Finding an Estimator that Minimizes Revisions in a Monthly Indicator Survey

Nicole Czaplicki, Yarissa Gonzalez, Laura Bechtel
U.S. Census Bureau

Disclaimer: Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau.
Outline

▪ Motivation
▪ Research purpose
▪ Background on MARTS and MRTS
▪ Current estimator
▪ Estimators considered
▪ Results
▪ Conclusions
Monthly Retail Trade at Census

- Two surveys: three releases
  - Advance Monthly Retail Trade and Food Services Survey (MARTS)
    - Advance release of sales
  - Monthly Retail Trade and Food Services Survey (MRTS)
    - Preliminary release of sales
    - “Final” release of sales

- MRTS data are also revised annually with changes from benchmarking to the annual survey.
  - Includes remaining late reported data
  - Data corrections
Retail Trade Indicator Release Example for January

February
Advanced January Estimates from MRTS

March
Preliminary January Estimates from MRTS

April
Final January Estimates from MRTS

Goal: minimize this revision
Motivation

- MARTS is an economic indicators
- Total retail trade and food services sales estimates are inputs to GDP
- Revisions to the estimates, particularly the seasonally adjusted month-to-month change, are highly scrutinized.
Research Purpose

Identify an estimator that minimizes revision error between the MARTS estimate and MRTS preliminary estimate

- Using only respondent data
- Including analyst imputes, which are a subset of influential nonrespondents imputed based on historical company performance (current method)
Research Questions

1. Is there an alternative estimator that we can use that would reduce revision error?
2. Are the analyst imputed values reducing revision error?
Challenges to minimizing revisions

1. Low response rates for MARTS
2. Differences in methodology between MARTS and MRTS
MARTS Unit Response Rates
<table>
<thead>
<tr>
<th>Sample size</th>
<th>5,000 companies</th>
<th>12,000 companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample frame</td>
<td>MRTS sample</td>
<td>Annual Retail Trade Survey sample</td>
</tr>
<tr>
<td>Sample design</td>
<td>Stratified PPS -WOR (subsample of MRTS)</td>
<td>Stratified SRS-WOR</td>
</tr>
<tr>
<td>Sample redesign cycle</td>
<td>Approximately every 2.5 years</td>
<td>Approximately every 5 years</td>
</tr>
<tr>
<td>Time to respond</td>
<td>Approximately 7 business days</td>
<td>Approximately 5 weeks</td>
</tr>
<tr>
<td>Imputation</td>
<td>Analyst impute for selected companies</td>
<td>Analyst imputes retained, ratio impute for remaining nonrespondents and edit-failing items</td>
</tr>
<tr>
<td>Estimation</td>
<td>Link relative estimator</td>
<td>Horvitz-Thompson estimator</td>
</tr>
<tr>
<td>Tabulation industries</td>
<td>30</td>
<td>83</td>
</tr>
<tr>
<td>Sample size</td>
<td>5,000 companies</td>
<td>12,000 companies</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Sample frame</td>
<td>MRTS sample</td>
<td>Annual Retail Trade Survey sample</td>
</tr>
<tr>
<td>Sample design</td>
<td>Stratified PPS -WOR (subsample of MRTS)</td>
<td>Stratified SRS-WOR</td>
</tr>
<tr>
<td>Sample redesign cycle</td>
<td>Approximately every 2.5 years</td>
<td>Approximately every 5 years</td>
</tr>
<tr>
<td>Time to respond</td>
<td>Approximately 7 business days</td>
<td>Approximately 5 weeks</td>
</tr>
<tr>
<td>Imputation</td>
<td>Analyst impute for selected companies</td>
<td>Analyst imputes retained, ratio impute for remaining nonrespondents and edit-failing items</td>
</tr>
<tr>
<td>Estimation</td>
<td>Link relative estimator</td>
<td>Horvitz-Thompson estimator</td>
</tr>
<tr>
<td>Tabulation industries</td>
<td>30</td>
<td>83</td>
</tr>
</tbody>
</table>
Link Relative Estimator

\[ \hat{Y}_{LR,t} = B_{t-1} \left[ \frac{\sum_{i \in C} w_i y_{t,i}}{\sum_{i \in C} w_i y_{t-1,i}} \right] \]

- \( \hat{Y}_{LR,t} \) = link – relative estimator of a total for period \( t \)
- \( B_{t-1} \) = benchmark value associated with period \( t - 1 \)
  - MRTS Preliminary total
- \( C \) = set of units with usable data for both periods \( t \) and \( t - 1 \)
- \( w_i \) = sample weight for unit \( i \)
- \[ \frac{\sum_{i \in C} w_i y_{t,i}}{\sum_{i \in C} w_i y_{t-1,i}} = C \hat{Y}_t \]
  - \( \hat{Y}_t \) = the screened sum ratio (link relative)
- Introduced by Madow and Madow (1978)
Link Relative Assumption Validation

- Key assumption $E[y_{t,i} | t - 1] = \beta_t y_{t-1,i}$
- Tested by estimating:
  \[ y_{i,t} = \beta_t y_{i,t-1} + e_{i,t}, \quad e_{i,t} \sim (0, \sigma_t^2) \]
Box plots of $R^2$ statistics by month for the model $y_{i,t} = \beta_t y_{i,t-1} + e_{i,t}$ for tabulation NAICS 441100, new and used car dealers. Source MARTS and MRTS: May 2010 – May 2016.
Tabulation NAICS 451XX0

- General Bookstores
- Specialty, Religious, and Used Bookstores
- School and College Bookstores
- Musical Instrument and Supplies Stores
- News Dealers and Newsstands
- Prerecorded Tape, Compact Disc, and Record Stores
Box plots of $R^2$ statistics by month for the model $y_{i,t} = \beta_t y_{i,t-1} + e_{i,t}$ for tabulation NAICS 451XX0. Source MARTS and MRTS: May 2010 – May 2016.
Modified Link Relative Estimator

\[ \hat{Y}_{MLR,t} = \left[ B_{t-1} - \sum_{i \in \hat{S}} \tilde{w}_i y_{t-1,i} \right] \left[ \frac{\sum_{i \in \hat{C}} w_i y_{t,i}}{\sum_{i \in \hat{C}} w_i y_{t-1,i}} \right] + \sum_{i \in S} \tilde{w}_i y_{t,i} \]

- \( \hat{C} \) = set of units with usable data for both periods \( t \) and \( t - 1 \) that are included in the calculation of the screened-sum ratio
- \( S \) = set of units with usable data for both periods \( t \) and \( t - 1 \) that are excluded in the calculation of the screened-sum ratio
  - Referred to as “estimate separately” units
  - Group 1: “high impact” units currently excluded from MARTS estimation
  - Group 2: Group 1 units plus the analyst imputed units
- \( \tilde{w}_i \) = final weight from benchmark survey
- Developed at Census for the Manufacturers’ Shipments, Inventories, and Orders Survey (M3)
Weighting Class Estimator

\[ \hat{Y}_{WC,t} = \sum_{p=1}^{P} \sum_{i \in r_p} w_i y_{t,i} \frac{\sum_{i \in s_p} w_i}{\sum_{i \in r_p} w_i} \]

- \( i \in r_p \) indicates unit \( i \) is in the respondent set for weighting cell \( p \)
- \( i \in s_p \) indicates unit \( i \) is in the sample for weighting cell \( p \)
- Estimate inverse-probability of selection as \( \frac{\hat{N}_{sp}}{\hat{N}_{rp}} = \frac{\sum_{i \in s_p} w_i}{\sum_{i \in r_p} w_i} \)
- Weighting cells will be defined by certainty status
Weighting Class Estimator

\[ \hat{Y}_{WC,t} = \sum_{p=1}^{P} \sum_{i \in r_p} w_i y_{t,i} \frac{n_p}{m_p} \]

- \( n_p \) = sample size in cell \( p \)
- \( m_p \) = number of respondents in cell \( p \)

Ratio Estimator

\[ \hat{Y}_{R,NAICS_{j,t}} = \sum_{p=1}^{P} \sum_{i \in r_p} w_i y_i \frac{\sum_{i \in s_p} w_i x_i}{\sum_{i \in r_p} w_i x_i} \]

- \( x_i = \) value of auxiliary variable for unit \( i \)
  - MRTS prior month’s sales
- For the ratio estimator, we assume
  - \( y_i = \beta x_i + \varepsilon_i \) where \( \varepsilon_i \sim (0, x_i \sigma^2) \)
## Considered Estimators

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Stratified or non-stratified</th>
<th>Include or exclude weights in adjustment factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link Relative</td>
<td>Both</td>
<td>n/a</td>
</tr>
<tr>
<td>Modified Link Relative group 1</td>
<td>Both</td>
<td>n/a</td>
</tr>
<tr>
<td>Modified Link Relative group 2</td>
<td>Both</td>
<td>n/a</td>
</tr>
<tr>
<td>Weighting Class</td>
<td>Both</td>
<td>Both</td>
</tr>
<tr>
<td>Ratio</td>
<td>Both</td>
<td>Include weights</td>
</tr>
</tbody>
</table>
Evaluation Criteria

- Regression: Ideally MARTS perfectly predicts MRTS so $\beta_t$ should be close to 1.0

$$\hat{Y}_{t,j,\text{MRTS prelim}} = \beta_t \hat{Y}_{t,j,\text{MARTS}} + \epsilon_t, \; \epsilon_t \sim (0, \sigma_j^2)$$

- Mean absolute revision error for month-to-month percent change
Distribution of slope parameters by estimator from the 30 tabulation NAICS regression models. Source: MARTS May 2010-May2016
Distribution of slope parameters for the link relative and modified link relative estimators from the 30 tabulation NAICS regression models Source: MARTS May 2010-May2016
<table>
<thead>
<tr>
<th>Estimator</th>
<th>Current analyst imputes</th>
<th>No analyst imputes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link Relative</td>
<td>0.309</td>
<td>0.374</td>
</tr>
<tr>
<td>Link Relative: Stratified</td>
<td>0.277</td>
<td>0.354</td>
</tr>
<tr>
<td>Modified Link Relative: ES group 1</td>
<td>0.411</td>
<td>0.462</td>
</tr>
<tr>
<td>Modified Link Relative: ES group 2</td>
<td>0.396</td>
<td>n/a</td>
</tr>
<tr>
<td>Modified Link Relative Stratified: ES group 1</td>
<td>0.377</td>
<td>0.435</td>
</tr>
<tr>
<td>Modified Link Relative Stratified: ES group 2</td>
<td>0.363</td>
<td>n/a</td>
</tr>
<tr>
<td>Ratio</td>
<td>6.399</td>
<td>6.146</td>
</tr>
<tr>
<td>Ratio: Stratified</td>
<td>66.203</td>
<td>5.995</td>
</tr>
<tr>
<td>Weighting class: Weighted</td>
<td>75.854</td>
<td>45.048</td>
</tr>
<tr>
<td>Weighting class: Unweighted</td>
<td>24.611</td>
<td>5.960</td>
</tr>
<tr>
<td>Weighting class: Stratified weighted</td>
<td>29.388</td>
<td>12.205</td>
</tr>
<tr>
<td>Weighting class: Stratified unweighted</td>
<td>17.825</td>
<td>3.906</td>
</tr>
</tbody>
</table>
Distribution of mean absolute revision error for month to month changes at the tabulation industry level for the link relative and modified link relative estimators. Source: MARTS May 2010-May 2016
Research Questions

1. Is there an alternative estimator that we can use that would reduce revision error?
2. Are the analyst imputed values reducing revision error?
Absolute revisions to month-to-month changes using the link relative estimator without analyst imputes – absolute revisions to month-to-month changes using the link relative estimator with analyst imputes for MARTS Source: MARTS May 2010-May 2016
Absolute revisions to month-to-month changes using the link relative estimator without analyst imputes – absolute revisions to month-to-month changes using the link relative estimator with analyst imputes for MARTS. Number of analyst imputes plotted on the second y-axis. Source: MARTS May 2010-May2016
Conclusions

- Provided strong evidence that the currently used version of link relative estimator is effective at reducing revision error
  - Modified link relative estimator and stratified link relative estimator had inconsistent performances
  - Nonresponse adjustment estimators did not reduce revision error
- Found evidence that the current imputation procedure reduces revision error
  - We looked at aggregate results – did not look at impact of individually imputed values
  - It is possible that the improvements may be due to a select few observations
Contact Information

- Nicole.Czaplicki@census.gov
- Yarissa.Gonzalez@census.gov
- Laura.Bechtel@census.gov
Acknowledgements

- Thank you to MARTS estimation research team
  - Adam Smeltz
  - Artaud Hourticolon Retzler
  - Andromache Mason