Imputing Across Interviews: Balancing Time Savings with Data Quality

by

Geoffrey Paulin, Ph.D.

Senior Economist

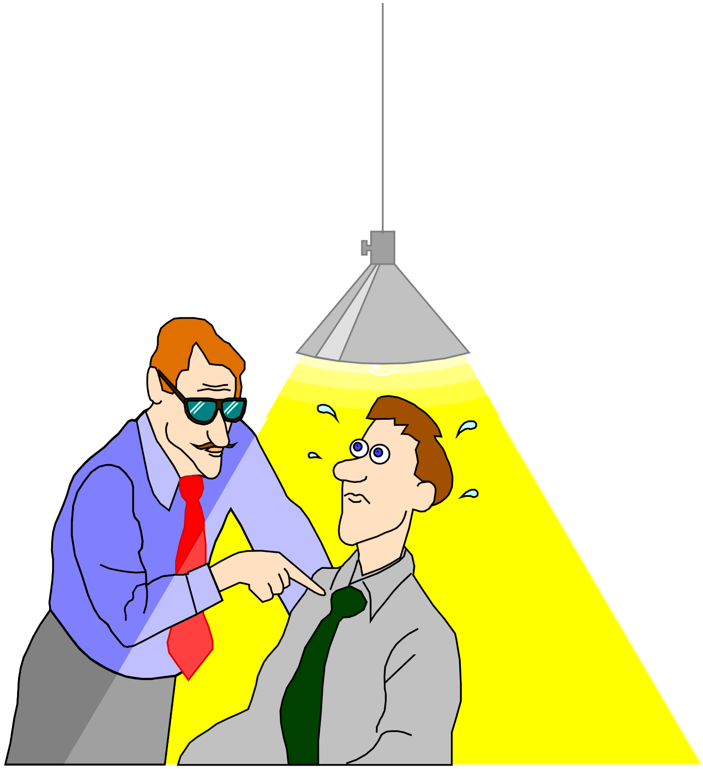
Consumer Expenditure (CE) Survey Program

CE Survey Methodology Symposium

July 16, 2013

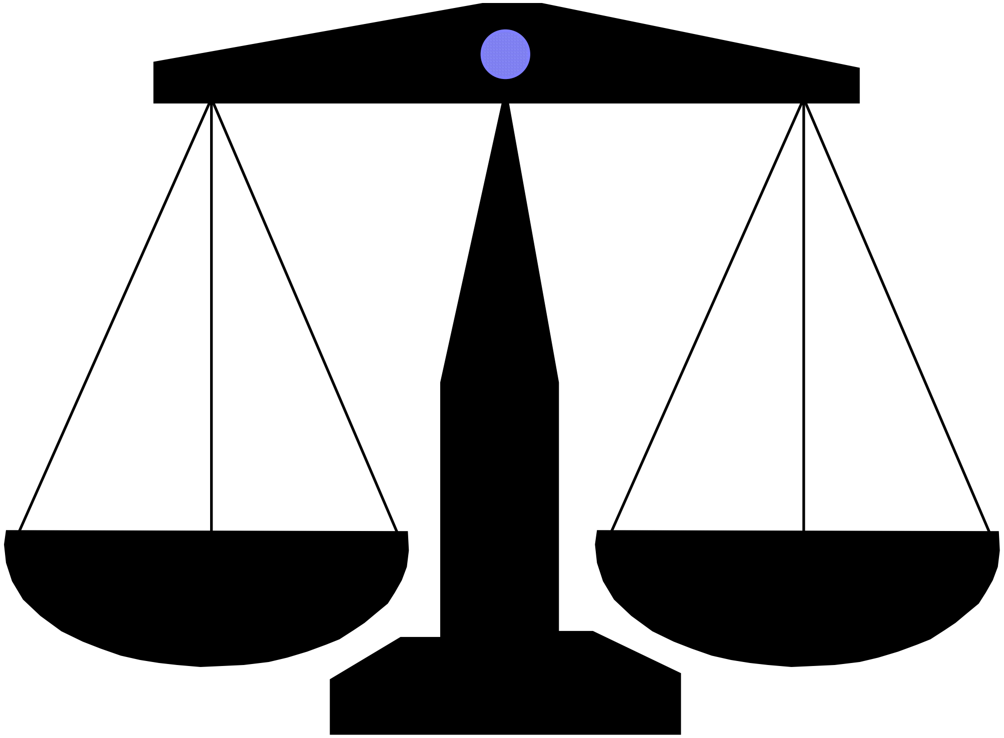
Washington, DC

Reducing respondent burden is an important goal…



…In 2011, the average quarterly interview was one hour; 10 percent exceeded 100 minutes.

