Imputing In-Kind Benefits in Survey Data

Nikolas Mittag

September 23, 2021

Introduction

- Accurate information on government benefits is crucial for policy
- However, **many programs are poorly reported** in household surveys
- Other programs are **missing from key surveys** altogether
- In this presentation, I will
    1. review **three strategies to impute** program receipt and amounts
    2. discuss key implications for **imputing in-kind benefits**, such as WIC, NSLP and LIHEAP, to measure consumption and poverty
- Similar strategies work for missing and misreported programs, though having reported receipt makes the latter simpler

Three Imputation Strategies: Basic Idea

Will discuss 3 (stylized) imputation strategies:

- **Predicting eligibility**:
  1. Predict eligibility based on survey information and program rules
  2. Assign receipt to (some of) those predicted to be eligible
- **Reported receipt**:
  1. Estimate a model of the probability to report receipt
  2. Assign receipt to additional units based on predicted probability to report receipt until survey estimates match aggregate receipt
- **Prediction equations**
  1. Estimate a model of receipt in a different data source
  2. Predict receipt in survey data based on estimated equation

Predicting Eligibility: Discussion

Key advantages

- Eligibility rules are often simple
- Restricts receipt to those (predicted) eligible

Key Problems

- Need accurate information on eligibility criteria
- Predicting eligibility is often very noisy (Scherpf, Newman, Prell, 2014)
- With incomplete take-up, need to decide which eligible units receive program

Reported Receipt: Discussion

Key advantages

- Improves underreporting (which is often severe)
- Preserves correlations of reported receipt

Key Problems

- Not possible if program is missing entirely from survey
- Cannot improve bias when misreporting is related to covariates

Prediction Equations: Discussion

Key advantages

- Can correct both levels and correlates of program receipt
- Can be consistent and theoretically optimal

Key Problems

- Requires additional data source with accurate information
- Predictors need to be comparable accross data sources
- Reproducing correlations hinges on availability of good predictors

Three Imputation Strategies: Evaluation

Mittag (2019) evaluates 3 ways to impute SNAP, which loosely correspond to the 3 types of strategies:

- **TRIM** assigns receipt to units predicted to be eligible based on CPS responses such that the recipient population matches program records
- Scholz, Moffitt and Cowan (2009, **SMC**) proposed assigning additional receipt to households with a high probability of receipt according to Probit models of reported receipt
- Mittag (2019) estimates the **conditional distribution** of receipt given reported receipt and covariates from survey data linked to administrative records and uses it to predict receipt in survey data

Three Imputation Strategies: Comparison

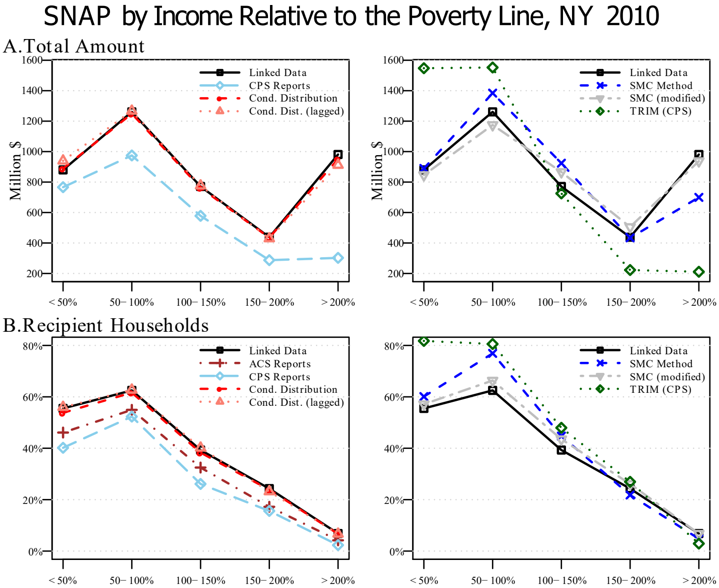
Comparison of Key Features of the Evaluated Methods

