Using imputation to reduce the cost of survey collection in the Current Population Survey

John Dixon

Bureau of Labor Statistics

Abstract

Survey cost has been of concern due to the increasing cost of interviewing, as increased effort has been needed to maintain sample size due to an increase in households that are reluctant to respond. The reluctance is seen in the increase in nonresponse in most household surveys. This study investigates whether imputing survey responses, rather than collecting the data, for select respondents in a longitudinal survey can reduce costs while maintaining estimate quality. The Current Population Survey (CPS) is designed to measure labor force characteristics of the United States. Since households are interviewed eight times, subgroups that are likely to have stable labor force status are of special interest for imputing. Some examples include older retired people (where the entire household would be not-in-labor-force), parents involved in childcare for young children (where one parent is employed, and the other is committed to childcare), and students during the school year.

In this study, we use models to predict households that could be good candidates for imputation using the person and household characteristics that predict no change in labor force status. To do this, the labor force status for the subsequent months is imputed for households after the first interview based on person characteristics, household characteristics, and labor force status collected in the prior interview. The effect of imputing is simulated by eliminating the responses of those predicted to have stable labor force status, and imputing their responses. The difference in estimates between the original and imputed data shows the effectiveness of the model.

Key Words: panel survey, imputation, survey cost, classification trees.

1. Introduction

The cost of surveys to collect information from households has steadily risen over time. The motivation of this study is to explore reducing costs by eliminating household interviews where it could be possible to impute labor force status, the key outcome variable for the CPS, which is the primary source of information on the labor force characteristics of the U.S. population.[1] Some characteristics of the survey design and administration include:

- The CPS consists of 8 separate interviews spread out over a 16 month period using a complex sample rotation design, which includes four consecutive months of interviews, followed by an eight month break, then four more consecutive months of interviews.
- The data collection period for the monthly CPS is 10 days.
- Months 1 and 5 (with an 8-month break in between) are designed to be in-person interviews.
- Months 2, 3, 4, and 6 through 8 are designed to be telephone interviews, either conducted by field interviewers in a decentralized manner (about 66 percent) or sent to a telephone call center (about 10 percent). The remaining 24 percent are conducted in person.

A number of studies have examined imputation in the context of survey costs (Haslett et al., 2010, Lee, 2015). Others have used classification trees to build models for imputation (Bechtel et al, 2015).

2. Study Design

The focus of this study is whether imputation can provide good data quality with fewer interviews. Utilizing data from 2010 through 2013 as a training dataset, classification trees were used to identify subgroups that do not change their labor force status. Separate models were used for each CPS labor force category, which include not in the labor force (NILF), employed, and unemployed. Those starting as "not in labor force" might be expected to be older (retired), or younger involved in caring for pre-school children. Those starting as "Employed" might be expected to be younger. Once the classification models were developed, a testing data set from 2014 was used to test the effect of imputation by randomly eliminating the responses for some interviews after the first interview based on the groups identified in the first step.

The predictors include age, race, ethnicity, gender, and labor force status reported in the previous wave. These predictors are used to report demographics in the CPS labor force tables. The labor force status for the subsequent wave is imputed for households after the first interview based on the person characteristics, household characteristics, and labor force status collected in a prior interview. The effect of imputing is simulated by eliminating the responses of those predicted to have stable labor force status, and then imputing their responses. The difference in estimates between the original data and the data with imputations shows the effectiveness of the model. The change in the standard error of the labor force estimates from the multiple imputation is used to indicate the quality of the imputation. The potential savings is estimated based on the number of

[1] Details about the CPS can be found in Technical Paper 66 (http://www.census.gov/prod/2006pubs/tp-66.pdf).

attempted contacts and the interview time of the survey for those households which are imputed.

In measuring the effect of the imputation, the percent eliminated was varied in the simulations, and the effect on estimates of labor force status was examined. While the percent eliminated from the "imputation cells" ranged from 10% to 70%, the graphs show the percent eliminated from the total sample, indicated as "fraction missing". This gives a better indicator of savings which might be possible. For example, for "not in the labor force" a 10% imputation only reduces the sample by 2%, while a 70% elimination rate only reduces the sample by 41%.

3. Findings

Building the models to find the subgroups provided some surprises. Stereotypes of the labor force proved unreliable. Older, presumably retired adults moved in and out of the labor force more than expected, although at a lower rate than younger adults.

