Project 8/9: User's Guide to Income Imputation in the CE

NOTE:

The content of this training presentation is derived from the "User's Guide to Income Imputation in the CE" (https://www.bls.gov/cex/csxguide.pdf)

In this project, you will learn:

- How income data are collected in the Consumer Expenditure Surveys (CE)
- Why and how these data are imputed when missing
- How imputation affects "you-sers"

Income Collection in the CE:

- Interview Survey
  - 1^{st} and 4^{th} interview (2015 onward; 2^{nd} and 5^{th} previously)
- Diary Survey
  - One time only (1^{st} or 2^{nd} 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.

Some Key Points:

- Reiteration:
  - Each income variable has not one, but five imputed values;
  - When reported, each imputed value equals reported value;
  - When bracket range is reported, imputed values differ, but all fall within the bracket range.
- New point:
  - When model-based imputation is used, values have no preset bound, but are always positive (except for variables like SEMPFRMX, for which losses are possible).

Some variables include summations of imputed values:

- FSMPFRX1:
  - Equals sum of SMPFRMX1 for each member of the consumer unit (CU); i.e., first “F” of “FSMPFRX1” is for “Family”
  - Some SMPFRMX1 within the CU may be valid reports, others imputed (bracket or model-based)
- FINCBTX1:
  - Sum of all components (including FSMPFRX1), imputed or not

NOTE: This applies to ALL "INCOMEm" variables:

- FSMPFRXm, m=imputation number, 1<=m<=5:
  - Equals sum of SMPFRMXm for each member;
  - Some SMPFRMXm within the consumer unit may be valid reports, others imputed (bracket or model-based)
- FINCBTXm, 1<=m<=5:
  - Sum of all components (including FSMPFRXm), imputed or not

To identify quantity of, and reason for, imputation, “flag variables” are available.

- Naming convention: End in “I”. Examples:
  - INTRDVXI
  - SEMPFRMI
  - FSMPFRMI

Possible Values:

- 100: No imputation on variable or subcomponents (i.e., variable or subcomponents are validly reported)
- 2nn: Only model-based imputation is performed on variable or subcomponents
- 3nn: Only bracket imputation is performed on variable or subcomponents
- 4nn: At least one model-based and at least one bracket imputation are performed (summary variables only)
- 5nn: "AVB" case

What does “nn” mean?

- Number of subcomponents imputed.
  - Always equal to “01” for non-summary variables (e.g., INTRDVX or SEMPFRMX).
  - Minimum of “01” for summary variables.
    - Example: A consumer unit has three members reporting receipt of SEMPFRMX.
      - One reports the value (SEMPFRMI=100).
      - Another provides no information (SEMPFRMI=201).
      - The third reports a bracket (SEMPFRMI=301).
      - FSMPFRMI is 402.

Important for CE Microdata Users:

- Once again, each income variable has not one, but five imputed values.
- Microdata users **must use all five values** to obtain valid results.

Computing Means

- Unweighted (i.e., Sample) Means

$$\left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)/(n \times m)$$

- *X* is the value of income from consumer unit i for imputation j *(where 1 <= j <= 5 in CE)*
- *n* is the number of rows *(varies by data set)*
- *m* is the number of columns *(always 5 in CE)*

Computing Unweighted Means

- Applied Example:

INTRDVX	INTRDVX1	INTRDVX2	INTRDVX3	INTRDVX4	INTRDVX5
100	100	100	100	100	100
D	50	250	300	20	80

  - INTRDVX is the value reported (or not).
  - INTRDVX1,…,INTRDVX5 are the values imputed.

1. Sum all imputed values, i.e., INTRDVX1,…,INTRDVX5 (100 + 100 + … 100 + 50 + … + 20 + 80);
2. Divide total (1,200) by total number of observations (n*m=2*5=10);
3. Mean = 120

Computing Unweighted Means

- Alternatively, use INTRDVXM:

INTRDVX	INTRDVX1	INTRDVX2	INTRDVX3	INTRDVX4	INTRDVX5	INTRDVXM
100	100	100	100	100	100	100
D	50	250	300	20	80	140

  - Find the mean of each row (INTRDVXM).
  - Add the row means: 100+140=240.
  - Dividing this by the number of rows (2).
  - Mean=120, the same value obtained by finding the mean of all 10 observations.

BONUS:

- INTRDVXM is already computed for you on the PUMD files!
- This makes it easy to compute either a sample mean (as just demonstrated), or a weighted mean (as you shall soon see…).

Computing Weighted Means: Estimating Population Means

- Consider the following data:

INTRDVX	INTRDVX1	INTRDVX2	INTRDVX3	INTRDVX4	INTRDVX5	INTRDVXM	FINLWT21
100	100	100	100	100	100	100	5,000
D	50	250	300	20	80	140	7,500

- Based on FINLWT21:
  - The first CU represents 5,000 units in U.S.;
  - The second CU represents 7,500.
- Using INTRDVXM and FINLWT21, the weighted mean is: [(100*5,000) + (140*7,500)]/(5,000 + 7,500) = 124.

Computing Variances: Unweighted (Sample) Data

- Five steps for computing variances for unweighted means:

INTRDVX	INTRDVX1	INTRDVX2	INTRDVX3	INTRDVX4	INTRDVX5
100	100	100	100	100	100
D	50	250	300	20	80

1. Compute the mean of each column of completed data (INTRDVX1 through INTRDVX5)
2. Calculate the average of the five means
3. Calculate the variance of the MEAN of each column of data
4. Calculate the average of these variances (of means)
5. Calculate the variance (actual, not variance of mean) *between* (or among) the five complete data mean estimates

Computing Variances

- Then, the final step is to insert them into the formula for total variance:

$$T_m = \overline{U}_m + \left(1 + m^{-1}\right)B_m$$

- T_m is total variance; SQRT(T_m) is the standard error
- U_m is the variance of the MEAN of the mth column; and $\overline{U}_m$ is the average of these variances (of means)
- m is the number of columns (5 columns in this case)
- B_m measures the variance of the five means (one for each column)

Computing Variances

- So, in this example:

INTRDVX	INTRDVX1	INTRDVX2	INTRDVX3	INTRDVX4	INTRDVX5
100	100	100	100	100	100
D	50	250	300	20	80
MEAN	**75**	**175**	**200**	**60**	**90**

1. **Compute the mean of each column of completed data (INTRDVX1 through INTRDVX5): 75; 175; 200; 60; 90**
2. **Calculate the average of the five complete data estimates: (75+175+200+60+90)/5 = 600/5 = 120**
3. **Calculate the variance of each column of data: 1,250; 11,250; 20,000; 3,200; 200; divide each by 2 (because n=2) to get each variance of mean: U_1=625, …, U_5=100**
4. **Calculate the average of these variances of means (U_1+…+U_5)/5=3,590**
5. **Calculate the variance *between* (or among) the five complete data mean estimates: Var (75,…,90)=[(75-120)^2+ … +(90-120) ^{2}]/m-1, where m=5; Var (75,…,90)=3,987.5**

Computing Variances

- Then, in the final step is to insert them into the formula for total variance:

$$T_m = \overline{U}_m + \left(1 + m^{-1}\right) B_m$$

- U_m is the variance of the MEAN of the mth column; and $\overline{U}_m$ is the average of these variances: 3,590
- m is the number of columns (5 columns in this case)
- B_m measures the variance of the means of each of the five columns: 3,987.5
- T is the total variance:

  3,590 + (1+0.2)*3,987.5 = 8,375

Computing Variances: IMPORTANT NOTICE

- Proper computation of variances using multiply imputed data is more complicated than computing means.
  - You **MUST use all five columns** of imputed data do obtain the correct variance.
  - You **MUST NOT compute the variance of any old column** (INTRDVX1 only; INTRDVX2 only), **or even of the "M" column** (i.e., INTRDVXM) and call it a day.
- All right, don't listen to me.
  - But your variances will be biased, possibly quite seriously.
  - And the direction of the bias ("too large" or "too small") is not predictable!

Computing Regression Results

- To compute regression coefficients and standard errors, use repeated-imputation inference (RII).
  - The proper estimation uses all five implicates for income by estimating the regression model once with each implicate.
  - Estimating coefficients with RII is similar to mean estimation.
  - Estimating standard error with RII is similar to variance estimation.
- RII applies to both weighted and unweighted regression analysis. However, for simplicity, only unweighted regressions are described herein.

Computing Regression Results: Coefficients

- Objective: Compute $y = \alpha + \beta I + \gamma X + \varepsilon$ using imputed income.
- To obtain estimates of the α, β, and γ, the regression model is estimated five times, once for each implicate:
  - $y = a_1 + b_1(\text{FINCBTX1}) + g_1 X$,
  - $y = a_2 + b_2(\text{FINCBTX2}) + g_2 X$,
  - $y = a_3 + b_3(\text{FINCBTX3}) + g_3 X$,
  - $y = a_4 + b_4(\text{FINCBTX4}) + g_4 X$, and
  - $y = a_5 + b_5(\text{FINCBTX5}) + g_5 X$.
- Average $a_{1 \text{ to } 5}$ to get α; $b_{1 \text{ to } 5}$ to get β; and $g_{1 \text{ to } 5}$ to get γ.

Computing Regression Results: Standard Error (SE) of a Coefficient

- Objective: Compute SEs for α, β, and γ.
- Same steps as computing variance of income, except the coefficient is treated as the column mean. For example, to compute SE(α):
  - Compute the **VARIANCE** of a_1, then a2, …, then a_5. (Your computer software may do this. If not, it should provide the SE of each. Square SE for each coefficient a_1, …, a_5 to obtain the variance of a_1, …, a_5.)
  - Compute the **average** of the variances of a_1, …, a_5; call the result $\overline{U}_{m.}$
  - Compute the **variance of the VARIANCES** of a_1, …, a_5; call the result $B_{m.}$
  - Compute $T_m = \overline{U}_m + \left(1 + m^{-1}\right)B_m$
  - The square root of T_m *is the SE*(α).

This concludes the “basics” of using imputed CE income data. For more applications, see:

“User’s Guide to Income Imputation in the CE”

(https://www.bls.gov/cex/csxguide.pdf)

ATTENTION SAS USERS:

You have a “macro” available to compute the following exercises. Details in “special topics” after feedback session.

Project 8

1. Create data set containing data for a collection year from 4 quarterly FMLI files.
2. Find the unweighted mean for income using FINCBTXM for all CUs by region of residence:
   a. Northeast (REGION=“1");
   b. Midwest (REGION=“2");
   c. South (REGION=“3);
   d. West (REGION=“4");
   e. Suppressed (REGION=“ “)—affects selected combinations of PSU and STATE, done to maintain confidentiality
3. See next slide to compute standard error

Project 8

3. Same steps as computing variance of income, except the coefficient is treated as the column mean. For example, to compute SE(α):
   - Compute the **VARIANCE** of a_1, then a2, …, then a_5. (Your computer software may do this. If not, it should provide the SE of each. Square SE for each coefficient a_1, …, a_5 to obtain the variance of a_1, …, a_5.)
   - Compute the **average** of the variances of a_1, …, a_5; call the result $\overline{U}_{m.}$
   - Compute the **variance of the VARIANCES** of a_1, …, a_5; call the result $B_{m.}$
   - The square root of T_m *is the SE*(α).

Project 8 Results

Unweighted Mean & Standard Error (FMLI211-FMLI214)

REGION	Mean Income Before Tax	Total Standard error of Income
Suppressed	$66,779.98	$2,778.30
Northeast (1)	$98,551.01	$1,669.53
Midwest (2)	$84,908.21	$1,284.43
South (3)	$78,894.41	$1,081.29
West (4)	$93,865.96	$1,285.97

Project 9

1. Create data set containing data for a collection year from 4 quarterly FMLI files.
2. Find the income coefficient (i.e., "β") and its SE for Food at Home:
   a) Add FDHOMEPQ and FDHOMECQ, and annualize the resulting variable FDHOME: FDHOME=4*(FDHOMEPQ+FDHOMECQ)
   b) Regress FDHOME on FINCBTX1, …, FINCBTX5
   c) Compute mean and standard error of "β"

Project 9 Results

Regressions using unweighted multiply imputed data (FMLI211-FMLI214)

Type	Estimate	Total Variance	Total Standard Error
INTERCEPT	4,897.59	1923.53100	43.85808
Income Coeff. (MPC)	0.019	1.250216E-7	0.00035358