	(1) **Model**	(2) **Required Data**	(3) **Key Assumptions**
TRIM	Eligibility Criteria Matching Moments	Reported eligibility criteria Info on recipient population	Accurate eligibility information “Selection on observed aggregates”
SMC	Probit (receipt) OLS (amounts)	Reported receipt	Never correct Random misreporting
Conditional Distribution	Conditional normal distribution	Accurate receipt in auxiliary data	Model and predictors comparable in aux. and main data

Three Imputation Strategies: Summary of key findings

- All methods drastically **improve levels** of program receipt (partly by construction)
- The **conditional distribution** method accurately reproduces uni- and multivariate statistics as well as the geographic distribution of program spending
- Carefully **extrapolating** from linked data accross time and geography appears promising
- The (modified) **SMC method** improves estimates, but less so especially for multivariate statistics
- **TRIM** improves simple statistics and the geographic distribution of spending, but sharply overcorrects below the poverty line

Three Imputation Strategies: Income Gradient



Three Imputation Strategies: Geographic Distribution

Extrapolating SNAP Statistics to the Entire U.S., 2010

	(1) Reports	(2) Conditional Distribution	(3)	(4) SMC Method modified	(5) TRIM
Data	CPS US	ACS US	ACS US	ACS US	CPS US
Parameters	–	NY	NY, adj.	by state	–
Mean Abs. Deviation of Total \$ Received (in Million \$) to Admin. Totals . . .					
by state	497.2	110.4	3.0	0.0	93.9
for large MSAs	210.0	54.2	21.8	125.5	55.6
for county groups	–	10.7	8.6	12.0	–

Imputing In-Kind Benefits to Measure Consumption

- To measure consumption at the consumer unit level, need to **add information on in-kind benefits** such as the national school lunch program, WIC, or LIHEAP to the CE
- Contrary to the study above, these programs are **missing from the survey entirely** and **linked administrative data is not available**
- The specific purpose of the imputation emphasizes specific aspects:
    - Need to impute **multiple programs**
    - **Correlation** with other consumption (and other programs) particularly important
    - Less important to reproduce correlation with other predictors for multivariate models?

Thoughts on Potential Imputation Strategies

- Imputation **based on eligibility** is a good start (Garner et al. 2015), but faces problems:
  - accurate information on eligibility criteria?,
  - predicting joint take-up of multiple programs
  - predicting how take-up varies with other consumption
- Imputation **based on survey information**
  - SMC method does not work for programs missing from the survey
  - However, could use a model of receipt estimated from a different survey (Garner and Hokayem 2011, 2012)
  - Could also impute from other surveys via matching (Short and Renwick) or via a conditional distribution
- The CE has not been linked (yet?), but can use information from similar (linked) surveys and unlinked administrative data

Combining Strategies I

It seems useful to **combine the elements from each strategy** that are likely to work well in the case at hand, for example:

1. Constrain imputation to those **eligible** whenever reliable information on (in)eligibility is available (e.g. presence of children)
2. Make best use of available **survey data**:
    - Estimate **prediction equations** from reported receipt in the most accurate survey available
    - Use information from administrative data to adjust for underreporting at the lowest feasible geographic/demographic level (CBO 2018)
    - Use surveys with extensive information on programs (e.g. the SIPP) to validate imputations

Combining Strategies II

3. Make use of **(linked) administrative data** whenever feasible
    - Use prediction equations from linked data whenever possible (Fox, Rothbaum, Shantz, 2020)
    - If subsamples, some years or other surveys can be linked, can
        - examine extrapolation
        - use them to validate methods, e.g. that imputations reproduce key correlations
        - acquire additional information (receipt by income, ethnicity, correlations, etc.) to use as constraints
    - Can also use aggregate statistics from unlinked administrative data as constraints

Three practical issues

1. Benefits need to be **imputed stochastically** rather than to the most likely recipients to avoid overimputing among the poorest
2. To estimate distribution of consumption, imputation needs to **capture dependence** in program receipt **and correlation** with other consumption. Potential solutions:
   - Impute sum of program benefits (requires availability in one source)
   - Condition on programs sequentially
   - Check whether imputations reproduce relevant correlations from other data
3. **Combine** all available sources of information: other surveys, other linked data, information from aggregate statistics or (un)linked administrative records, . . .