values in replicates in order to increase the amount of variance when compared to the naive method. Since the Rao and Shao adjustment produces a consistent estimate of variation, the estimates calculated in this manner should on average be less than the re-imputed jackknife estimates from Section 2.3 assuming the sample size is sufficiently large but in any given sample it is possible that the Rao-Shao adjusted jackknife estimate can exceed the Burns method jackknife estimate.

As with the naive estimates, imputation will only be run once on the full survey sample prior to estimation. For the purposes of this research, the full sample was only imputed once and the output from this imputation run was then used for both the naive estimates and the Rao and Shao adjusted estimates. This same imputation was also used as the full sample imputation for the Burns method. In this way, the only differences to be found in the three variance estimation methods will be between replicates. The full samples, and thus the estimates they produce, will be the same regardless of the variance estimation technique.

The Rao- Shao adjustments will be made to all imputed values, with specific adjustments made to different imputation cells. In imputation, class variables are used whereby a donor and recipient must be from the same class, i.e. they must share the same value of the class variable. These same classes will be used as cells for making the Rao and Shao adjustments such that all recipients that share the same values of these class variables will be adjusted the same way.

This adjustment, made to all imputed values, is described as follows. The cell means need to be calculated for both the full sample, and for each of the replicates using only the imputation donors for the variable being estimated. Let y_{lj}^{*} equal the imputed value for the j^{th} observation of variable y from the l^{th} stratum. Let $\bar{y}_{liv}$ equal the donor mean of the variable y from the l^{th} stratum, i^{th} replicate and v^{th} cell, and let $\bar{y}_{lv}$ equal the donor mean of the variable y from the l^{th} stratum in the v^{th} cell. Using this information, calculate:

$$z_{lij}^{*} = \begin{cases} y_{lj} & if\ y_{lj}\ not\ imputed \\ y_{lj}^{*} + (\bar{y}_{liv} - \bar{y}_{lv}) & if\ y_{lj}\ imputed \end{cases}$$

From here, the jackknife variance formula from Section 2.2 is applied. The only difference is in how the replicate estimates are calculated. The l^{th} stratum estimate, Y_{l0}, is calculated, as in Section 2.2, using unadjusted data. However, the replicate estimates, Y_{li}, are calculated using the z_{lij}^{*}'s derived above.

As was mentioned in section 1.2, it was observed that in some cases this adjustment would produce a negative estimate of proportion at the replicate level. As a simple example that this is mathematically possible consider the following. A simple random sample of size ten is taken and the values observed are eight zeroes, a single one, and a single missing value. Assume that the missing value is to be imputed at random from the nine observed values so there is a 1/9 probability of imputing a value of 1 and an 8/9

probability of imputing a value of 0. The mean of all of the donors is 1/9. In the replicate in which the lone one is to be deleted, the mean of the remaining donors is 0. Thus in this replicate the adjustment factor to be added to the imputed value is (0 – 1/9). Since the imputed value has an 8/9 probability of being 0 there is an 8/9 probability that the adjusted replicate estimate is -1/9 but on average across all imputations the adjustment is 0.

3. Results

The simulation results were somewhat consistent with the results in Rao and Shao (1992). The results presented here are related to the standard error of the proportion of employees taking insurance in the establishments. The results indicate that as the number of strata increased the Rao-Shao adjusted estimate of variance more accurately estimated the true variance. The bias in the Burns estimate of variance did not increase through the sample sizes considered here. That may simply mean that sufficiently large sample sizes were not used in the simulation to detect this property of the Burns method.

Negative estimates of adjusted proportion of employees taking insurance were detected in about ten percent of the samples. However, in these cases the Rao-Shao adjusted estimate exceeded the Burns method estimate about 10%-14% which is almost exactly the percent of time it exceeded the Burns method estimate in the total number of samples. This would seem to indicate that the event of a negative estimate of adjusted proportion is a red herring and should not be truncated. In any given sample it may make the estimate look more reasonable but on average, as a procedure, it can introduce a bias into the estimation.