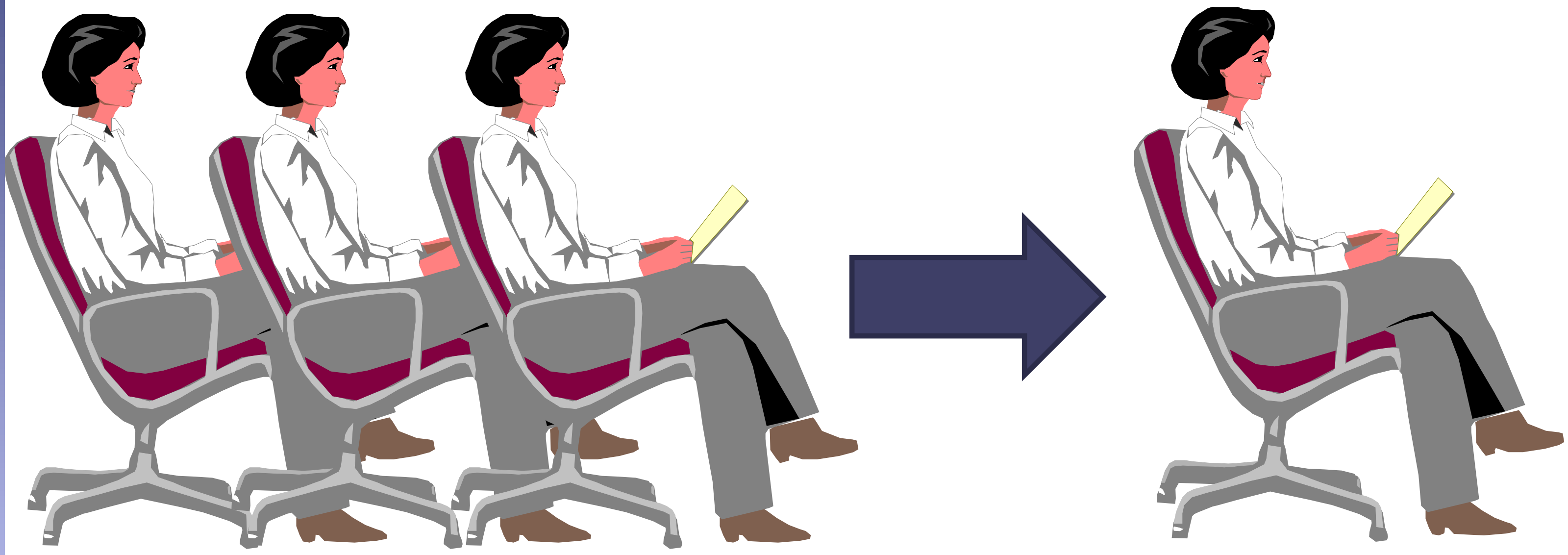
However, this must be balanced with maintaining high quality of data.



In 2015, the Bounding Interview will discontinue.

- CONSEQUENCES:
  - Need to add bounding information to current 2nd interview
  - Current 2nd interview time will increase, which was already shown to be a concern
- QUESTION: Can expenditures collected in the (current) 2nd interview be successfully imputed from (current) 3rd, 4th, & 5th interviews to minimize response burden?

To achieve this, the CE program is investigating the feasibility of imputing results from later interviews to the current second interview.



3rd 4th 5th

2nd

This presentations includes:

1. The conceptual framework currently being investigated
2. Problems encountered or anticipated
3. A request for comments

At present, there are three basic categories of expenditure under consideration:

1. Utilities (Electricity, Natural Gas, Fuel Oil/Other Fuel, Telephone, and Water)

2. Apparel

3. Those for which Interview respondents are asked are about “usual” weekly/monthly expenditures

1. Utilities

Reasons for considering:

- Section 4 is the most time consuming
- Expenditures are expected to occur each month, which makes processing easier (no need to decide in which month to place an expenditure; just allocate across the three)
- Expected to be highly correlated with explanatory variables already collected (housing size, types of appliances, region/State/PSU, urban/rural, city size)

2. Apparel

Reasons for considering:

- Burden reduction.
  - In 2011, 75 percent of consumer units interviewed reported expenditures for apparel and services (Section 9). In these cases:
    - Section 9 accounted on average for 6 percent of total interview time (almost 4 minutes), and increased with family size for consumer units up to 6 members.
    - 25 percent of reporters required more than 4½ minutes to complete the section; 10 percent required over 7½ minutes.
- Many items collected in both surveys are selected from the Diary for integrated publications.

