

Imputation and Allocation in CE Microdata

Evan Hubener

Consumer Expenditure Survey
Microdata Users' Conference
July 15-18, 2014



Components

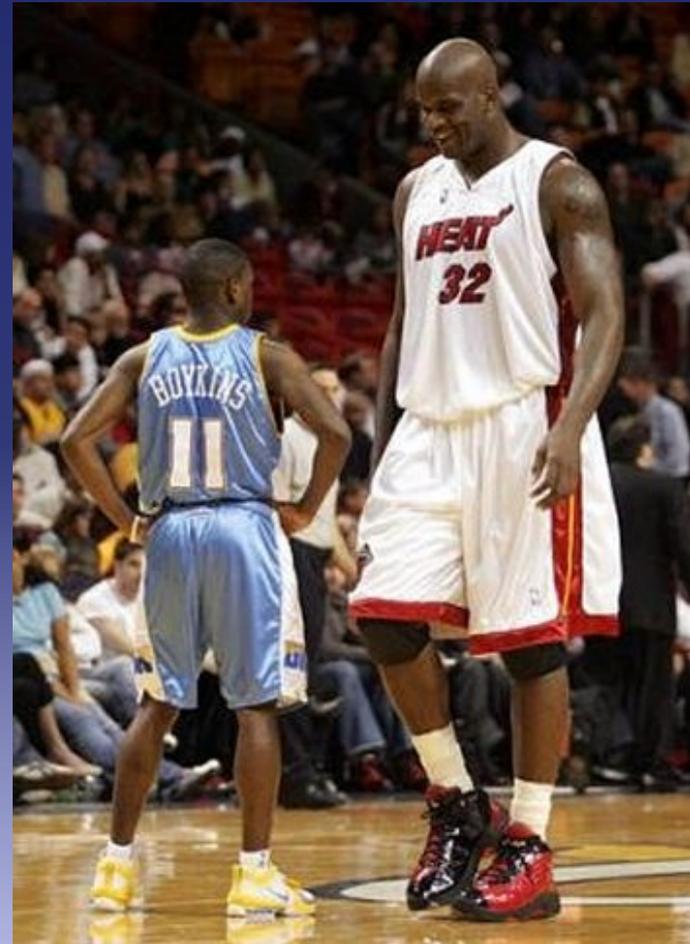
Data Screening

Imputation

Allocation

Data Screening

Screen for outliers
and
misclassifications



Data Screening

- Reviews of data are based on
 - Consumer Unit characteristics
 - Interview audit trail report, field notes, and descriptions
 - Range of the category expenditures
- Data may be updated by either:
 - Manually correcting the cost or classification
 - Imputation and Allocation processes

Objectives

CE Objective: Map all expenditures to

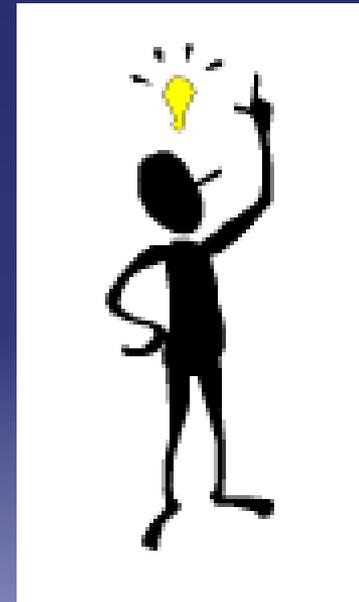
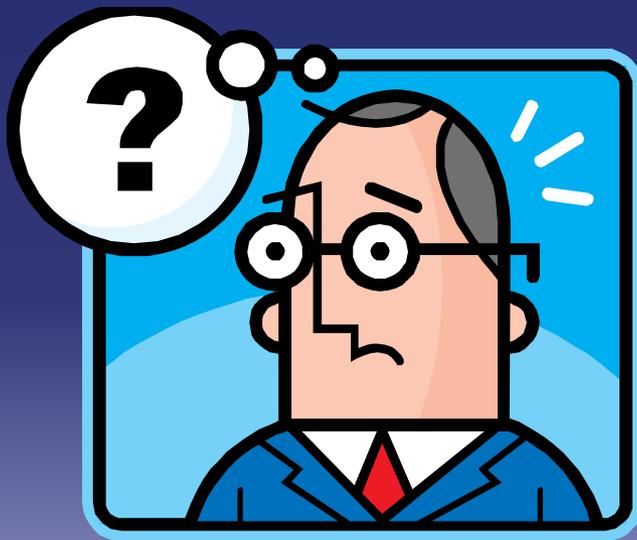
1. a specific Universal Classification Code (UCC)
2. a specific month and year

What if we are missing any details?

Consumer Unit does not provide the detail

Field representative does not collect the detail

So what do we do?



Solutions

CE uses two techniques to fill in missing data

Imputation

Allocation

Apply controls to limit imputation and allocation

Certain 'by variables'

Sufficiency criteria

minimum number of observations in a subset, etc

Imputation



Imputation Methods

Hot Deck Imputation
Weighting Class Imputation
Percentage Distribution Imputation

Imputation Methods

Hot Deck Imputations

Uses valid entries from the current quarter to assign values in missing fields
The source records are those with similar characteristics (Income, PSU, etc)
Some variation in methods, depending on the section

At Item Level

Match to a record with similar characteristics and same type of expenditure

At UCC Level

For records with multiple components

Match to a record with at least two components in common

Hot Deck Imputation Examples

A respondent reports that a purchase of a man's jacket, but the price is missing:

A random expenditure report for this item (jacket) and age-sex classification (male, 16+) is selected from a like subset of values and assigned to the missing record

Hot Deck Imputation Examples

A respondent reports financing a vehicle purchase, but the loan principal is missing:

Imputation steps:

- Other records with similar attributes are used to create a similar subset of data
 - Vehicle Type
 - New or Used Vehicle
 - Business Use Percentage
 - Income Class
- The principal amount borrowed and other financing information from the randomly selected record is assigned to the record needing imputation (overlying reported values in some cases)

Hot Deck Imputation Examples

A respondent reports three monthly cable television bills, but only reports the values for two months:

Imputation steps

- A random expenditure report is selected from one of the valid reports for the same consumer unit and same expenditure code
- The expenditure amount from the randomly selected record is assigned to the original record

Weighting Class Imputation

Weighting Class uses the mean of reported expenditures for a given subclass to impute missing data

Data is again subset so that source data have similar characteristics

Weighting Class Example

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

Solution:

Subset the data by income, family size, and PSU

Calculate the weighted mean expenditure

Assign it to the original invalid record

Percentage Distribution Imputation

Used for Variable Attributes

age/sex, gift/non-gift, new/used

Value is assigned randomly according to the distribution for the given attribute

Percentage Distribution Example

A respondent does not report whether a vehicle purchase is new or used:

Solution:

- The percentage distributions for the attribute are derived for values with similar characteristics
- Based on the distributions, a random number is selected, and the corresponding attribute is assigned

Allocation



Allocation Procedures

Distribution Ratio

Fixed Ratio

Probability Distribution Ratio

Distribution Ratio Allocation

Used for allocation of both reported and imputed combined expenditures with up to five components

This method randomly selects a valid value from the weighted distribution based on weighted means of source record values

Source variables have similar characteristics to the record eligible for allocation, and components with reported values

Distribution Ratio Example

A respondent reports a combined clothing expenditure , but does not report the components

First, the percentage distributions are derived for all target items

Second, the expense is allocated to selected targets based on the percentage distribution

Fixed Ratio 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

For reported and imputed expenditures with six or more component codes

Uses the distribution for each component item in determining which component items are assigned in allocation

Target components are randomly selected from the weighted cumulative distribution ratio

Probability Distribution Ratio Example

- A respondent reports combined appliance purchases but the items in the purchase are not specified
- Allocation steps:
 - Calculate the percentiles for all possible target components for the combined item code in the cell defined by matching values of:

The selection process randomly selects 6 of 12 targets

The combined expense is allocated to selected targets based on the percentage distribution

Distribution Ratios to Reported Targets

The Consumer Expenditure Survey has collected components of combined expenses since 2005Q2

This eliminates the need to randomly select targets, improving data quality

This method is used for allocation of both reported and imputed combined expenditures to the two or more reported targets

Distribution Ratios to Reported Targets Example

A respondent reports a clothing expenditure of \$500 that includes pants, shirts and shoes

The percentage distributions of the weighted means are derived for the specified items for a subset of like variables

The combined expense is allocated to the specified targets based on the percentage distribution

Evan Hubener

Economist, Consumer Expenditure Survey

www.bls.gov/cex/

hubener.evan@bls.gov



www.bls.gov