Figure 1 shows the beginning of the classification tree for predicting no change in labor force status for those starting as "not in labor force" (NILF). The right branches are for no change, the left for change. Other variables are further down the tree. The width of the branches indicates the relative sample size. Age had the largest effect, with older (>64.5 node E) and higher educated (some college or higher node C) less likely to change out of NILF.

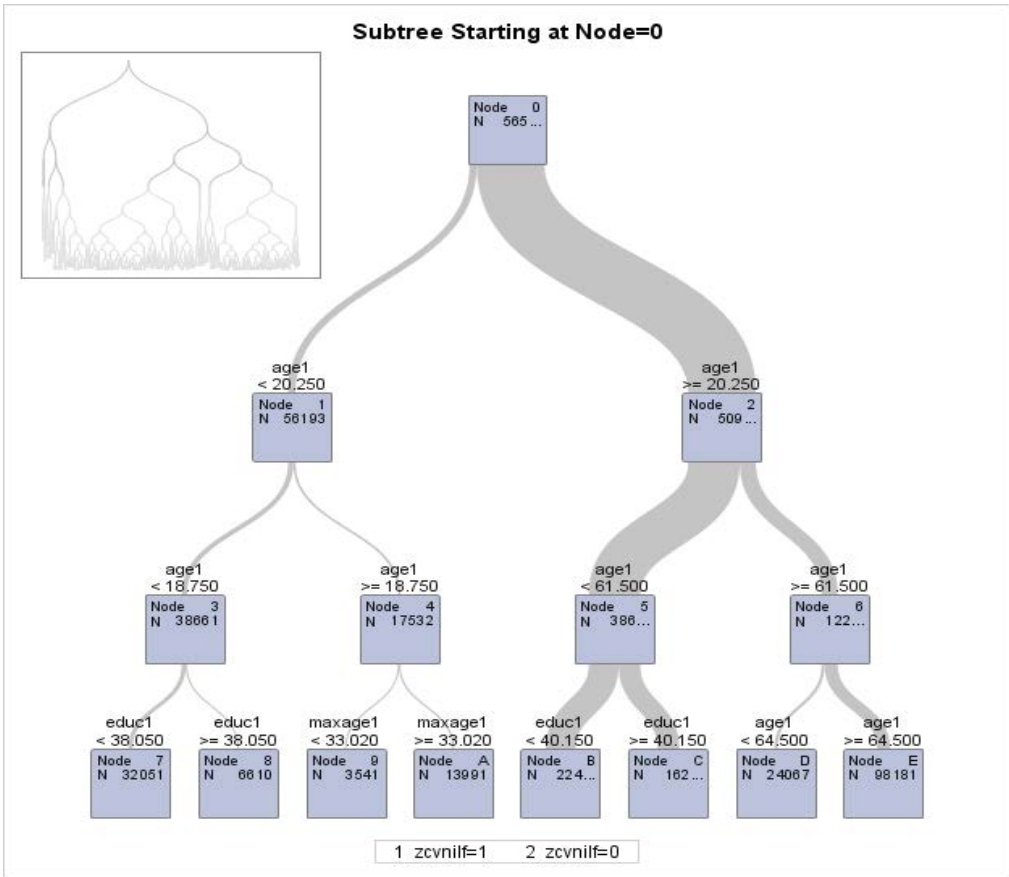
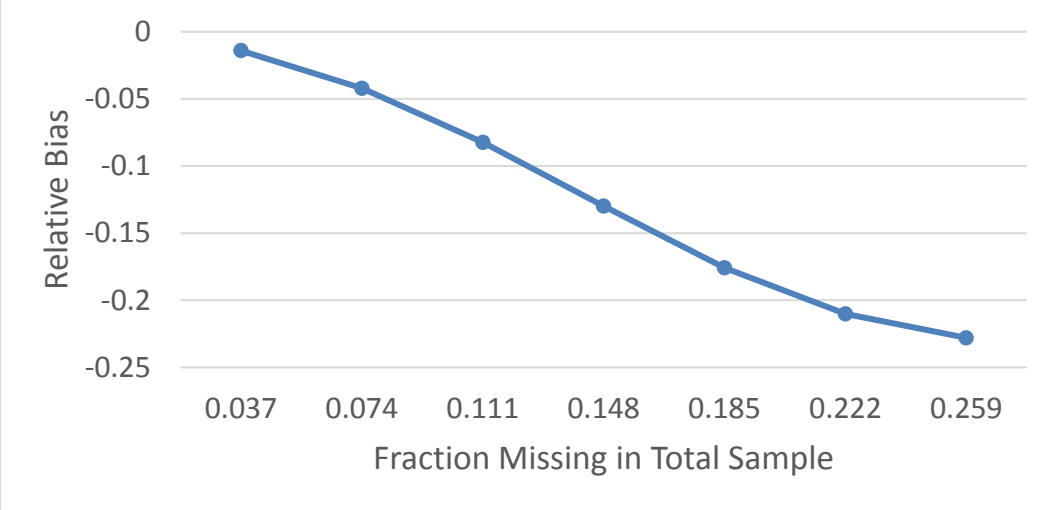