3. Respondents asked about “usual” weekly/monthly expenditures

- Food at home
- Food away from home (except on trips)
- Alcoholic beverages at home
- Alcoholic beverages away from home (except on trips)

Reasons for considering (GLOBALS):

- In 2011, Section 20 is the second most time-consuming expenditure section
- All food and alcohol items are published from Diary
  - This indicates that quality of collected data is higher in Diary Survey than in Interview Survey.
  - Imputing data for food at home from Diary was investigated, but dropped due to poor quality of imputed results. However, matching from other interviews may produce higher quality estimates.
  - Food expenditures from the Interview Survey are required for supplemental poverty measures, and therefore information must be collected where possible.

Procedural Concerns and Clarifications:

- "Back Imputation"—that is, using reports from a specific consumer unit's 3rd, 4th, and 5th interviews to impute that consumer unit's 2nd interview is not feasible as it:
  - Causes delays in production (process cannot start until subsequent interviews have been completed;
  - Is still subject to nonresponse. (What happens if the unit participates in the 2nd, but no subsequent, interview?)
- For these reasons, regression using data from ALL consumer units participating in 3rd, 4th, and 5th interviews will be performed. Collection periods will be matched for source data. For example: 3rd, 4th, and 5th interviews from January of a given year will be used to impute 2nd interview values collected in January of that same year.

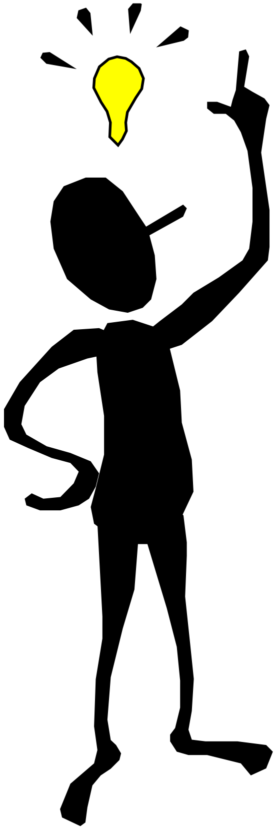
The “Yes/No” question will still be asked.

- That is, respondents will be asked whether or not the consumer unit incurred each expenditure, but not how much was spent if the answer is yes.
- This eliminates the need for a two-stage estimation procedure where the first stage predicts whether or not a purchase took place.
- As a result, the estimation process is much easier, and less prone to error, since the first stage is reported, not estimated.

As noted, expenditures shall be estimated by regression analysis.

- **Hot decking has been considered but rejected.**
  - Currently, hot decking is used when respondents report that an expenditure occurred, but not the amount. The team investigated the possibility of adopting this approach for the larger project.
  - However, the limitations of hot decking are well-documented (e.g., ability to use few predictor variables; effects on variance).
  - The limitations are less problematic for filling in nonresponse blanks, especially when item nonresponse rates are low. But in this case, all expenditures would be imputed.
  - The inability to properly preserve correlations among expenditures and independent variables would be detrimental to microdata users.
- **This means regression will be used.**

The first item considered is electricity.



Status:

- Regressions in progress
- Models are:
  - The most complicated so far. They include:
    - Standard demographics (age, family size, income)
    - Special variables such as—
      - Number/type of appliances in household, where known
      - Detailed geographic data as described earlier
      - Type of housing (detached, townhome, highrise, dormitory, mobile home, etc.)
  - Currently run separately for homeowners and renters, but may require further breakdown (e.g., housing tenure by region).

Next up:

Apparel

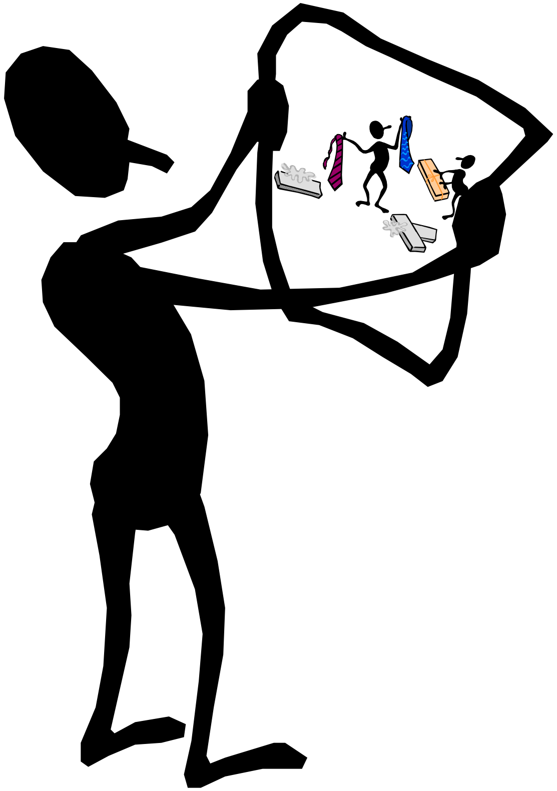
Apparel status

- So far, models require only standard characteristics variables
- However, may need separate models for different family types, etc.

