Detecting & Treating Verified Influential Values in a Monthly Retail Trade Survey

Mary H. Mulry
Broderick E. Oliver
Stephen J. Kaputa
U.S. Census Bureau

WSS JOS ICES-IV Special Issue Seminar, February 4, 2015

Definition: **Influential Value**

An observation whose reported value is ***correct*** but whose ***weighted*** contribution has an excessive effect on estimated total or period-to-period change

- ‘Verified’ in title emphasizes not an error
- Treat to avoid bias in estimates

U.S. Monthly Retail Trade Survey

- Monthly survey ≈ 12,000 retail businesses with paid employees
- Collects sales, e-commerce and end-of-year inventory data

Current approach

- The returns for a monthly business survey have been edited and reviewed for:
  - Consistency of individual returns (micro-edits)
  - Outliers using the Hidiroglou-Berthelot Edit (macro-review)
- Analysts review “outliers” identified by the Hidiroglou-Berthelot edit
  - Legitimate/validated values are retained for potential use in the tabulations
- Tabulations are compared to prior month levels
  - Large shift (low or high) observed
  - Analysts must determine reason for shift
  - How should these **influential value(s)** be treated?

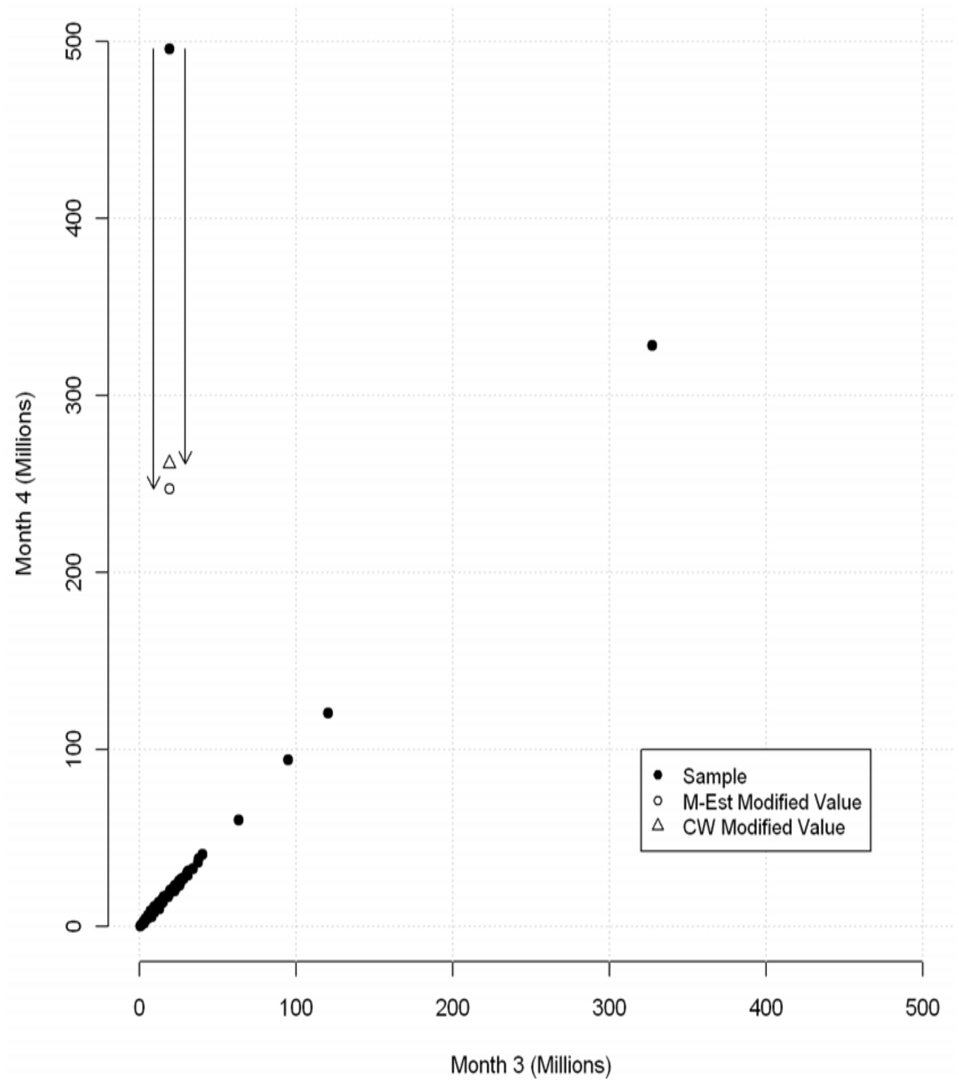
Objective of Research

To investigate and recommend an automated statistical procedure for detecting and treating influential values.

- Investigated
  - **M-estimation** (Beaumont & Alavi, 2004)
  - **Clark Winsorization** (Kokic & Bell, 1994; Clark, 1995; Chambers et al., 2000)
- Simulation Study
  - Data modeled after the Monthly Retail Trade Survey (MRTS, U.S. Census Bureau)

Example of influential value & adjustments

(weighted, simulated data)



Methods

Clark Winsorization

- Modifies the value of the influential observation(s)
- One-sided
  - Detects/treats high infl. values
- Fits robust regression (LMS); uses residuals in detection
- L-estimator method finds adjusted value by minimizing MSE
- Detected values determined by current data set
  - No predetermined parameters

M-Estimation

- Modifies the value or the weight of the influential observation(s)
