Consumer Expenditure Survey Microdata Overview and Practical Training

Aaron Cobet
Brett Creech
Scott Curtin
Bill Passero
Geoffrey Paulin
Arcenis Rojas
Division of Consumer Expenditure Survey

CE Microdata User's Workshop
July 15-17, 2015

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

www.bls.gov

Overview

- Introduction to the CE Microdata files
- Microdata basics
  - What's on the PUMD release?
    - Files
    - Naming conventions
    - Documentation
    - Detailed File structure
- Projects
  - Overview of theme: Healthcare expenditures in the CE
  - Projects:
    1) Basic statistics using FMLI file
    2) Introducing MEMI file
    3) Detailed expenditures from MTBI file
    4) Diary estimates using FMLD / Integration
    5) Expenditure characteristics in IHB file
    6) Population estimates / Weighting
    7) Calendar year estimates
    8) Standard errors

What’s included in the PUMD Release

Zipped Data Files:

- Interview “INTRVW” files, incl. paradata files
- Interview “EXPN” files (detailed expenditures)
- Diary Files

- File types available:
  - SAS (*.sas7bdat), STATA (*.dta), SPSS (*.sav), ASCII comma-delimited (*.csv)

Interview Survey Files

- INTRVW files
  - 5 quarterly files each
    - FMLI – CU characteristics, incl. summary expenditures
    - MEMI – Member characteristics
    - MTBI – Monthly expenditures
    - ITBI - Income
    - ITII – Imputed income
  - 2 9-quarter paradata files
    - FPAR – CU-level paradata – interview outcome
    - MCHI – Contact history
- EXPN files (1 5-quarter file each)
  - Information about the expenditure
  - ≈50 files, following the interview structure

Diary Survey Files

- Diary files (4 quarterly files each)
  - FMLD – CU characteristics
  - MEMD – Member characteristics
  - EXPD – Detailed expenditures
  - DTBD – Income
  - DTID – Imputed income