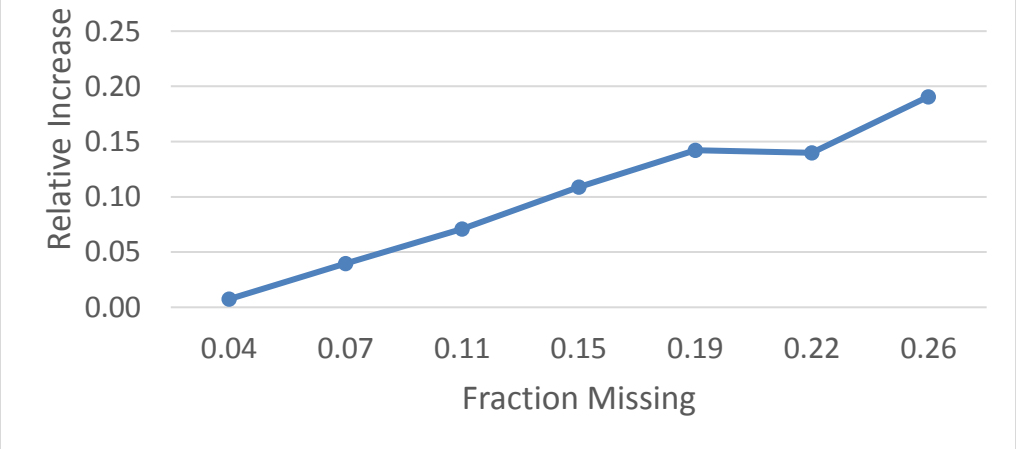
Figure 1: Classification Tree for "Not in Labor Force" Change.

Graph 1a shows the result of the simulated missing data. The horizontal axis shows the proportion eliminated (represented by fraction of missing data); from the 10% to 70% imputation cells with the least change from "Not in labor force" (NILF). Since the NILF category comprised about 33% of the sample, if the low change cells contain too many of the other labor force categories, then the estimates will be biased downward. The relative bias is the difference in the NILF estimate from the imputation and the actual estimate. At the 0.037 fraction of missing would represent a savings of only 0.37 percent of interviews for NILF. The total savings would depend on the other labor force imputations.



Graph 1a: Bias in "Not in Labor Force" estimates for different amounts of imputation.

Graph 1b shows the increase in standard error estimates from the multiple imputation. This is useful as an indicator for how successful the imputations were, with larger increases associated with poorer imputations. The increase in variance was gradual, similar to the increase in bias seen in Graph 1a.



Graph 1b: Standard Error increase in "Not in Labor Force" estimates for different amounts of imputation.

Figure 2 shows the classification tree for the change in labor force status for "employed". Again, age and education were the best predictors of change from "employed".

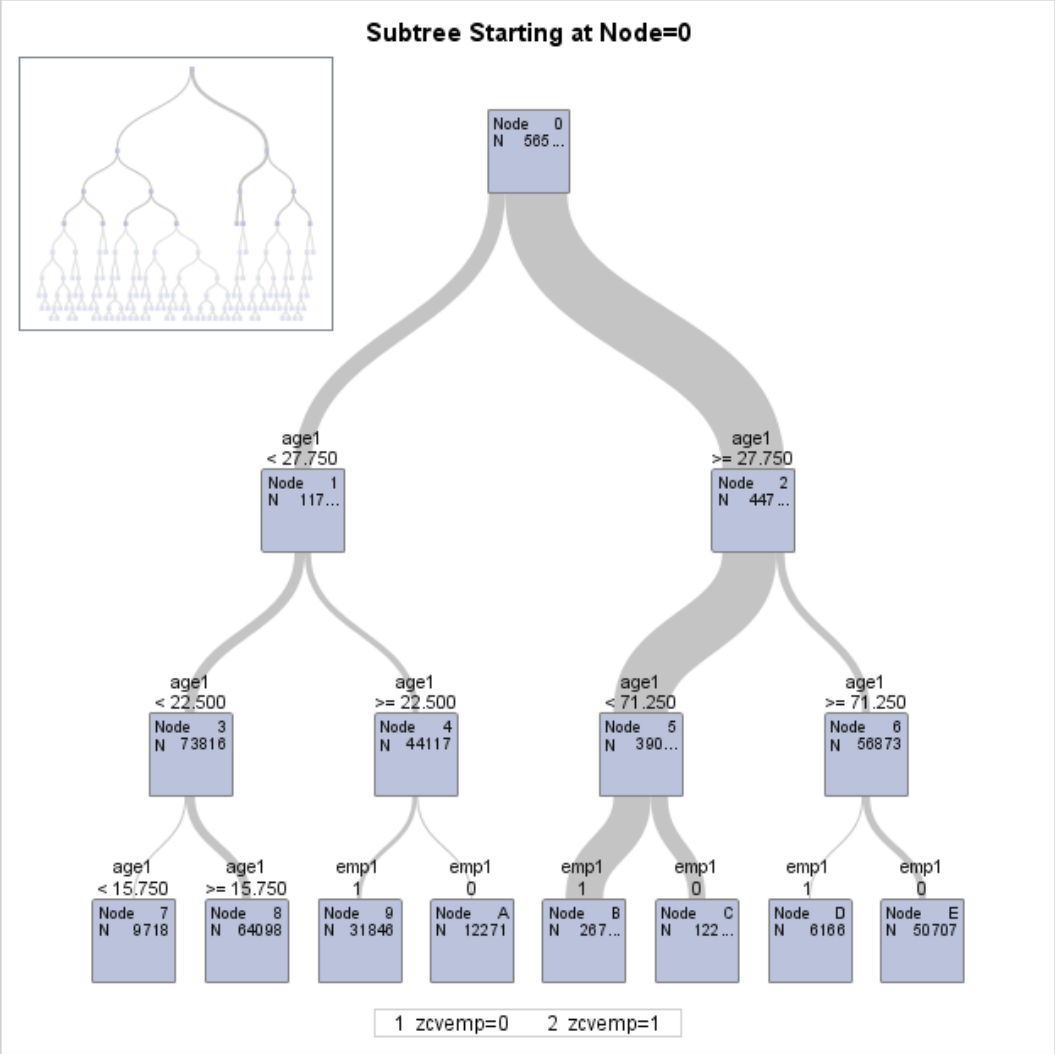
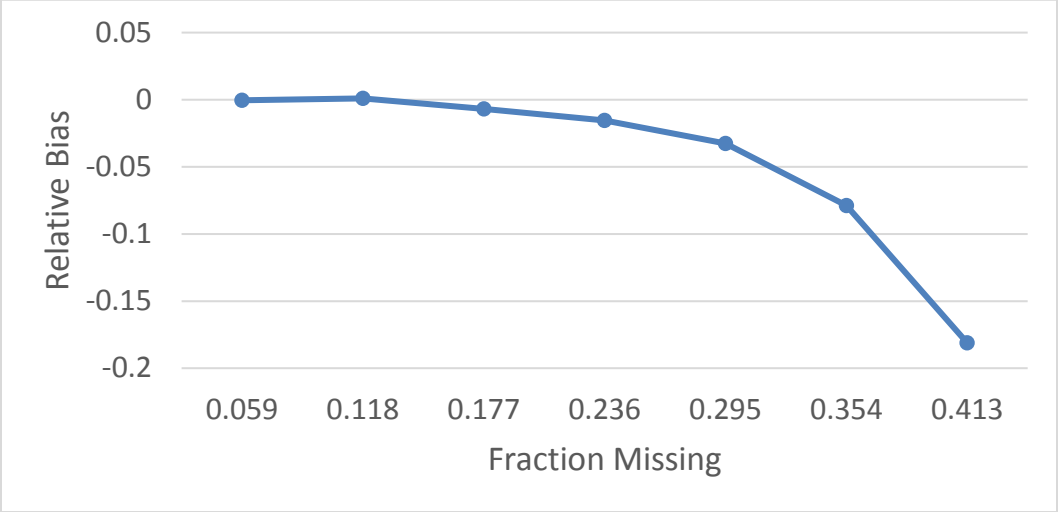


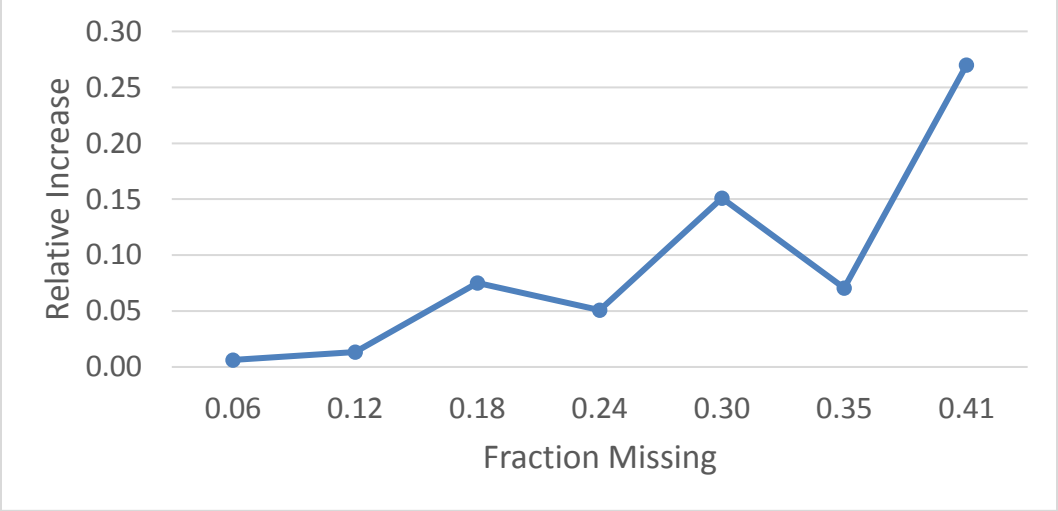
Figure 2: Classification tree for "Employed" change.

Graph 2a shows the bias for different amounts of imputation. The curve is much flatter than the other labor force categories, so more imputation could be done with this group. At least 10% to 20% of interviews could be saved with minimal impact on estimates.



Graph 2a: Bias in "Employment" for different amounts of imputation.

Graph 2b shows the relative increase in standard errors due to imputation. It was low for the first two levels, but rises noticeably after that. This was also reflected in the bias in Graph 2a.



Graph 2b: Standard Error increase in "Employment" for different amounts of imputation.

Figure 3 shows the classification tree for change in labor force status from unemployed. Age and education are the main determinants of change, where younger less educated are more likely to stay unemployed.

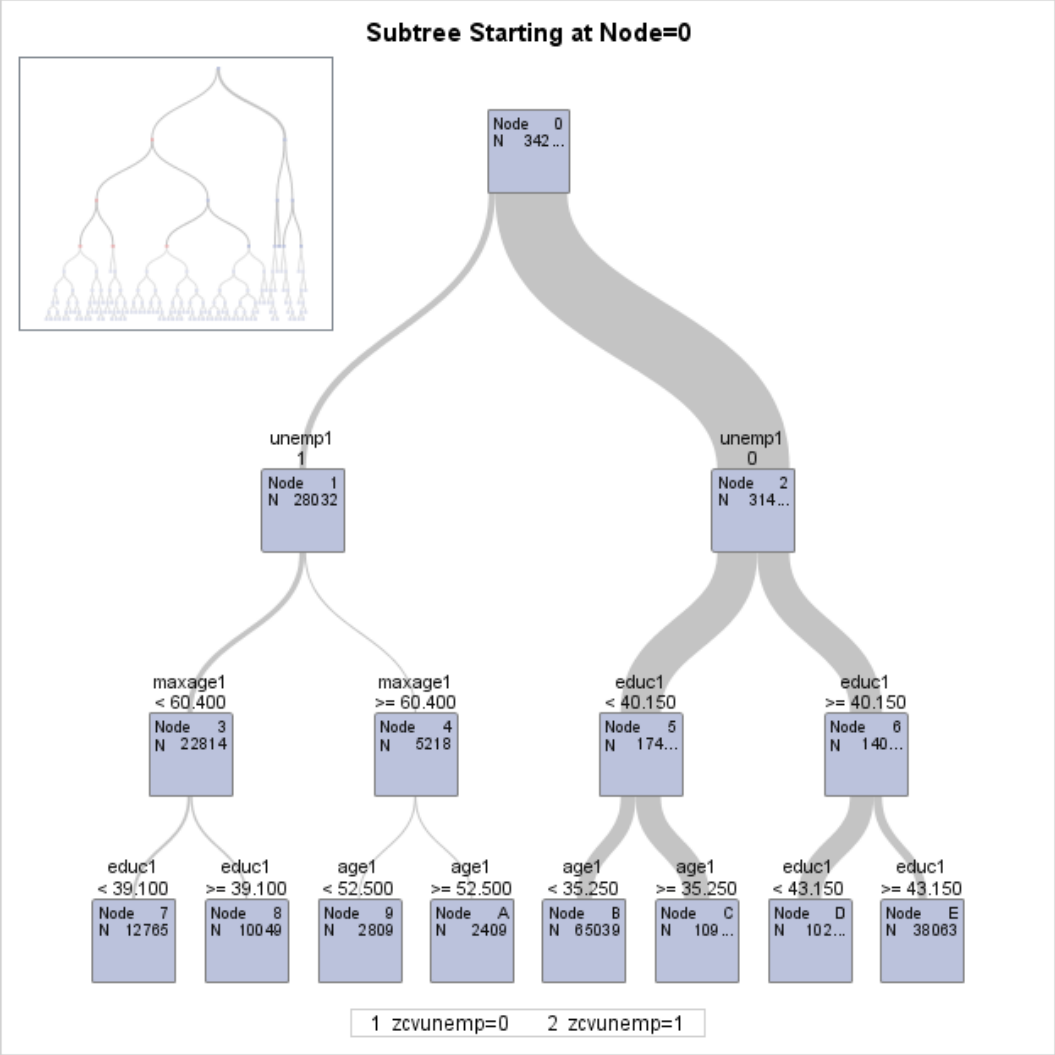
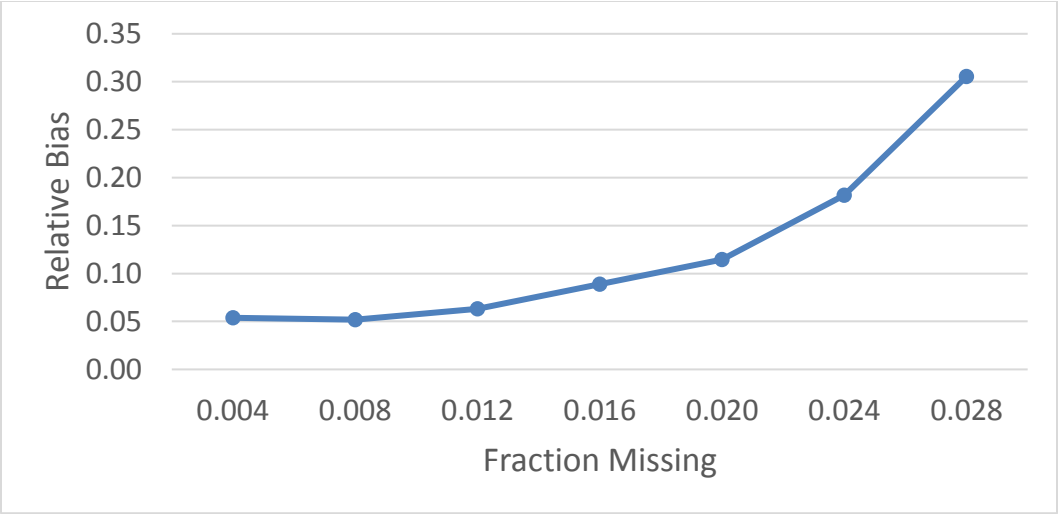