And finally:

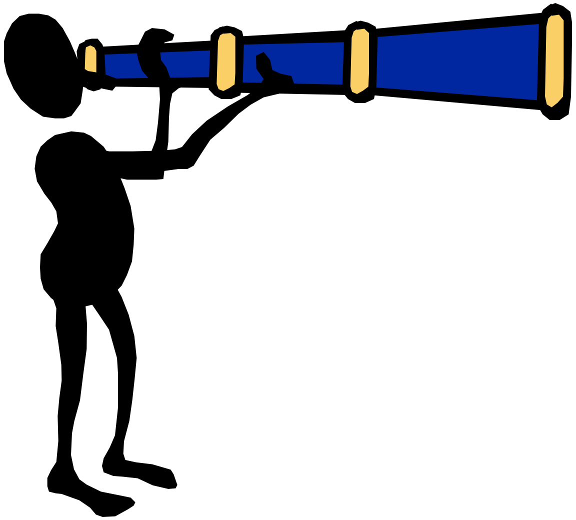
Food and Alcoholic Beverages

Globals status



Models have not yet been constructed, but they are expected to be similar to apparel.

Looking ahead:



General questions

- Quality assessment: Will the imputed values fall within acceptable ranges when subjected to testing?
- Are there qualitative differences in current 3rd, 4th, and 5th interviews (e.g., means or variance) at least when compared to current 2nd interviews for which the models need to account?
- What are the unintended consequences of replacing reported with imputed data?
  - Can/will covariate relationships be preserved (e.g., food at home with apparel)
  - If not, what are the implications for the supplemental poverty measures, and other important uses of the data?

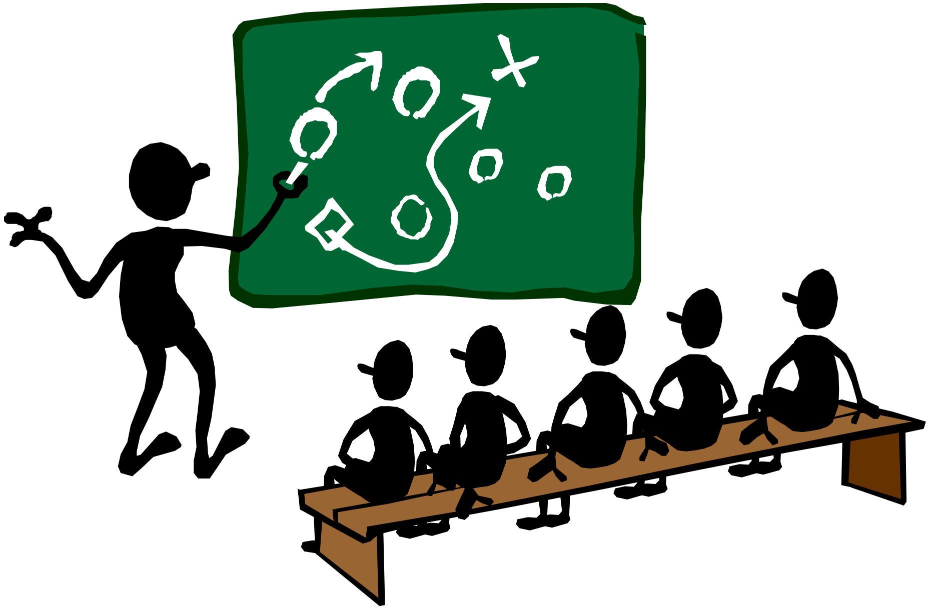
Technical Questions

- Should single or multiple imputation be used?
- If multiple imputation is used, what is the proper way to use income, which is itself multiply imputed?
  - Use average imputed income for each consumer unit the same way as a non-imputed variable would be used, generating five imputations of the Interview expenditure variable
  - Obtain a regression estimate using the five imputed income values; shock; repeat four times. In this way, 25 regressions yield 5 imputed expenditure values per consumer unit.

Next Steps:

- Continue refining models and assessing quality of results
- Receive and incorporate comments and suggestions from experts like you(!)
- Prepare an interim report on feasibility (October 2013)

If you have any suggestions, comments, or questions of your own…



…The team looks forward to hearing from you.

Contact Information

Geoffrey Paulin, Ph.D.

Senior Economist

Consumer Expenditure Survey Program

www.bls.gov/cex

202-691-5132

Paulin.Geoffrey@bls.gov

BLS

BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR