Summary of Problem: Interview to Interview Imputation Methods Team[1]

Prepared for NISS Workshop on Analyzing Complex Survey Data with Missing Item Values

Presentation Date: 10/17/2014

By Geoffrey Paulin, Bureau of Labor Statistics, Consumer Expenditure Survey Program

Introduction

The Interview component of the Consumer Expenditure Survey (CE) currently consists of five interviews conducted over consecutive quarters. In addition to detailed expenditure and demographic information, the Interview Survey also collects information on income (second and fifth interviews), as well as assets and liabilities (fifth interview). Therefore, completing an interview can take time: In 2011, the average quarterly interview was 1 hour; 10 percent of interviews exceeded 100 minutes.

Reducing respondent burden and improving data quality are perennial concerns for the CE program. As part of this mission, the bounding interview (currently first interview) will be dropped from the survey in 2015.[2] This will necessitate collecting some information from this interview into the current second interview. Unless other questions are removed from current second interview, the estimated time required to complete the new first (currently second) interview will increase. The Interview to Interview Imputation Methods Team was chartered to see whether some items collected in the current second interview can be imputed using expenditures reported in the current third, fourth and fifth interviews in an effort to address both goals.

About the Data

Each quarter, the CE interviews approximately 7,000 *consumer units*, which are similar to families or households.[3] Designed to collect information on big-ticket or recurring expenditures, the Interview Survey also asks global questions on certain items, such as food at home.[4] Other items, such as utilities, are collected at a more detailed level.

Once collected and processed, data are published in many formats. This includes standard tables including means and standard errors for many demographic groups. However, another important format

---

[1] This research was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed here do not necessarily reflect the views of the BLS.

[2] The purpose of a bounding survey is to ensure information requested is reported accurately: For example, if the respondent reports the purchase of a major appliance in the first interview, and reports purchase of an identical appliance in the second, the interviewer can confirm with the respondent that the item reported in the second interview is indeed unique, and not a repeat report of the item actually purchased prior to the first interview.

[3] In essence, consumer units include members of a household who are either related by blood, marriage, or other legal arrangement, or who pool resources to fund purchases in major expenditure categories. Because a household is a physical dwelling, a household may comprise more than one consumer unit, as in roommates sharing an apartment.

[4] Global questions are general, such as, "What was your typical grocery bill," as opposed to "How much did you spend on lettuce, bread, etc."

is public use microdata files. Using these files, researchers can conduct their own explorations of relationships between expenditures and consumer unit characteristics. Any imputation method adopted would have to produce data satisfactory to all data users (tabular and microdata).

Expenditures Evaluated

The team evaluated food at home and electricity. Models for each of these items produced low-quality results, meaning that expenditures for food at home and electricity cannot be imputed across interviews.

The team started to evaluate telephone expenditures. The problems encountered, and those anticipated from the food and electricity analyses, led to a discontinuation of the investigation.

Methodology:

In theory, it would be possible to "impute backward." That is, on a consumer-unit-by-consumer-unit basis, reported expenditures from third, fourth, and fifth interviews are used to impute expenditures for the second interview. This would be especially useful for expenditures that presumably have low variance over time, such as cable television. However, in practice, such a method is infeasible. Production of all second interview data would have to wait until all completion of the fifth interview before it could be completed. In addition, due to attrition, some consumer units do not complete any or all interviews after the second. Therefore, this method was not investigated.

The team briefly investigated using the hot-deck approach for imputing values. One reason for doing so is that this technique is currently used in production to fill in cases where the consumer units reports that a purchase occurred, but not the amount of the purchase. However, for reasons well-documented in the literature, the team ruled out using this procedure.[5] Therefore, regression-based imputation was selected for further investigation. This technique also had precedent in production-based systems, such as income imputation.

Regarding the regressions:

1) Models were run for each expenditure separately by family type, and other categorical breakdowns where appropriate, for example, by housing tenure and geographic area for electricity.

2) Models were evaluated through R-square, mean absolute deviations, and absolute percentage by which predicted values deviate from reported values.

3) Total outlays less expenditure of interest were used as a proxy for income for processing and theoretical reasons detailed in the main text.

[5] Among the concerns is the fact that while hot decking preserves means, it artificially minimizes variance of estimated values, because, even when samples are fairly large, the number of donor cases used for matching can be relatively small. Furthermore, because it is usually necessary to limit the donor class to a small number of variables (e.g., matching by region, age, and income only), correlations between expenditures and demographic characteristics not included (e.g., education, family size, and housing tenure) are not preserved. This is particularly important because under the proposed system, the imputations would not be performed only for cases where there are many reports and some missing values, but would replace all the data that would have been collected in the course of an interview.

REGRESSION ANALYSIS

Food at Home Model Details:

Cases were analyzed where reports for purchases of food at home were positive for participants in 3rd, 4th, and 5th interviews. The sample size was found to be large regardless of category (e.g., family type) tested. This was consistent with *a priori* expectations: Publications show that the percent reporting food at home expenditures through a global estimate in the 2011 Interview Survey is almost 99 percent.[6]

Models were run based on family type: single males and females are modeled separately from each other, as are husband/wife only consumer units, and other consumer units with differently-aged children. Each model included appropriate demographic variables, such as age, outlays, and family size (except for singles and husband/wife only models).

