Imputation and Allocation of CE Data

Clayton Knappenberger
Economist
Division of Consumer Expenditure Surveys
2017 CE Microdata Users' Workshop
July 19 - 21

Outline

1. Process Overview
2. Data Screening
3. Imputation
4. Allocation
5. Questions and Contact Information

Process Overview

- CE's goal is to map expenditures
  - As monthly amounts
  - To specific Universal Classification Codes (UCCs)
  - In a specific month and year
- However, data quality is not always sufficient to meet this goal
  - Respondent does not know or refuses to provide
  - Collected information has mistakes

Process Overview

1. **Data Screening** – check data for errors
   - Misclassification
   - Outliers
2. **Impute** missing values
3. **Allocate** combined expenditures to components for mapping.

Data Screening

Misclassified Records

- Specific keyword lookups for “hard to classify” items
  - iPad/iPhone/iPod
  - “Glasses”/“Cable”/”Nails”
- Identified through outlier reviews
- New process in development to use text descriptions to identify misclassified records.

Outlier Review

- Three different methods are used to identify expenditures with extreme values
  1. Largest Gap
  2. P-Index
  3. Z-Score

Outlier Review

- Correction of an outlier is based on:
  1. Consumer Unit characteristics: *income, demographics, geographic location*
  2. Text description of the expense
  3. Interview metadata
  4. Historical range of the expense
- Updates are made by:
  1. Correcting value based on available information
  2. Flagging the expenditure for later imputation

Imputation

1. Hot Deck Imputation
   - Use valid records with similar characteristics to replace missing values
2. Weighted Mean Imputation
3. Percent Distribution Imputation
   - Randomly select a valid value based on the percent distribution of reported values

Hot Deck Imputation Example

- A respondent reports buying a men's jacket, but does not know the cost
- Imputation steps:
  - Select a valid random men's jacket expenditure from all such purchases with the same:
    - Region
    - Area Type
    - Income Class
  - The selected record's expenditure amount is copied to the record being imputed

Weighted Mean Imputation

- Use valid records with similar characteristics to define cells
- Calculate the weighted mean of that cell
- Assign the weighted mean of reported expenditures within a given cell to missing or invalid expenditures in the same cell

Percent Distribution

- A respondent is unable to say how many people are covered by their insurance plan
- Imputation steps:
  - Create weighted percent and cumulative percent distributions for “number of people covered” by matching values of income class
  - Generate a random number between 0 and 1
  - Find the value for “number of people covered” whose range includes the random number
  - Assign that value to the original record

Allocation

- Example: Respondent reported spending $500 on clothing
- Two main kinds of allocation:
  1. Reported Targets
  2. Unreported Targets

Reported Targets

- A Respondent reports a $500 clothing expense that includes (A) Pants (B) Shirts and (C) Shoes
- Allocation steps:
  - Derive percent distribution ratios using weighted medians for the specified targets by matching values of:
    - Age-Sex Classification
    - Income Class
    - Region
  - Allocate the $500 to each of the targets based on the percent distribution ratio

