# In an era of declining response rates, a note on the importance of estimation procedures

Martin Klein, Joanna Fane Lineback, and Joseph L. Schafer

U.S. Census Bureau

June 10-12, 2019

International Total Survey Error Workshop

The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. CBDRB-FY19-CED002-B0012



# Outline

- 1. Motivation
- 2. Missing Data in the Monthly Wholesale Trade Survey
- 3. Simulation Study
- 4. Discussion
- Alternative Estimation Procedures



#### **Refusal Rate Trends**



census.gov

Bureau

3

# 1. Motivation

- Increasing nonresponse in U.S. federal surveys
- Focus on response rates and the potential for nonresponse bias
- Less attention to estimation procedures, which may have been in place for a long time
- Estimators should be working well without missing data
- Example: Ongoing, multi-year research effort to evaluate missing data and estimation procedures (and offer alternatives) in the Monthly Wholesale Trade Survey (MWTS)



# 2. Missing Data in the MWTS

- Monthly survey of employer wholesale businesses
- Businesses report their sales and inventories data for the month just ending
- Provides estimates and associated standard errors of:
  - monthly sales
  - end-of-month inventories
  - month-to month percent change in sales
  - month-to-month percent change in inventories
  - inventories to sales ratio



# Current MWTS Methodology

- Stratified random sample, with strata defined by industry class (NAICS) and measure of size (MOS)
- Wholesale units are first selected for the Annual Wholesale Trade Survey (AWTS), subsample of AWTS is selected for the MWTS
- Total MWTS sample size approx. 4,200 units (about 40% selected with certainty)
- Selected companies are followed for 5+ years
- Missing values filled in by a simple ratio imputation method  $(\hat{y}_i = \hat{R}x_i)$
- Total sales and inventories estimated by Horvitz-Thompson, variances by random groups
- Estimates are benchmarked and seasonally adjusted



### MWTS Item Missingness Rates by Month



## Previous Research

- Earlier work suggested the potential for nonresponse bias
- Concerns with current imputation method
  - implicitly assumes linear model with intercept 0
  - only predictor is value from previous month
  - in some imputation cells, estimated ratios based on very few cases
  - relationship between sales and inventories may be distorted
  - doesn't make use of other information (e.g., from economic census)
  - no attempt to assess missing-value uncertainty or include it in variance estimates
- Investigated multiple imputation (MI) of key items in MWTS under an explicit multivariate model (see Lineback and Schafer, 2013) that
  - preserves distributional shapes
  - preserves relationships within months
  - preserves trends across months
  - makes efficient use of other covariates (e.g., frame variables)
- Observed large discrepancies between methods; new imputation methods produced smaller point estimates and larger standard errors than existing methods



# 3. Simulation Study

- Goal: To evaluate missing data procedures in hypothetical population and sample that resembles the MWTS
- Design:
  - Construct plausible population
  - Draw samples
  - Impose patterns of missing values
  - Impute missing values
  - Compute estimates and standard errors



# Data for this Study

- 25 months of MWTS data from December 2008 to December 2010
- Key predictors from the 2007 Economic Census (EC)
  - Sales
  - Payroll
  - Employment
  - Industry
- Simplifications:
  - Ignored "births" and "deaths" (i.e., fixed units, constant weights)
  - Assumed each unit operates in only one industry group
  - Ignored benchmarking and seasonal adjustment



# Simulation

- Constructed an artificial population (lots of work!) using 25 months MWTS data and 2007 EC data
- For validation, compared simulated population totals to published MWTS estimates
- Selected 1,000 random samples using design that mimics actual MWTS stratified random sample design
- Assigned random groups (for variance estimation later)
- Imposed missing values to mimic actual MWTS missing data pattern



### Simulated Population vs. Sample





U.S. Department of Commerce Economics and Statistics Administration U.S. CENSUS BUREAU *census.gov* 

12

# Estimates and SEs

- Computed Horvitz-Thompson (HT) estimates and random group SEs for
  - total sales
  - total inventories
  - percent change in sales
  - percent change in inventories
- Estimates computed for
  - complete sample (i.e., without imposing missing data)
  - each imputation method
    - current imputation method (imputed values treated as actual data)
    - MI method



# Complete Sample Analysis

- Expect:
  - estimates of population quantities (point estimates) would be approximately unbiased for the population true values
  - variance estimates (squared standard errors for the point estimates) would accurately reflect the true variability in the point estimates
  - confidence intervals would have actual coverage close to nominal levels (e.g., point estimate plus or minus two SEs would cover the population true values about 95% of the time)



# Complete Sample: Total Sales



Var. of point estimates

- National level ٠
- Inferential procedures assume sample ٠ estimates approximately normally distributed
- Not normally distributed ٠
- Highly skewed, heavy tailed ٠
- Industry level
- If both average point estimate and population value lie on 45 degree line, then unbiased
- Point estimates unbiased
- Variance estimates unbiased ٠



**U.S. Department of Commerce** Economics and Statistics Administration **U.S. CENSUS BUREAU** census.gov

Population value

## Complete Sample: Total Sales



- Industry level
- Combines point and variance estimates
- There is undercoverage; the confidence interval does not cover the "true" population value 95% of the time
- Nonlinear functions of the totals might not be well-behaved



# Complete Sample: Percent Change in Sales



- National level
- Still not normally distributed
- Less skewness, but very heavy tails, which will cause a problem for variance estimation

- Industry level
- Point estimates biased
- Variance estimates biased, not lying close to the 45 degree line
- Not show here, but coverage approximately 73%



# Current Imputation Method: Total Sales



- National level
- Not normally distributed
- Highly skewed, heavy tailed

- Industry level
- Point estimates slight upward bias
- Variance estimates unbiased
- Coverage still poor



# Current Imputation Method: Percent Change in Sales





# MI Method: Total Sales



- National level
- Not normally distributed
- Highly skewed, heavy tailed
- Point estimates biased
- Coverage goes up to about 81%



# Findings

- Theory assumes that without missing data estimators working well
- This assumption seems to be violated:
  - Highly skewed (i.e., non-normal) and/or heavy-tailed point estimates
  - In some cases, biased point and/or variance estimates
  - Coverage much, much less than 95%
- No missing data technique will mitigate problems with estimators themselves (i.e., HT estimator and method of random groups)



# 4. Discussion

- Sampling and estimation procedures have stayed relatively the same for decades
- Advances in statistics and computing
- Access to other data sources and prior months' data
- Synthetic population allows us to investigate different sampling, missing data, and estimation procedures
- New estimation procedures seem promising



#### Questions:

- 1. Are there gains to be made with improvements to sampling, missing data, and estimation procedures?
- 2. Survey candidates?



# Alternative Estimation Procedures

- Improve performance of complete-data estimators
- Take advantage of advancements in statistics and computing
- More rigorous simulation study to compare performance of methods under realistic conditions
- Updated the artificial population using
  - 52 months of MWTS data (December 2008-December 2013)
  - 2017 Economic Census of Wholesalers data
  - the BSR07 sampling frame
- Constructing realistic models using Bayesian methods for surveys and a class of functional models



#### Simulated Population vs. Sample



(d) Percent change in inventories

United States<sup>\*</sup> Bureau

#### Example

- Stratified sample of n = 209 Apparel Wholesalers from our simulated population of N = 14,747, with patterns of missingness imposed by hot deck
- Response shown is the tenth root of (monthly sales / MOS)
- A few outlying series (those with zeros) were removed





#### Fitted long-term trend



Fitted periodic trend





#### One "observed" series from simulated sample



Same series, with fitted curve and 100 imputations