While some variables included in the models were statistically significant, the model results as a whole did not produce large R-squared statistics, and yielded low predictive powers as measured by absolute deviations. This result held even when the regressions were performed using Box-Cox transformations.[7] This may be because the models are underspecified, or due to factors beyond the control of the model (for example, if the dependent variable is by nature unpredictably variable). The team did not discover any definitive explanation.

Electricity Model Details:

Electric bills are generally charged for use occurring at a particular address, rather than to individual consumer units therein. Therefore, for simplicity, the models of electricity expenditures were run only for households comprising one consumer unit. For cases where two or more consumer units share a household, some method of allocation of expenditures imputed at the household level would need to be developed. An obvious suggestion is to impute the expenditure at the household level and divide the expenditure by the number of consumer units therein.

More complicated is the geographic problem. General weather patterns will clearly affect use of electricity demand for heating and cooling. However, variables that are presumed to be predictive, such as "degree-days" in the area for which the consumer unit resides, are not available even for testing purposes.[8] Therefore, while the model may be underspecified, there is no way to know for sure.

Related to this, even if two geographic areas are near each other, and therefore experience similar weather patterns, there is rarely competition within a geographic location among electricity suppliers. Therefore, households that are similar in size, location, and usage, but serviced by different providers, may receive

---

[6] Source: 2011 Interview PPUB, Age of Reference Person, All Consumer Units column.

[7] The formula involves raising the dependent variable (Y) to the power "lambda" (L), substracting one, and dividing the result by "lamba": That is, transformed $Y = (Y^L - 1)/L$.

[8] While "degree-day" data may be available from outside sources, they are not collected in the Interview Survey. In production, relying on data collected from outside (i.e., non-CE program) sources is problematic, as changes in definition, data availability, or other factors can occur without notice. Therefore, only data collected in the Interview Survey were considered for this project.

substantially different electric bills. Therefore, it is not clear at what level of geography each model should be run to ensure highest quality results. For example, in theory, separate models by each PSU makes sense. However, practically, this would result in a large number of models, each of which would have a small (if any) amount of source data, thereby rendering the results, if any can be obtained, meaningless.

Another consideration is data quality. The section may require substantial time to answer because respondents are using records more frequently than in other sections. To the extent that they are willing to do so may indicate high quality data with relatively little burden (if burden is measured by discomfort or other reluctance to use records, rather than time of interview alone). In this way, imputation of electricity expenditures might replace high quality data with imputed data, and thus reduce overall quality of the results. Indeed, the data show that 38 percent of all single-consumer-unit households interviewed in 2011 used some type of record (for example, checkbooks, bank statements, or bills) to answer questions in section 4.[9] At the same time, it may be that the high quality of reported data yields high quality imputations, and therefore it is less problematic to replace the data. In addition, even when only binary variables identifying types of records used are included in the model (done to minimize multicollinearity), they are generally not statistically significant, at least not for the models tested. If there are no errors in coding, this could indicate that respondents recall their expenditures accurately for electric bills. This may also indicate some other as yet undiscovered factor is at work.

Telephone Model Details:

The team started, but soon discontinued, investigating imputation of expenditures for telephones, including residential phones (henceforth, landlines), cellular phones (henceforth, cell phones), phone cards, and voice over IP service. One reason was the finding, consistent with food and electricity models, of low R-square values for regressions. However, there were other considerations that led to the discontinuation of the investigation.

First, cell phone and landline expenditures are presumed to be highly correlated for consumer units that have both. In such cases, it is not clear which expenditure should be imputed first (cell or landline), as the results may be used to impute the other expenditure. It is also possible that some type of more complicated, simultaneous equations modeling would be in order. Both selecting the type of model and ascertaining the implications for production would be difficult, and perhaps fruitless based on the preliminary results obtained.

Second, even if the order of imputation were clear (e.g., use landline expenditures to impute cell phone expenditures), the second item imputed would involve using first-item expenditures that are themselves imputed, meaning that the second-item imputations would be even less accurate than those imputed using reported values for the first item.

[9] The variable RECSEC02 indicates whether or not the respondent used some kind of record in answering Section 4. The variables TYPEREC1 (bills) through TYPEREC9 (none) identify specific types of records, if any, used at some point during the interview, but not necessarily for Section 4. The team did not ascertain whether it is possible to identify what types of records were used for each section; this remained under investigation at the time the work was discontinued.

Third, similar to the electricity imputations, calling plans for both landline and cell phones may differ substantially by area of the country in which the consumer unit is located, and there are so many different types of plans available that it is impossible to predict what levels of expenditures would be, even if the name and type of company (e.g., cable television, telecommunication, or internet-based firm) offering the plan were collected. Related to this, bundling plans, where phone service is combined with other services, such as internet access, increase the difficulty of imputing phone services alone.

Summary of Findings:

The models tested (food at home, electricity, and telephone) exhibited low R-square values, and large mean absolute deviations. Furthermore, expenditures for electricity and telephone were too complicated to model in a production environment, due to the numbers of models required for accurate imputations, and, for telephone, the potential for correlation among types of expenditures (e.g., landline and cell phone). Therefore, the investigation was discontinued.

Related Projects and Future Plans:

The CE program is currently investigating the feasibility of imputing asset and liability data when it is missing due to nonresponse. Suggestions for improvement of the Interview-to-Interview imputation results may be helpful in this project.