Investigating the Imputation of Assets and Liabilities in the CE Interview Survey

Geoffrey Paulin, Ph.D
Senior Economist
Consumer Expenditure Surveys (CE) Program
Technical Advisory Committee (TAC) Meeting
May 19, 2023
Virtual Event

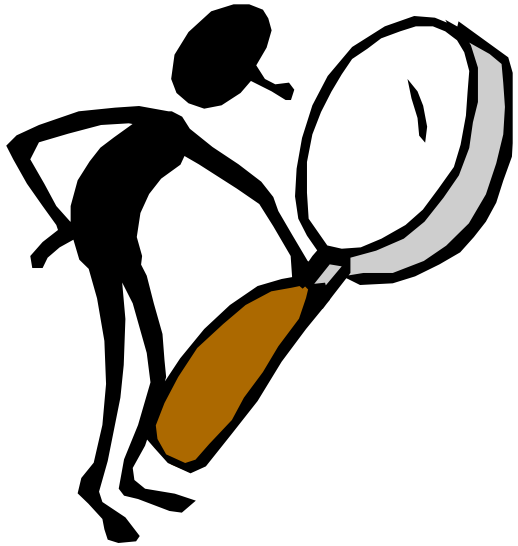
The Consumer Expenditure Survey (CE) collects information on:

- Expenditures
- Income
- Taxes (income and other)
- Assets and Liabilities

Nonresponse is a problem for each. However, corrections are in place for most of these items:

- Expenditures: Since the 1980s
- Income: Since 2004
- Taxes (income only): Since 2013

Assets and Liabilities are currently under investigation.



Assets and Liabilities

Project Overview

- "The purpose of this team is to initiate and conduct a research project designed to impute missing Interview asset and liability data, leveraging models from income imputation and other relevant procedures."
- "The goal is to implement this into production with 2017 Quarter 2 data."

Source: Charter for the Asset and Liability Imputation Team, 9-9-2014

This presentation describes three aspects of the project:

- What asset and liability data are collected?
- What processes have been considered for imputation of missing values?
- What are the next steps in the investigation?

But first….

Income Collection in the CE:

- Interview Survey
  - 1st and 4th interview (2015 onward; 2nd and 5th previously)
- Diary Survey
  - One time only (1st or 2nd week, at interviewer's discretion)

In both surveys, components are:

- Collected for:
  - The consumer unit as a whole (e.g., INTRDVX),

  or

  - Each member 14 or older (e.g., SEMPFRMX)
- Summed to consumer unit total (FINCBTAX, Interview; FINCBEFX, Diary)
- Subject to nonresponse. This leads to biased statistics (means, standard errors, etc.).

First, respondents are asked for each component: "Did you or any member of your household receive [type of income]?"

- If yes, then asked: "What was the amount?" To which respondent reports:
  - Actual value; If unknown or refused:
    - Bracket value; if unknown or refused:
      - No information ("invalid blank")
- If no:
  - Next source is collected.
  - But if all "no," the respondent is an "All Valid Blank" (AVB) reporter.

How are missing data handled?

- Historical Data:
  - 1972-73, 1980-2003:
    - "Complete Reporter" definition is in effect:
      - Complicated: "Reference person"-based, but not always;
      - Does not mean "all valid" reporters of income.
  - 2001: Brackets introduced to Interview Survey.
  - 2004: Brackets introduced to Diary Survey.
- Current Data:
  - 2004-present: Missing incomes are imputed.

Income Imputation Highlights

- Enables CE to fill in blanks due to nonresponse;
- Particular methodology is called “multiple imputation,” because there is more than one imputed value for each income source not reported.

Why "multiple" imputation?

- Technical reasons, related to variance.
  - From the User's Guide:
    - Multiple imputation "yields variance estimates that take into account the uncertainty built into the data from the fact that some observations are imputed, rather than reported." (P. 1, section I.A.)
  - In other words: Multiply imputed data are designed to have larger variances than "singly" imputed data because, by definition, imputed data are "best guesses," not actual values.

How are data multiply imputed?