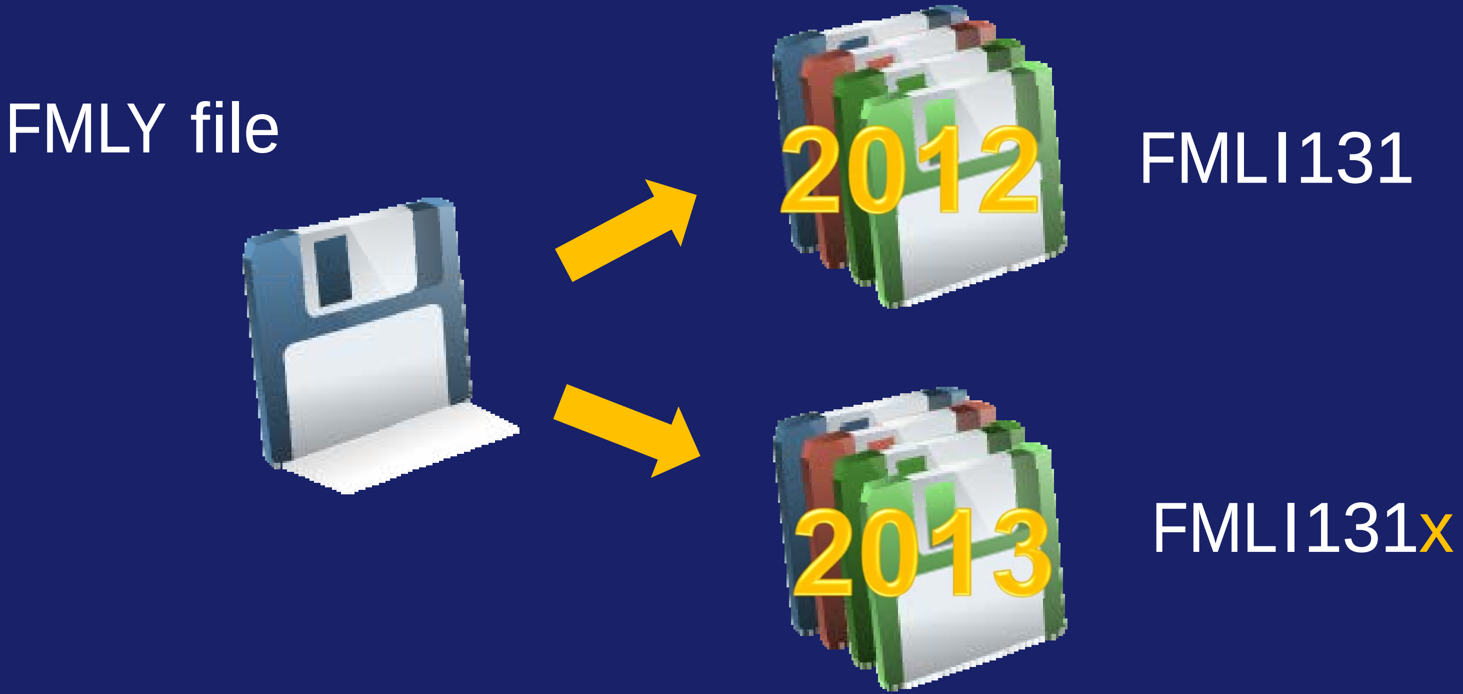
“X” Factor in Interview Survey

- Files for the first quarter of any calendar year will appear on two releases of CE microdata
  - “Fifth” file in the previous calendar year’s release
  - “First” file in the current calendar year’s release

- Are the files identical? Check the data set name.
  - If the files are identical, the data set name will be the same for both files
  - If the files are different, the data set names will be different and, in general, an “X” will be added to the end of the name for the current year’s release

“X” Factor in Interview Survey

- For example: Data *collected* in 2013q1:



FMLY file

FMLI131

FMLI131x

“X” Factor in Interview Survey

- What’s the difference
  - Topcoding/disclosure methods with processing year (means preserved for processing year)
  - Sometimes variables introduced/deleted between processing years
  - Sample redesign years (1995,2005,2015…)
    - *Datasets will be from separate sample designs*
- Which do I use with a time-series?
  - Your call, but we recommend being consistent.

Documentation

Documentation

- User's documentation
- Data Dictionaries
- Sample programs
- Other documentation/support files

User’s Documentation

- User’s Documentation
  - Interview User’s Documentation
  - Diary User’s Documentation
  - User’s Guide to Income Imputation in the CE
  - Getting Started Guide

User’s Documentation

- One for each survey
  - Things that change year to year:
    - Topcoding values, notes on disclosure requirements
    - Record counts
    - Variable additions, deletions, and modifications
  - Things that don’t change
    - Information about files
    - Definitions of flags (EXPN flags, COST_, Imputation flags)
    - Estimation procedures/formulas

Record Counts

- Number of observations on a data file
- For Interview CU Characteristics files (FMLI), denotes number of quarterly interviews completed (2013 record counts = 32,591)
- For Diary CU Characteristics files (FMLD), denotes number of diaries completed (2013 records counts = 12,335)
- Interview Monthly Expenditure, Income, and Imputed Income files are particularly large (400K – 600K+ records) per quarter (2013 record counts = Approx 4.7 million combined annual records)

Data Flags

- Interview EXPN files (from 2006)
  - “A” = valid blank in field where no response is expected
  - “B” = invalid blank indicating nonresponse that is inconsistent with other data reported by CU
  - “C” = invalid blank from a refusal, “don’t know” or other type of nonresponse
  - “D” = valid or good data value that is unadjusted
  - “E” = valid or good data value that has been allocated
  - “F” = valid or good data value that has been imputed or in some other way adjusted

Data flags cont’d

- EXPN Data flags cont’d
  - “G” = valid or good data values that has been imputed and allocated
  - “T” = data value has been topcoded or suppressed
  - “U” = data value has been allocated and then topcoded or suppressed
  - “V” = data value has been imputed or in some other way adjusted and then topcoded or suppressed
  - “W” = data value has been allocated and imputed, and then topcoded or suppressed
  - “H” = data value has been allocated to other records, original expenditure blanked out

Data flags cont’d

- Interview MTBI cost flag (COST_) (2007 – forward)
  - “D” = valid or good data value that is unadjusted
  - “E” = valid or good data value that has been allocated
  - “F” = valid or good data value that has been imputed
  - “G” = valid or good data values that has been imputed and allocated
  - “T” = data value has been topcoded or suppressed
  - “V” = data value has been imputed or in some other way adjusted and then topcoded or suppressed
  - “W” = data value has been allocated and imputed, and then topcoded or suppressed

Percentage of records reported directly, imputed, allocated, or imputed and allocated by record type, Interview Survey, 2013 Q1-2014 Q1

	Percent			
Record type	Directly reported	Strictly imputed	Strictly allocated	Allocated and imputed
APA	87.4	1.5	11.0	0.1
APB	96.4	0.5	3.1	0.0
CLA	44.1	0.4	54.9	0.6
CLD	98.2	0.5	1.3	0.0
CNT	97.6	2.4	0.0	0.0
CRA	94.9	1.1	4.0	0.0
CRB	90.0	1.1	8.9	0.0
EDA	95.8	0.8	3.4	0.1
ENT	98.3	1.7	0.0	0.0
EQB	97.1	0.7	2.2	0.0

Percentage of records reported directly, imputed, allocated, or imputed and allocated by record type, Interview Survey, 2013 Q1-2014 Q1 – Cont.

	Percent			
Record type	Directly reported	Strictly imputed	Strictly allocated	Allocated and imputed
FN2	92.9	7.1	0.0	0.0
FNA	94.1	5.9	0.0	0.0
FNB	57.3	42.7	0.0	0.0
FRA	88.9	0.5	10.4	0.2
FRB	99.1	0.9	0.0	0.0
IHB	87.4	12.6	0.0	0.0
IHC	100.0	0.0	0.0	0.0
IHD	56.0	44.0	0.0	0.0
INB	86.5	6.8	6.6	0.0
LSD	91.6	8.4	0.0	0.0

Percentage of records reported directly, imputed, allocated, or imputed and allocated by record type, Interview Survey, 2013 Q1-2014 Q1 – Cont.

Percent				
Record type	Directly reported	Strictly imputed	Strictly allocated	Allocated and imputed
MDB	92.4	1.3	6.3	0.1
MDC	93.3	2.0	4.2	0.4
MIS	97.6	0.9	1.4	0.0
OPB	80.8	19.2	0.0	0.0
OPD	79.9	20.1	0.0	0.0
OPF	79.8	20.2	0.0	0.0
OPH	77.6	22.4	0.0	0.0
OPI	78.1	16.1	5.8	0.1
OVB	93.0	7.0	0.0	0.0
OVC	91.1	8.9	0.0	0.0

Percentage of records reported directly, imputed, allocated, or imputed and allocated by record type, Interview Survey, 2013 Q1-2014 Q1 – Cont.

	Percent			
Record type	Directly reported	Strictly imputed	Strictly allocated	Allocated and imputed
RLV	99.1	0.9	0.0	0.0
RNT	95.7	4.3	0.0	0.0
SUB	99.1	0.9	0.0	0.0
TRB	83.4	3.0	13.6	0.0
TRD	0.0	0.0	97.6	2.4
TRE	0.0	0.0	99.2	0.8
TRF	78.8	2.4	18.8	0.0
UTA	58.8	37.8	3.4	0.0
UTC	67.4	2.2	29.8	0.6
UTI	64.1	1.0	34.5	0.5

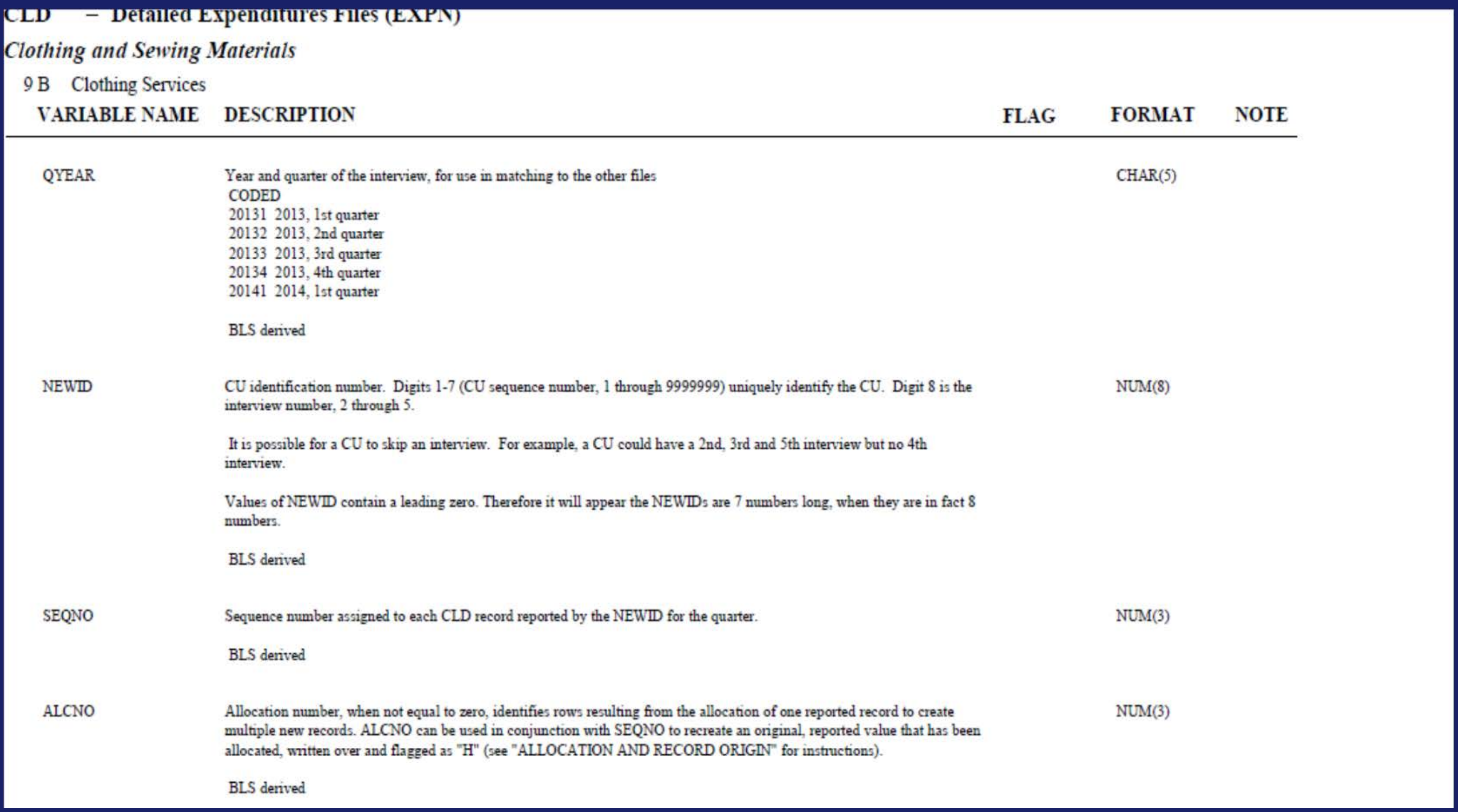
Percentage of records reported directly, imputed, allocated, or imputed and allocated by record type, Interview Survey, 2013 Q1-2014 Q1 – Cont.

Percent				
Record type	Directly reported	Strictly imputed	Strictly allocated	Allocated and imputed
UTP	99.1	0.9	0.0	0.0
VEQ	80.6	0.7	18.6	0.1
VLR	92.3	1.9	5.8	0.0
VOT	97.9	2.1	0.0	0.0
XPA	95.8	4.2	0.0	0.0
XPB	97.5	2.5	0.0	0.0

Documentation – Data Dictionaries

- Data Dictionaries
  - Interview Data Dictionary
  - Diary Data Dictionary
  - Access data dictionary database

Documentation – Data dictionaries



CLD – Detailed Expenditures Files (EXPN)

Clothing and Sewing Materials

9 B Clothing Services

VARIABLE NAME	DESCRIPTION	FLAG	FORMAT	NOTE
QYEAR	Year and quarter of the interview, for use in matching to the other files CODED 20131 2013, 1st quarter 20132 2013, 2nd quarter 20133 2013, 3rd quarter 20134 2013, 4th quarter 20141 2014, 1st quarter BLS derived		CHAR(5)	
NEWID	CU identification number. Digits 1-7 (CU sequence number, 1 through 9999999) uniquely identify the CU. Digit 8 is the interview number, 2 through 5. It is possible for a CU to skip an interview. For example, a CU could have a 2nd, 3rd and 5th interview but no 4th interview. Values of NEWID contain a leading zero. Therefore it will appear the NEWIDs are 7 numbers long, when they are in fact 8 numbers. BLS derived		NUM(8)	
SEQNO	Sequence number assigned to each CLD record reported by the NEWID for the quarter. BLS derived		NUM(3)	
ALCNO	Allocation number, when not equal to zero, identifies rows resulting from the allocation of one reported record to create multiple new records. ALCNO can be used in conjunction with SEQNO to recreate an original, reported value that has been allocated, written over and flagged as "H" (see "ALLOCATION AND RECORD ORIGIN" for instructions). BLS derived		NUM(3)	

Documentation – Data Dictionaries

- N(Yyyq) – indicates new variable introduced in yyq
- D(Yyyq) – indicates deleted variable, no longer available starting in yyq
- C(Yyyq) – indicates a change within the variable effective yyq

Documentation – Data Dictionaries

- Access database
  - Search by variable name
  - Search by RECTYPE
  - Search by keyword in description

Documentation – Sample Programs

- Sample programs
  - SAS :
    - Means and SE (Interview, Diary, and Integrated)
      - Goal is to replicate publication totals
    - Intrvw Sumvars (Means using summary expenditures in the interview FMLY file)
    - CE Macros (including documentation)
  - STATA:
    - Diary Means and SE
    - CE is working on increasing STATA knowledge and offer more microdata-related STATA products in the future.

Documentation - Other

- Other documentation/supporting files
  - Stub Parameter files
    - DStubYYYY.txt, IntStubYYYY.txt, IStubYYYY.txt,
    - Information on table aggregation, source selection, and UCC labels
  - UCC Title files
    - UCCIYY.txt, UCCDYY.txt with Interview and Diary files respectively
    - These files are a crosswalk with each UCC with its corresponding title

Documentation - Other

- Other documentation/supporting files
  - CAPI Vehicle codes
    - CAPIVEHI13.txt with the EXPN files
    - List of codes for the variable MAKE

File Structure

Interview Files

FMLI files

- One row = one household*
- NEWID is the unique identifier
- Each household surveyed with a completed interview will have a record
- Types of data: Income, demographics, geography (limited), weighting, summary of expenditures at the household level

Meet the Jones Family

- Family of 3:
  - Husband/Wife
  - 3 year old boy
- Renters
- Mr. Jones is self-employed
- Mrs. Jones works full time for an employer

BLS

Where does their information go?

- Mr. Jones was home for the interview and was identified as the reference person.

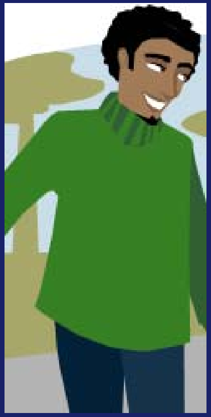
FMLI Files

NEWID	AGE_REF	AGE2	CUTENURE	FINCBTXM	REGION	BLS_URBN	TOTEXPPQ	TOTEXPCQ	FOODPQ	FOODCQ	FINLWT21
1928075	55	52	2	$16,408	2	1	$3,803	$11,153	$737	$1,623	15389
1928405	51	50	1	$77,800	2	1	$20,254	$0	$3,965	$0	19400
1928425	41	38	4	$41,500	4	1	$7,217	$0	$2,484	$0	12133
1928735	43	39	1	$68,000	1	1	$19,109	$6,541	$737	$368	15337
1928775	70	73	1	$80,000	2	1	$2,370	$5,896	$368	$737	15852

MEMI Files

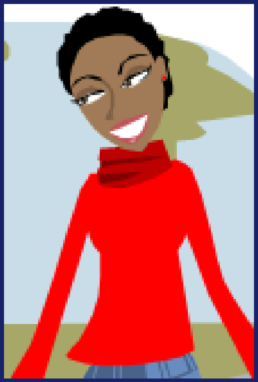
- One row = one member in a household
- NEWID and MEMBNO are the unique identifiers
- Types of data: Member income and employment information, member demographics, relationship to reference person

Where does their information go?



- Mr. Jones is a self-employed technician and had a loss of $7,100 in the past 12 months

- Mrs. Jones worked full time as a counselor and earned $48,600 in the past 12 months



MEMI Files

NEWID	MEMBNO	CU_CODE	AGE	EDUCA	MARITAL	EARNER	INCWEEKQ	OCCUPREV	SALARYXM	NONFARMM
1928425	1	1	41	43	1	1	52	103	.	- $7,100
1928425	2	3	3	.	5	.	.	.	.	.
1928425	3	2	38	44	1	1	52	201	$48,600	.

EXPN files

- Each file is called a “RECTYPE” (record type) and represents one section of the questionnaire
- The file variables follow closely with the way the questions are asked
- One row = IT DEPENDS!
- NEWID is *a* unique identifier, other identifiers are determined by the file

Where does their information go?

- The Jones' own two vehicles – so they will have two records in the OVB file

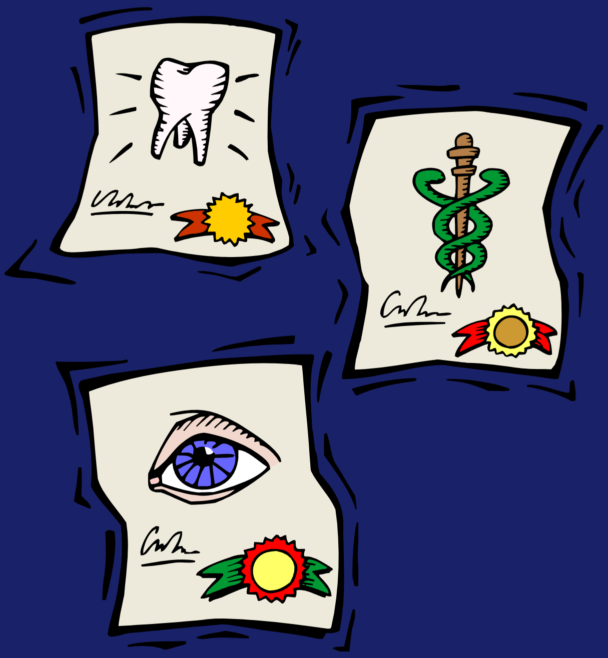
EXPN files: OVB file

QYEAR	NEWID	SEQNO	ALCNO	VEHICB	VEHICYB	VEHICYR	MAKE	QTRADEX	QVINTM3X	QADINT3X
20122	1928425	18	0	01	100	27	TOY	$26,000	$269	$66
20122	1928425	19	0	.	100	21	DOD	.	.	.

Additional OVB info:

- Information on purchase price is collected only when:
  - Vehicle purchase falls within the reference period
  - Vehicle is financed with payments remaining

Where does their information go?



- The Jones' hold 3 medical insurance policies: Dental, Vision, and a Fee for Service plan
- There will be a record for each policy in the IHB file

EXPN files: IHB file

QYEAR	NEWID	SEQNO	ALCNO	REC_ORIG	HHIPDLIB	HHICOVQ	HHICODE	HHIBCBS	HHIFEET	HHISPECT	QHI3MCX
20122	1928425	56	0	5	01	3	2	1	1	.	$236
20122	1928425	57	0	5	02	3	4	1	.	2	$25
20122	1928425	58	0	5	03	2	4	1	.	1	$100

- Some questions asked only under specific conditions, so many of the variables will have valid missing data.

EXPN files: SEQNO, ALCNO

- Common to all RECTYPES
- SEQNO assigned sequentially as each expenditure record is recorded into the database.
- ALCNO is assigned sequentially for each record that has been allocated from one expenditure during processing

Allocation example

“I bought a shirt and a jumper and spent $45”

EXPN files: SEQNO, ALCNO

QYEAR	NEWID	SEQNO	ALCNO	CLOTHYA	CLOTHMOA	CLOTHXA	CLOTHXA_
20132	1928425	43	0	270	02	43	.

QYEAR	NEWID	SEQNO	ALCNO	CLOTHYA	CLOTHMOA	CLOTHXA	CLOTHXA_
20132	1928425	43	0	270	02	.	H
20132	1928425	43	1	190	02	18	E
20132	1928425	43	2	150	02	25	E

MTBI files

- One Row = One expenditure by household
- NEWID, UCC, SEQNO, ALCONO, REFMO, REFYR, and UCCSEQ are the unique identifiers
- Types of data: Expenditures converted to monthly estimates categorized by a “Universal Classification Code” (UCC) with identified reference month and year.

BLS

Where does their information go?

- The Jones’ had 3 months of expenditures – regular bills and the like.

MTBI Files

NEWID	UCC	REF_MO	REF_YR	RECTYPE	EXPNAME	PUBFLAG	GIFT	COST	COST_
1928425	270310	03	2013	UTA	QADCAB1X	2	2	54	D
1928425	270310	04	2013	UTA	QADCAB2X	2	2	54	D
1928425	270310	05	2013	UTA	QADCAB3X	2	2	54	D
1928425	380333	03	2013	CLA	CLOTHXA	1	2	35	D
1928425	400310	04	2013	CLA	CLOTHXA	1	2	180	D
1928425	380313	05	2013	CLA	CLOTHXA	1	2	30	D
1928425	380333	04	2013	CLA	CLOTHXA	1	2	80	D

ITBI files

- One Row = One income source by household
- NEWID, UCC, REFMO, and REFYR are the unique identifiers
- Types of data: Income (imputed and reported values) and characteristics converted to monthly estimates (divided by 12) categorized by a “Universal Classification Code” (UCC) with identified reference month and year.

ITBI Files

NEWID	UCC	REF_MO	REF_YR	Value	Value_
1928425	900010	03	2013	-$591.667	
1928425	900010	04	2013	-$591.667	
1928425	900010	05	2013	-$591.667	
1928425	900000	03	2013	$4050	
1928425	900000	04	2013	$4050	
1928425	900000	05	2013	$4050	

Additional ITBI info:

- Based on the imputed versions of household level data
- Value_ only indicates whether value is topcoded

ITBI Files

- Multiple records per CU
- One record per income item per month/year
- Unique Records defined by NEWID, UCC, REFMO, REFYR
- Monthly amounts mapped from the corresponding CU-level imputed income mean values

NEWID	REFMO	REFYR	UCC	PUBFLAG	VALUE	VALUE_
1928075	03	2013	002030	2	243	
1928075	04	2013	002030	2	243	
1928075	05	2013	002030	2	243	
1928075	03	2013	800931	2	427	
1928075	04	2013	800931	2	427	
1928075	05	2013	800931	2	427	

Note: This ITBI sample shows ALL variables that are in the full ITBI file

ITII Files

- One Row = One impute for one income source by household
- NEWID, UCC, REFMO, REFYR, and IMPNUM are the unique identifiers
- Types of data: Five Imputes of income UCCs converted to monthly estimates categorized by a “Universal Classification Code” (UCC) with identified reference month and year.

ITII Files

NEWID	UCC	REF_MO	REF_YR	IMPNUM	Value	Value_
1928425	900010	05	2013	1	-$591.667	
1928425	900010	05	2013	2	-$591.667	
1928425	900010	05	2013	3	-$591.667	
1928425	900010	05	2013	4	-$591.667	
1928425	900010	05	2013	5	-$591.667	
1979322	900080	03	2013	1	$11833.17	
1979322	900080	03	2013	2	$7833.31	
1979322	900080	03	2013	3	$15221.44	
1979322	900080	03	2013	4	$9865.112	
1979322	900080	03	2013	5	$10372.877	

Diary Files

Diary Files

- Most files are similiar to the Interview files
- 4 quarters of weekly expenditures
- Slightly different variable names
- FMLD File summary variables – not as extensive as Interview file (no “Total Expenditures”)
- No detailed expenditure files

Diary EXPD files

- Similiar to MTBI files in Interview
- One Row = One expenditure by household
- Types of data: Expenditures as reported by a “Universal Classification Code” (UCC) with identified date of purchase (QREDATE)

So many files, where do I start??

Where to start

- Depends on the type of information and level of detail you need
  - Interview, Diary, or both?
    - Usually you can use Interview for most analysis (keeping in mind it will miss some smaller expenditures)
    - Detailed Food analysis, personal care expenses → Diary survey

Where to start, cont'd

- Which file in Interview?
  - FMLI will have demographics and summarized expenditures
  - MTBI will have detailed expenditures and can be joined with FMLI for additional demographics and weighting
  - EXPN files will have detailed expenditures and additional details about the expenditures (but you should familiarize yourself with the survey questions)

BLS