Table 1 gives the percent of samples for which the Burns method exceeded the Rao-Shao method, the percent of samples for which the Rao-Shao method exceeded the Burns, and the number of times the two methods were equal. Note that the two methods could be equal in samples with a large number of zero counts for employees taking insurance since the adjustment would be zero or small and the reimputation would be reimputed with a zero value.

Table 1. Percent for maximum variances

Number of Strata	10	20	30	40
RS<Burns	19%	50%	71%	61%
RS>Burns	12%	10%	14%	12%
RS=Burns	69%	40%	15%	23%

The theoretical results for the Burns method indicate that for a sufficiently large sample it is upwardly biased and should exceed the Rao-Shao adjusted method. But in practice this is not a mathematical bound and for samples in which the Rao-Shao exceeds the Burns method it can be concluded that the sample size is not sufficiently large for the bias to enforce the asymptotic relationship.

Table 2 gives the percent change in the standard errors under the three methods relative to the gold standard. For the naïve method the percent of underestimation increases as the number of strata increase. This may be because as the number of strata increase there are more missing values but the percent of missingness as a proportion of the sample did remain constant across the columns. The Burns method did appear to decrease in bias but

theory tells us that as the sample size continues to increase it must start to increase in bias again. It may be that for small sample sizes it can produce reasonable estimates but at some point it must become upwardly biased. The Rao-Shao adjustment method did become more accurate as the sample size increased and the theory tells us that it is a consistent estimator. However, from the previous work and from the simulation we can see that it can have large variability.

Table 2. Standard Errors: Percent above or below the gold standard

Number of Strata	10	20	30	40
Naive	-8%	-38%	-54%	-58%
Burns	+121%	+72%	+30%	+28%
Rao-Shao	+109%	+26%	-0.9%	-7%

4. Conclusions

The goal of the previous project was to assess the level of bias in the MEPS-IC variance estimates due to missing data. However, that research raised questions about the use of adjustments for proportional data. This project attempted to address those issues by simulation where a gold standard could be calculated. The results of this research indicate that the Rao-Shao adjustment technique is unbiased but can be highly variable. This is not a reassuring result for use of the adjustment technique.

5. Future Research

Before a decision can be made to report the results to the end data user clarification of the issues with proportions must be resolved. In the literature there are very few methods for addressing increase in variance due to imputation. Multiple imputation is known to have multiple biases so it was not initially considered. Perhaps it could be investigated as well. There are techniques for adjustment in the presence of nearest neighbor imputation due to Chen and Shao (2001) that could be investigated. This would also be closer to the production method of imputation used in the MEPS-IC. It may be useful to consider more variables but that would not address the fact that an important variable is already known to be problematic. Further simulations can provide information but will not fix the problem with the proportions. The only way to solve the issue with proportions is to refine the adjustment formula or find a new formula that can provide reasonable estimates.

Currently standard errors for the MEPS-IC estimates are reported without accounting for the potential impact of missing data. After analyzing additional estimates and assessing the bias introduced by these missing data procedures, an informed decision can be made as to how best to inform data users of the impact of missing data on the reported standard errors.

Acknowledgements

The authors would like to thank Prof. JNK Rao and Dr. Phil Kott for helpful suggestions. The authors would like to thank the many reviewers for their careful review.

References

Burns, R.M. “Multiple and Replicate Item Imputation in a Complex Sample Survey.” *Proceedings of the Sixth Annual Research Conference*, U.S. Bureau of the Census. (1990): 655-65.

Chen, Jiahua, and Jun Shao. "Jackknife Variance Estimation for Nearest-Neighbor Imputation." *Journal of the American Statistical Association*. 96.453 (2001): 260-69.

Rao, J.N.K., and J. Shao. "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation." *Biometrika*. 79.4 (1992): 811-22.

Thompson, Matthew S. and Kearney, Anne T., “Assessing the Impact of Missing Data on Variance Estimates for the Medical Expenditures Panel Survey – Insurance Component” *Proceedings of the Business and Economics Section*, American Statistical Association (2010).