- If respondent reports actual value:
  - Five “imputations” appear in the dataset, replicating the amount reported.
    - Example: Respondent reports value of INTRDVX to be $100. INTRDVXm = $100 (where m is number of imputations, and 1<=m<=5 in CE)

How are data multiply imputed? (Continued)

- Bracket Reports:
  - Through an algorithm, a random value within the bracket range is drawn, and serves as the first imputation.
  - Process is repeated four times.
  - Example:
    - Respondent reports $0<INTRDVX<=$999.
    - Values as small as $1 and as large as $999 are plausible (e.g., $10; $494; $384; $875; and $132 is a plausible string of imputed values for INTRDVX1-5)

How are data multiply imputed? (We're nearly done…)

- Regression-based, when respondent reports no information beyond receipt
  1. Income reported by similar consumer units is regressed on independent variables.
  2. Coefficients are "shocked" (i.e., random noise is added to each).
  3. Predicted values are produced using the "shocked" model coefficients.
  4. Predicted values from first "shocked" model are each "shocked"; The resulting values are used to fill in invalid blanks where they occur.
  5. This process is repeated four times, starting at step 2.

How are data multiply imputed? (Exciting Conclusion!)

- In case of AVB:
  1. Impute receipt (or lack thereof) for each source of income.
  2. If receipt is imputed, treat observation as a standard "model-based" case.

We now return to the topic at hand

Asset/Liability Data

Assets:

- Retirement accounts
- Stocks, bonds, mutual funds
- Checking, savings, money market, CDs
- Whole life insurance
- Other, including annuities, trusts, royalties

Liabilities:

- Credit cards
- Student loans
- Other loans, including medical and personal

Collection

- Questions are asked in the final survey (4th interview)
- Most are asked in two parts: Did you have ____? If yes, how much?
- For some items, only a total value is collected. In these cases, it is not clear whether $0 means:
  - No, I did not have such an account or
  - Yes, I had an account, but it is empty.

Collection, continued:

- For each asset/liability, the total value/balance/amount owed is collected:
  - As of today
  - As of one year ago today
- Bracket questions are asked when the respondent cannot provide a specific value.

The team considered several methods:

- Survey of Consumer Finance method (multiple imputation, iterative process)
- Regression trees
- Hotdeck

…But none is feasible.

Going back to the original motivations (charter):

A system based on income imputation processing is being investigated.

- Regression-based, multiple imputation of each component asset/liability, from which "total change in" values can be derived.
- For each component, separate models are run for demographic groups across which large variation in parameter estimates is observed or expected.

For example:

Consider IRAX.

- Amount reported when asked: "What is the total value of all retirement accounts such as 401(k)s, IRAs, and Thrift Savings Plans that you or your household own/owns?"
- Expected to vary considerably by age
- Preliminary tests support use of one model for each age group (group 1: age<A; group 2: A<=age<B, etc.)

Bracket imputation will also be used:

- Respondent identifies range in which asset/liability falls (e.g., less than $X; $X to $Y; etc.)
- Five values are selected based on current methods used in income imputation; each falls within the specified bracket range.
- Open-ended brackets ($Z or over) also are treated in income imputation

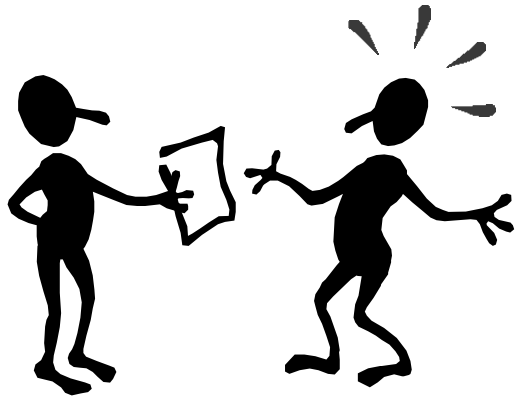
Related Challenges:

- How to distinguish $0 meaning no balance from $0 meaning no account.
- How to identify groups upon which to base models. That is:
  - Grouping variable: Is IRAX (e.g.) based on age, occupation, or something else?
  - Variable range: If age, where do the breaks occur—under 35, 35 to 64, 65 and older, or under 25, 25 to 34, etc.?

Work in progress:

- Identifying groups, and selecting variables to include within each model.
  - ANOVA/Chow tests have been used so far to test differences/pooling potential across groups.
  - Variables used in income imputation are considered the "starter group," with some to be added, deleted, or redesigned. (Example: Age ranges used in binary variables could be widened or narrowed.)

Comments/Suggestions are welcome!



Meanwhile, here are some questions for the Technical Advisory Committee (TAC):

1. Before implementation, BLS would like TAC's comments regarding whether there are aspects of the Westat recommendations that BLS should evaluate further or consider during implementation?
2. Does TAC have suggestions for validating imputations / ensuring imputations are within expected ranges as Westat recommends?
3. Does TAC have suggestions for imputing missing values for accounts with a zero balance?

