Imputation and Allocation in Consumer Expenditure Survey Data

Meaghan Smith
Division of Consumer Expenditure Surveys
CE Microdata Users Workshop
July 14 – 17, 2015
Washington, DC

BLS
BUREAU OF LABOR STATISTICS
U.S DEPARTMENT OF LABOR

www.bls.gov

Outline

- Outlier Review
- Process Overview
- Imputation
- Allocation

Outlier Review

- Check for and validate extreme values
- Confirm that the record is not misclassified

BLS

Outlier Review

- Outliers are determined using three methods.
  - Largest gap
  - P-index
  - Z-score

Outlier Review

- Correction of an outlier is based on a review of:
  - Consumer Unit characteristics like Income, Demographics, Area, Family size
  - Description of the expense
  - Audit trail investigation
  - Historical range of the category expense
- Values updated by either:
  - Correcting value based on examination of available information in the Consumer Unit
  - Set to invalid and impute later in processing

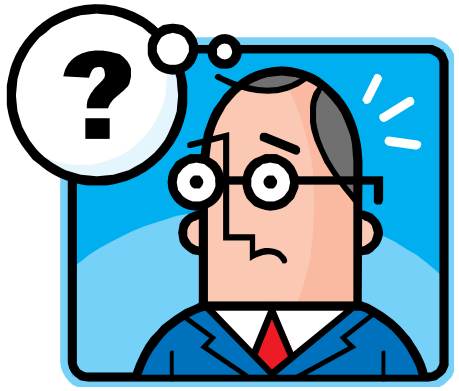
BLS

Process Overview

- The goal of our expenditure processing system is to map an expenditure report
  - As a monthly amount
  - To a specific Universal Classification Code (UCC) and
  - To a specific month and year
- However, we do not always get the detail needed to map the expenditures.
  - Consumer Unit cannot or will not provide the detail
  - Field representative does not collect/enter the detail

BLS

So what do we do in these cases?



Process Overview

- The Consumer Expenditure Survey uses two techniques to fill in missing data
  - Imputation
  - Allocation
- Within these techniques we use:
  - Cell defining variables
  - Sufficiency criteria
    - Both controls vary by questionnaire section and variable

BLS

Process Overview

- Imputation procedures:
  - Hot Deck Imputation
  - Hot Deck Imputation at Item Level
  - Weighting Class Imputation
  - Percentage Distribution Imputation
  - Month Imputation
- Allocation Procedures
  - Allocation using Distribution Ratios
  - Allocation using Fixed Ratios
  - Allocation using Probability Distribution Ratios
  - Allocation using Distribution Ratios of Reported Targets

BLS

IMPUTATION

Imputation

Hot Deck Imputation

- This method uses valid entries from the current quarter to assign values in fields containing invalid values
- The source records for this imputation are records with similar characteristics

BLS

Hot Deck Imputation Example

- A respondent reports that a consumer unit member purchased a men's jacket, but does not know the amount

- Imputation steps:
  - A random expenditure report for this specific item (JACKET) and specific age-sex classification (MEN) is selected by matching values of:
    - REGION
    - AREA TYPE
    - INCOME CLASS
  - The expenditure amount from the randomly selected record is assigned to the invalid cost variable on the original record

BLS

Method 1 in Concert Example

- A respondent reports vehicle purchase, but does not know the principle amount borrowed

- Imputation steps:
  - A random expenditure report is selected from the cell defined by matching values of:
    - Vehicle Type (car, truck, motorcycle, etc.)
    - New or Used Vehicle
    - Business Use Percentage
    - Income Class
  - The principle amount borrowed (along with the other financing data) from the randomly selected record is assigned to the record needing imputation (overlaying reported values in some cases)

BLS

Method 1 Same Schedule Example

- A respondent reports three monthly cable television bills, but only knows the bill amount for two of the months

- Imputation steps:
  - A random expenditure report is selected from the valid reports for the same consumer unit and same expenditure code (cable television)
  - The expenditure amount from the randomly selected record is assigned to the original record

BLS

Weighting Class Imputation

- Applies the mean of reported expenditures within a given cell to impute the missing or invalid expenditures in the same cell
- Source records for this imputation are records with similar characteristics

Weighting Class Imputation Example

- A respondent is unable to provide their usual monthly amount for dining out

- Imputation steps:
  - The weighted mean expenditure is calculated for the cell defined by matching values of:
    - Income Class
    - Family Size
  - The weighted mean expenditure amount is assigned to the original record