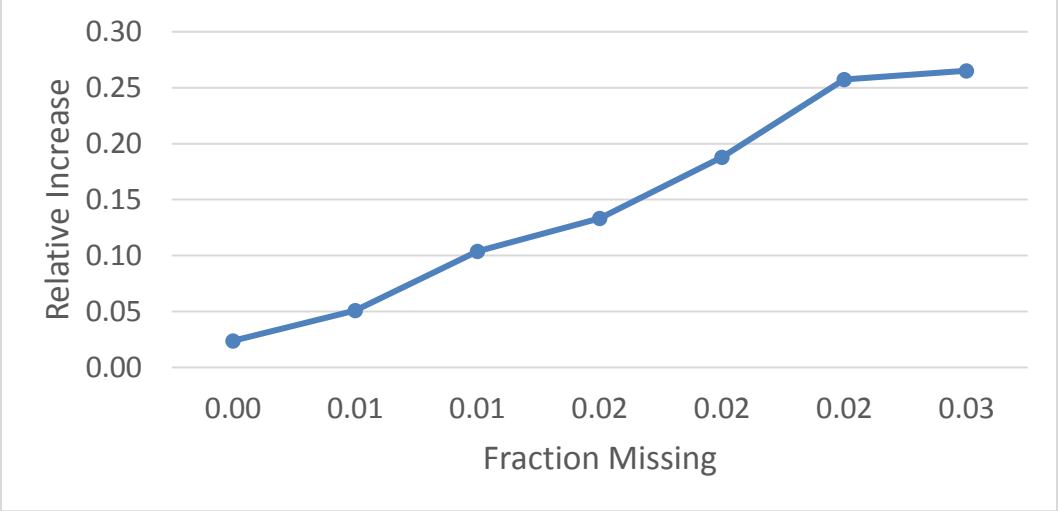
Figure 3: Classification Tree for "Unemployed" Change.

Graph 3a shows the relative bias with increasing percent imputed. Because the proportion unemployed is small in the sample, the larger proportions of other labor force categories would be expecting in the imputation cells. This would make for more bias in the imputations, particularly as the proportion imputed increases.



Graph 3a: Bias in "Unemployment" for different amounts of imputation.

Graph 3b shows a gradual increase in the relative increase in standard error. It was comparable to the other labor force categories.



Graph 3b: Standard Error increase in "Unemployment" for different amounts of imputation.

4. Summary and Suggestions

Based on the study results, there may be enough predictability of labor force status to impute a small fraction of the sample. Those who are employed could be imputed from 10 to 20 percent without appreciable bias. That produced and estimated 0.005 to 0.006 difference from the estimate of 0.585. This would be on the order of rounding. Those not in the labor force are more sensitive to the fraction missing, but 10 percent showed little bias. The bias was -0.006 difference from the estimate of 0.37, again within the rounding bias. The unemployed constitute such a small part of the sample it wouldn't save much to impute for them.

5. Limitations

The time period studied had relatively stable employment. During times of rapid change imputation may not work well. This study was based on imputing for the second interview using information from the first interview. The effect on the 3rd, 4th, and 6th through 8th interviews needs to be studied. Although the 2-4 and 6-8 interviews cover a potential 75 percent of the sample, much of the cost of interviewing is in the 1st and 5th interviews, since they are more often in person.

6. Future research

Refinements of the classification models used to develop the imputation cells may improve the efficiency of the imputation model. Additional variables to consider would include seasonal effects, exposure to the annual supplement (a long series of financial questions), and region of the country. Since nonresponse constitutes nearly 15% of the CPS sample, imputation models for nonresponse might reduce cost through adaptive design, although weighting procedures already serve some of the function with weighting cells similar to the imputation cells studied here. Multivariate imputations should be used to include other variables collected in the CPS.

References

Bechtel, L., Morris, D.S., and Thompson, K.J. (2015). "Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study." In *FCSM Proceedings*. Washington, DC: Federal Committee on Statistical Methodology.

Haslett, Stephen, Jones, Geoffrey, Noble, Alasdair, and Ballas, Dimitris (2010), "More for Less? Comparing small area estimation, spatial microsimulation, and mass imputation", Paper presented at the Joint Statistical Meetings, https://ww2.amstat.org/sections/srms/Proceedings/y2010/Files/306741_57091.pdf

Lee, Shin-Jung (2015), "Optimizing Survey Cost-Error Tradeoffs: A Multiple Imputation Strategy Using the Census Planning Database", Paper presented at the Joint Statistical Meetings.

U.S. CENSUS BUREAU, Current Population Survey Design and Methodology Technical Paper 66, October 2006.