Contact Information

Geoffrey Paulin, Ph.D.
Senior Economist
Consumer Expenditure Survey Program
www.bls.gov/cex
paulin.geoffrey@bls.gov

Investigating the Imputation of Assets and Liabilities in the CE Interview Survey

Discussant: Rebecca Andridge, PhD

BLSTAC Meeting – May 19, 2023

General Comments

- Complex, nested data structure challenging for imputation
- Love to see multiple imputation used (vs single)!
- Westat evalulation included some complex methods… that didn't outperform the simpler approach
  - Curious as to why hot deck (or PMM) was ruled out as an option
- Practicality a hugely important component (e.g., stick with SAS)
- The pessimist in me suspects that observed A&L values are not observed without error, so trying to find the "perfect" imputation model may be a fruitless endeavor – strive for "good enough"

1. Aspects of the Westat recommendations that BLS should evaluate further/consider

- Separately impute types of A&L "independently and in any order"
  - Is there correlation between A&L holdings? This is ignored if impute independently (unless zero correlation after conditioning on variables used in imputation model/groupings, e.g., age).
  - If substantial correlations, consider using previously imputed A&L to impute subsequent A&L (doesn't have to be fully FCS-style)

- Whether or not to multiply impute the previous year's indicators
  - Theoretically, these should be multiply imputed (as in Westat methods 2 & 3)
  - Creates more complex data structure

1. Aspects of the Westat recommendations that BLS should evaluate further/consider

- Transformation of Y when imputing account values (w/o brackets)
  - Westat evaluated multiple methods, including normalization and more complex transformations (and many model selection methods)
  - My opinion: simpler is better – reduce heteroskedasticity but don't overfit

- Imputing untransformed account values within brackets with uniform probabilities over the range
  - Appears this is the plan? (as in Westat evaluation)
  - May not be ideal – is a uniform distribution within brackets reasonable assumption? (It's not a good assumption for income….)
  - Consider alternative methods (e.g., lognormal interpolation, hot deck) if uniform distribution not empirically supported

1. Aspects of the Westat recommendations that BLS should evaluate further/consider

- Consider important domains
  - Make sure imputation procedure is congenial to key analyses
  - Estimates by occupation? Use occupation to define imputation cells.

2. Suggestions for validating imputations

- Really hard problem!
- Track donor pool sizes (and number of recipients imputed)
- For regression models, track model fit measures
- Comparing imputed to observed distributions can be misleading (if MAR, distributions will differ)
- Visualizations of imputed distributions (for each MI dataset) to look for outliers/obviously “weird” results
  - Indicators may be “easier” to monitor than values
- Fraction of Missing Information (FMI) for key outcome measures (proportions with each A&L type, totals) may be a useful metric, though with small number of imputations this is a noisy measure

3. Suggestions for imputing missing values for accounts with a zero balance

- Questions:
  - Does this happen with *current* values? (not per protocol)
    - Does not appear that $0 current value can be reported
    - Westat modified bracket to include $0
    - I don’t love imputing a value that cannot be directly observed ($0 current)
  - For previous year value, can we assume if CU provided a (non-zero) bracket that the account is not $0?

- For specific types of A&L where $0 makes sense, could impute 3-level indicator instead of 2-level for indicator {No, Yes/$0, Yes/>$0}

Imputing Assets and Liabilities in the CE Interview Survey

BLS TAC Meeting, May 19, 2023

Kevin B. Moore
Federal Reserve Board

The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors of the Federal Reserve System.

Why multiple imputation?

- To create a “complete” dataset that has no missing values
  - Alternative is to remove observations with missing values from estimates
- Imputations condition on existing relationships in the data
  - But what about the error in the imputed estimate?
- Multiple imputation provides a way to measure the variability in the imputed estimates (Rubin, 1987)

- CE income data have been multiply imputed since 2004

Implementation

- Westat report finds that all three imputation methods produce similar results
  - Recommend using the method that is similar to the current income imputations methodology
  - Recommend possibly updating the code/software
- Benefits of building on existing code and methodology
  - Easier to graft into the data production process
  - Still additional work for current staff, impact on production schedule?
- Users are already familiar with imputed data

Imputation issues – bracket vs. missing

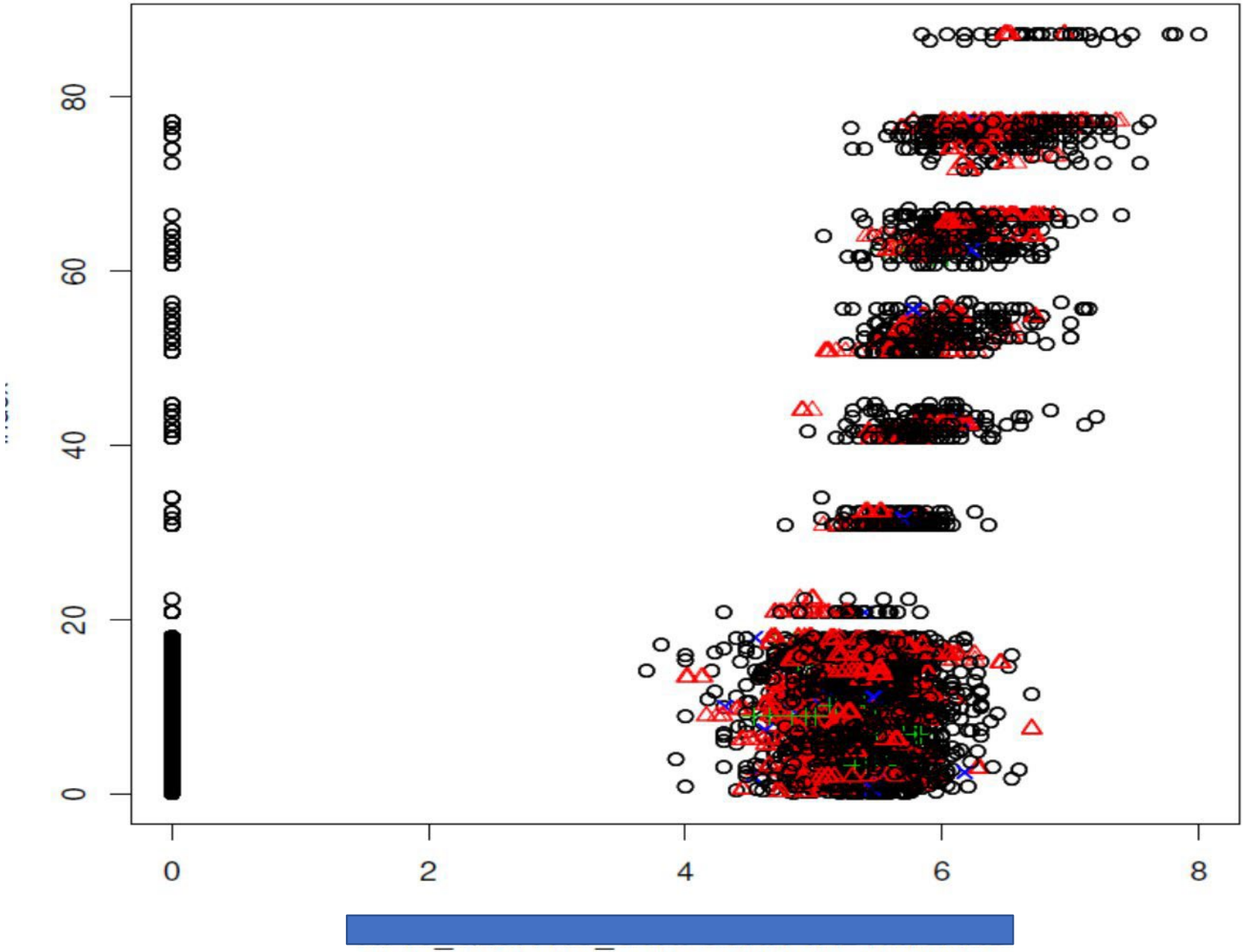
- Bracket values
  - Algorithm designed to preserve the mean value of observations that reported an actual value within the bracket range
- Missing values
  - Regression-based approach, coefficients and predicted values are both “shocked” to add random noise to the imputed value
- Why use two different approaches?
  - Incorporate bracket info into the regression-based approach
  - For predicted value outside the bounds, repeated draws from the error distribution

Imputation issues – zero balance accounts

- Zero balance is a legitimate response for a few items
  - Credit cards
    - 2019 SCF = 79% of hhs had some type of credit card, and 56% of cardholders had a revolving balance (not including convenience use)
    - CE question appears to ask “total amount owed today”, if includes convenience use than less zero balances
  - SCF imputations for credit card balances (for hhs with a credit card)
    - Impute binary outcome balance/no balance
    - If balance, then impute balance, otherwise set balance to zero
  - Also possible for checking/savings/money market/CDs, but very unlikely

Imputation issues – validation

- Logical constraints
  - Built into imputation routines whenever possible, based on logical relationships in the data
- Out of bounds imputations
  - SCF uses same imputation procedure for range values and missing values
  - Flagged in imputation output, warnings if value is “pushed” to upper or lower bound
- Graphical analysis
  - R programs



O = original value
X = edited value
+ = imputed value
△= range value

Thank you

kevin.b.moore@frb.gov