- One or two-sided
  - Focus on one-sided here
- Fits robust regression; uses residuals in detection
- Iterative method finds adjusted value by minimizing MSE
  - Set max iterations at 5
- Detected values determined by current data set with parameters specified by user

Implementation challenges

- Large number of industries in monthly economic surveys
  - Some industries volatile, others not
- Methodology must fit into tight schedule for production of monthly estimates
  - Minimize false detections to control staff time for checking
- Estimation is at industry level, but survey design is at industry by size category

Model misspecification for MRTS data

- Stratified SRS-WOR Design
  - Strata defined by primary industry and unit size

$$y_{hi} = \beta_h x_{hi} + \xi_{hi},\ \xi_{hi} \sim (0, \sigma_h^2)$$

Stratum *h* and unit *i*

Stratum *h* within industry

- Data are tabulated at industry level & models underlying methods at industry level

$$y_i = \beta x_i + \xi_i,\ \xi_i \sim (0, \sigma^2) \text{ or } \xi_i \sim (0, x_i \sigma^2)$$

Unit *i*

Industry

Simulation design for MRTS industry

- Generate 20 months of population
  - AR(1) stationary time series
    - Avoids confounding with trends & seasonality
- Induce influential value in Month 4
- Draw samples until have 200 with induced influential value in Month 4

Simulation Procedure

Apply the M-estimation and Clark Winsorization algorithms to each sample

- Data from prior month serves as auxiliary variables
- Any changes in a given month carry over to the next month
- Estimate the industry sales total over all replicates
- Assess the statistical properties of each treatment with respect to total and to month-to-month change

Industries simulated

- Industry 1, volatile
  - $46.1 billion sales; 1,161 sample size
  - 10,742 samples
- Industry 2, stable
  - $2.5 billion sales; 147 sample size
  - 11,931 samples

Performance measures

- Type I Error Rate (false positive)
- Type II Error Rate (false negative)
- Relative Bias
- Relative Root Mean Square Error (RRMSE)
- Nonconvergence Rate

M-estimation algorithm parameters: settings based on Type I & II error rates

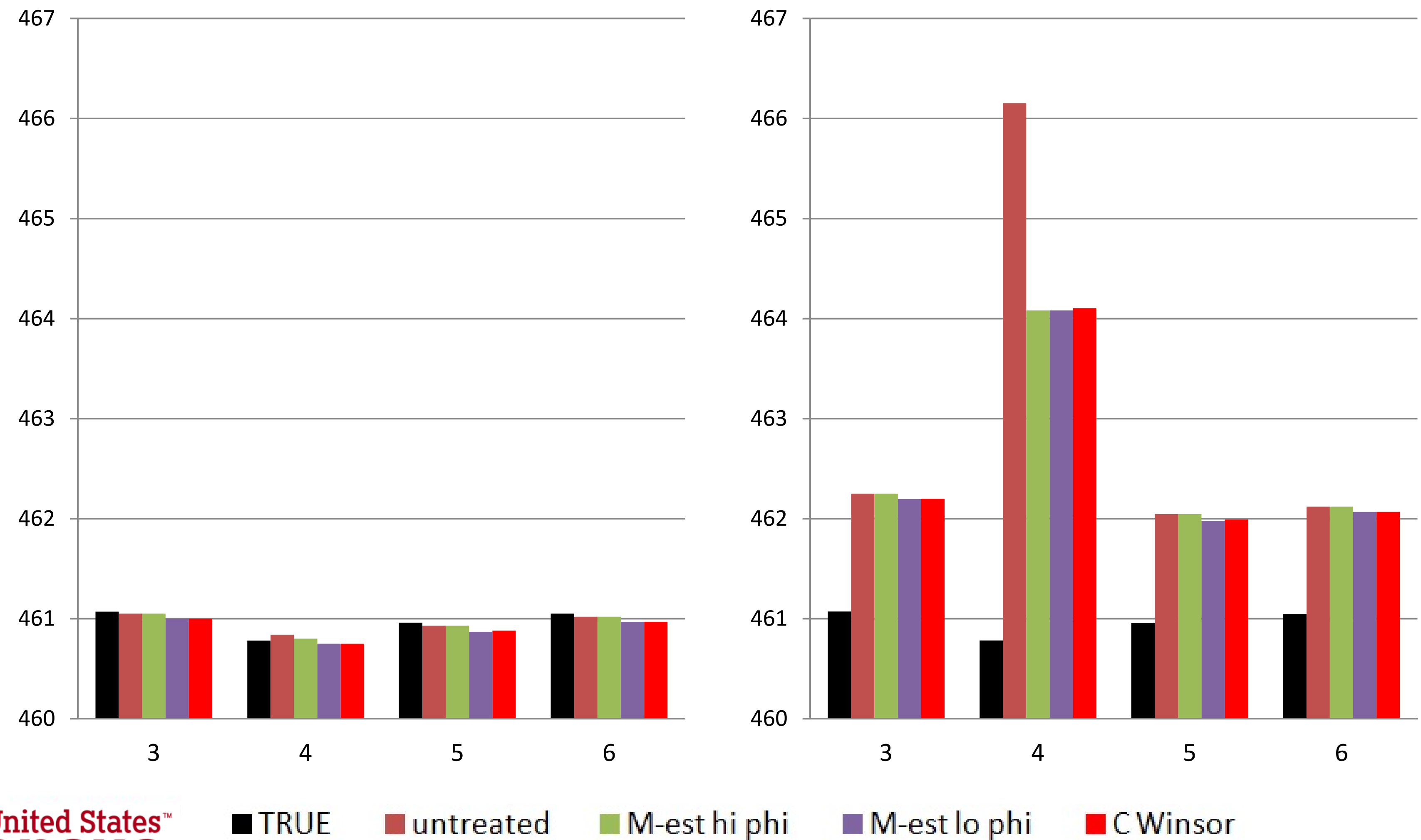
Parameter	Parameter Function	Values
v_i	Model error underlying regression estimator	$= 1$ or x_i
ψ	Reduces influence of units with large weighted residual	Huber I or Huber II
φ	Tuning constant • User provides initial value and program calculates optimal value	$*$ Low initial φ $*$ High initial φ

Estimated Sales Total by Month

(100 millions)