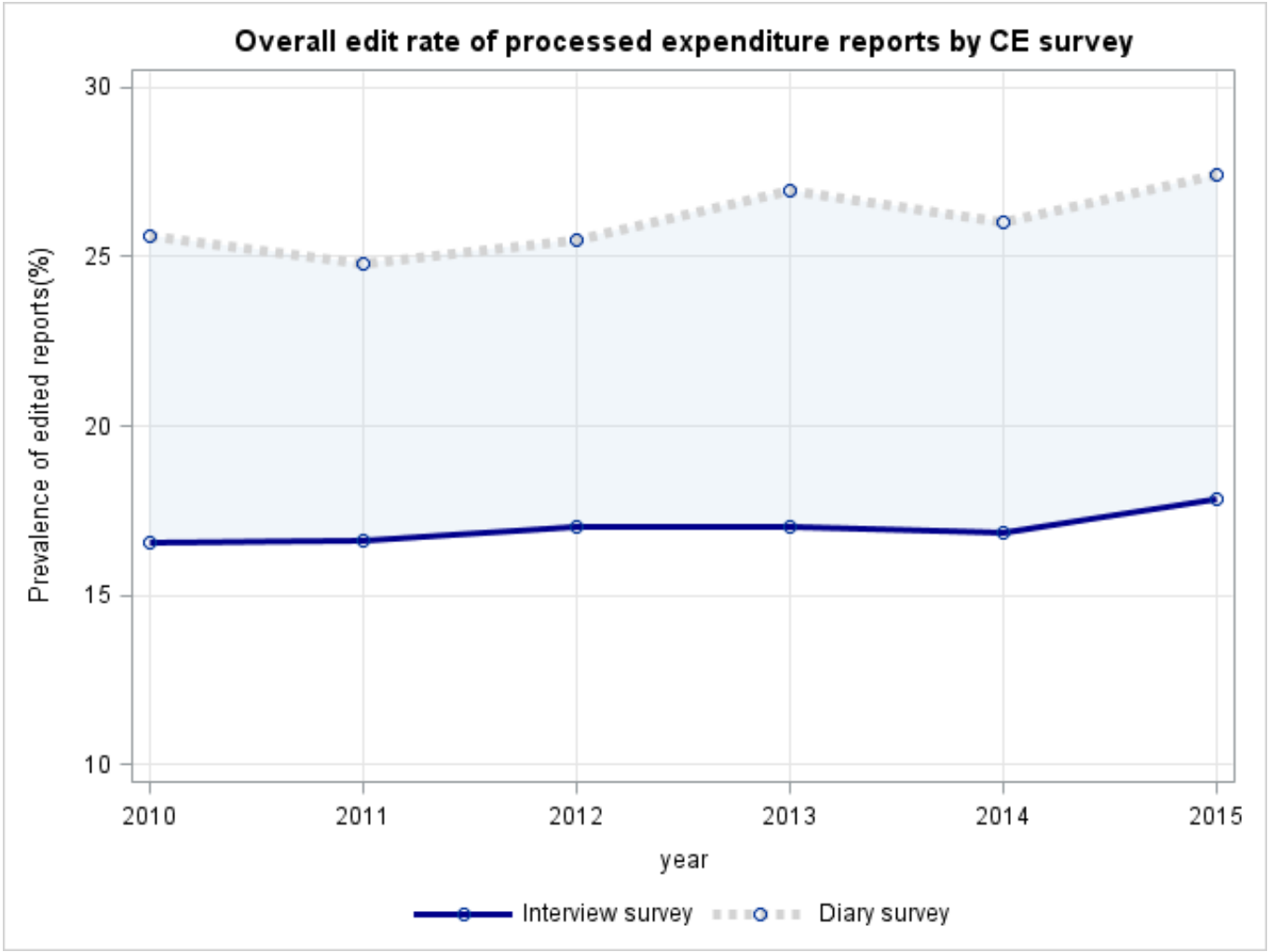
Unreported Targets (targets <= 5)

- A respondent reports $500 for clothing but does not specify what is included
- Allocation steps:
  - Derive weighted percent distributions for all target items by matching values of:
    - Income Class
    - Region
  - The $500 is allocated to all targets based on each target's allocation share in the percent distribution

Unreported Targets (targets > 5)

- Select Two or more targets
  - Calculate weighted cumulative frequency distributions for all the target items
  - Generate a random number between 0 and 1 to select the first target
  - Do this until the sum of the weighted medians is greater than or equal to the reported amount.
- Carry out allocation using percent distributions and allocation shares

Imputation and Allocation Rates



Why Impute and Allocate?

Benefits

- Meet internal needs for mapping
- Provide complete datasets to users
- Unbiased mean and variance

Concerns

- Our methods rely on MAR assumption
- Potential for underestimated variance

Contact Information

Clayton Knappenberger
Economist
Division of Consumer Expenditure Surveys
www.bls.gov/cex
202-691-6236
Knappenberger.clayton@bls.gov