BLS

Percentage Distribution Imputation

- This method randomly selects a valid value from the weighted distribution of source record values
- The source records for this imputation are records with similar characteristics
- Primarily used for attribute variables or variables used to classify expenditures into more detailed levels

BLS

Percentage Distribution Imputation Example

- A respondent is unable to say whether some of the average monthly gasoline expenditure is for business

- Imputation steps:
  - The percentage and cumulative percentage distributions of calibration weights are derived for the attribute code in the cell defined by matching values of:
    - Income Class
  - A random number is selected between 0 and 1 and the attribute with the range containing the random number is assigned to the original record

BLS

ALLOCATION

BLS

Distribution Ratio Based Allocation

- This method is used for allocation of both reported and imputed combined expenditure to five or less component items
- Randomly selects a valid value from the weighted distribution based on weighted means of source record values
- The source records for this imputation are records with similar characteristics where the consumer unit reported separate values for each item

BLS

Distribution Ratio Based Allocation Example

- A respondent reports a combined expense for infant's clothing of $150, but does not know the specific items included in the purchase

- Allocation steps:
  - The percentage distributions of weighted means are derived for all target items in the cell defined by matching values of:
    - Income Class
    - Region
    - Area Type
  - The combined expense is allocated to all targets based on the percentage distribution

BLS

Fixed Ratio Based Allocation

- This method is used for gas and electric proportions for non-vacation properties if distribution ratio based allocation fails
- The allocation proportions are derived from the previous calendar year's Interview survey data

Probability Distribution Ratio Based Allocation

- This method uses percentiles for each component item in determining which component items are eligible for allocation
- This method is used for allocation of both reported and imputed combined expenditures to more than five component items

BLS

Probability Distribution Ratio Based Allocation

- If the combined expenditure is less than the sum of the weighted median, allocate to a subset of component codes. Otherwise, allocate to all component codes

- For records with combined expenditures less than the sum of the weighted median, select a subset of target components from the percentile that has at least two item codes with expenditures less than or equal to the expenditure of the record requiring allocation

BLS

Probability Distribution Ratio Based Allocation

- Target components are randomly selected without replacement from the weighted cumulative distribution ratio
- After each selection, we subtract the median expenditure of the selected component from the reported expenditure requiring allocation and continue with the next selection
- Once the selection process is complete, allocation is carried out based on percent distributions derived from weighted expenditures for each component item which met the selection criteria

BLS

Probability Distribution Ratio Based Allocation

- Once the selection process is complete, allocation is carried out based on percent distributions derived from weighted expenditures for each component item which met the selection criteria

Probability Distribution Ratio Based Allocation Example

- A respondent reports combined appliance purchases but did not specify the items included in the purchase

- Allocation steps:
  - Calculate the percentiles for all possible target components for the combined item code in the cell defined by matching values of:
    - Income Class
  - The selection process randomly selects 6 of 12 targets including stove, refrigerator, microwave, dishwasher, garbage disposal, and range hood
  - The combined expense is allocated to selected targets based on the percentage distribution

BLS

Distribution Ratios to Reported Targets Allocation

- Beginning with 2005Q2, CE began collecting the specific item components of a combined expense
- This data allows CE to allocate expenditures to the specific items purchased by the Consumer Unit in lieu of applying one of the previous allocation methods that use all components or a random sample
- This method is used for allocation of both reported and imputed combined expenditure to two or more specified targets

BLS

Distribution Ratios to Reported Targets Allocation Example

- A respondent reports a $500 clothing expense that includes pants, shirts, and shoes

- Allocation steps:
  - The percentage distributions of weighted medians are derived for the specified target items in the cell defined by matching values of:
    - Age-Sex Classification
    - Income Class
    - Region
    - Area Type
  - The combined expense is allocated to the specified targets based on the percentage distribution

BLS

Imputation of Rental Equivalence

- Beginning with processing of 2007Q2, a regression model was implemented to impute for missing rental equivalence
- Two different models
  - A model for owned homes
  - Another model for vacation properties

Imputation of Rental Equivalence

- Both models use property value and a number of demographic and housing characteristic variables
- Property value and demographic variables may have to be imputed.
- Top coding is performed after imputation to account for any outliers

Meaghan Smith
Edit and Estimation Section Chief
Division of Consumer Expenditure Surveys
www.bls.gov/cex
smith.meaghan@bls.gov

BLS
BUREAU OF LABOR STATISTICS
U.S DEPARTMENT OF LABOR