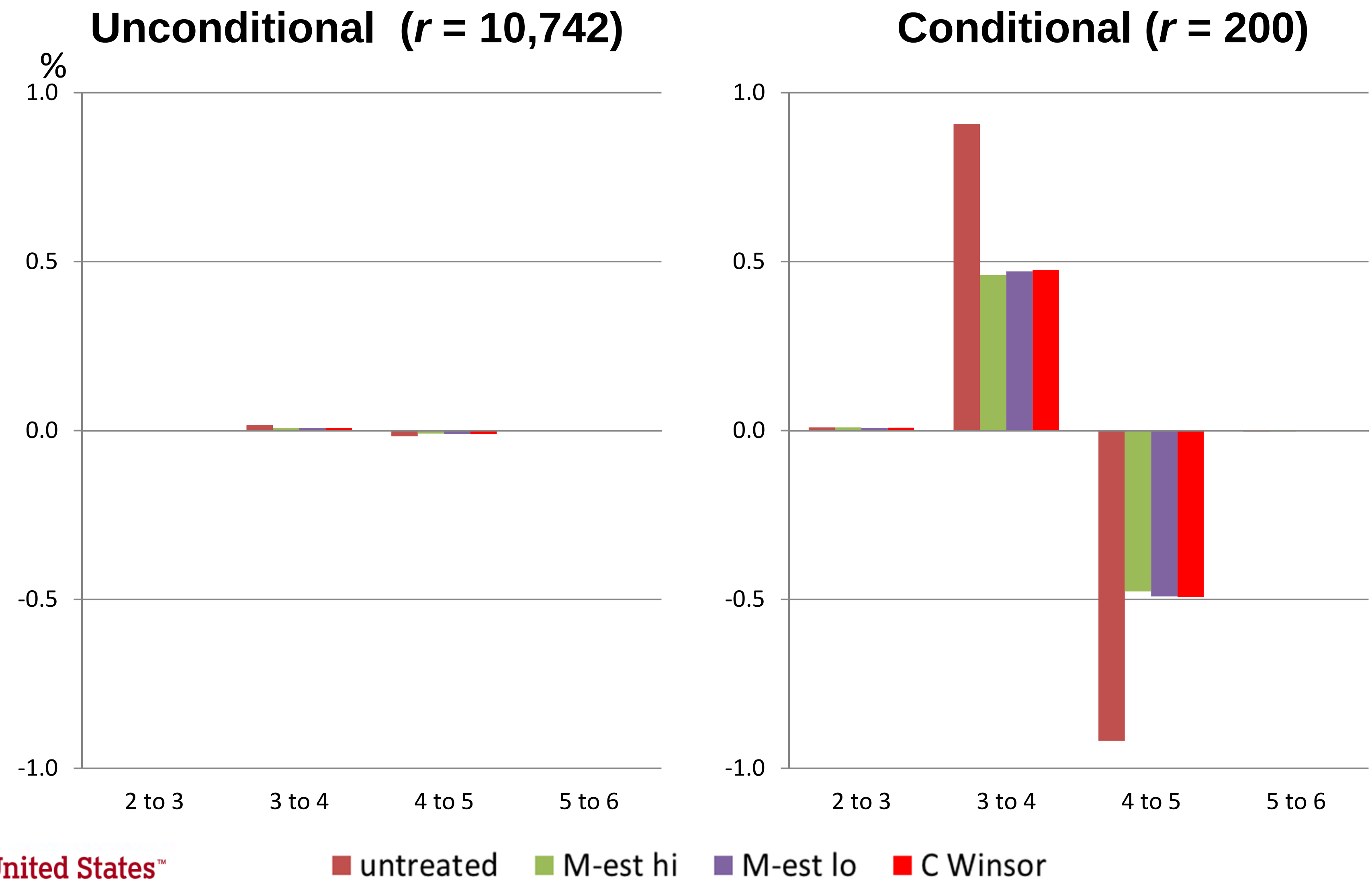
Industry 1, Huber II

Unconditional (*r* = 10,742) **Conditional (*r* = 200)**



Relative Bias for Month-to-Month Change

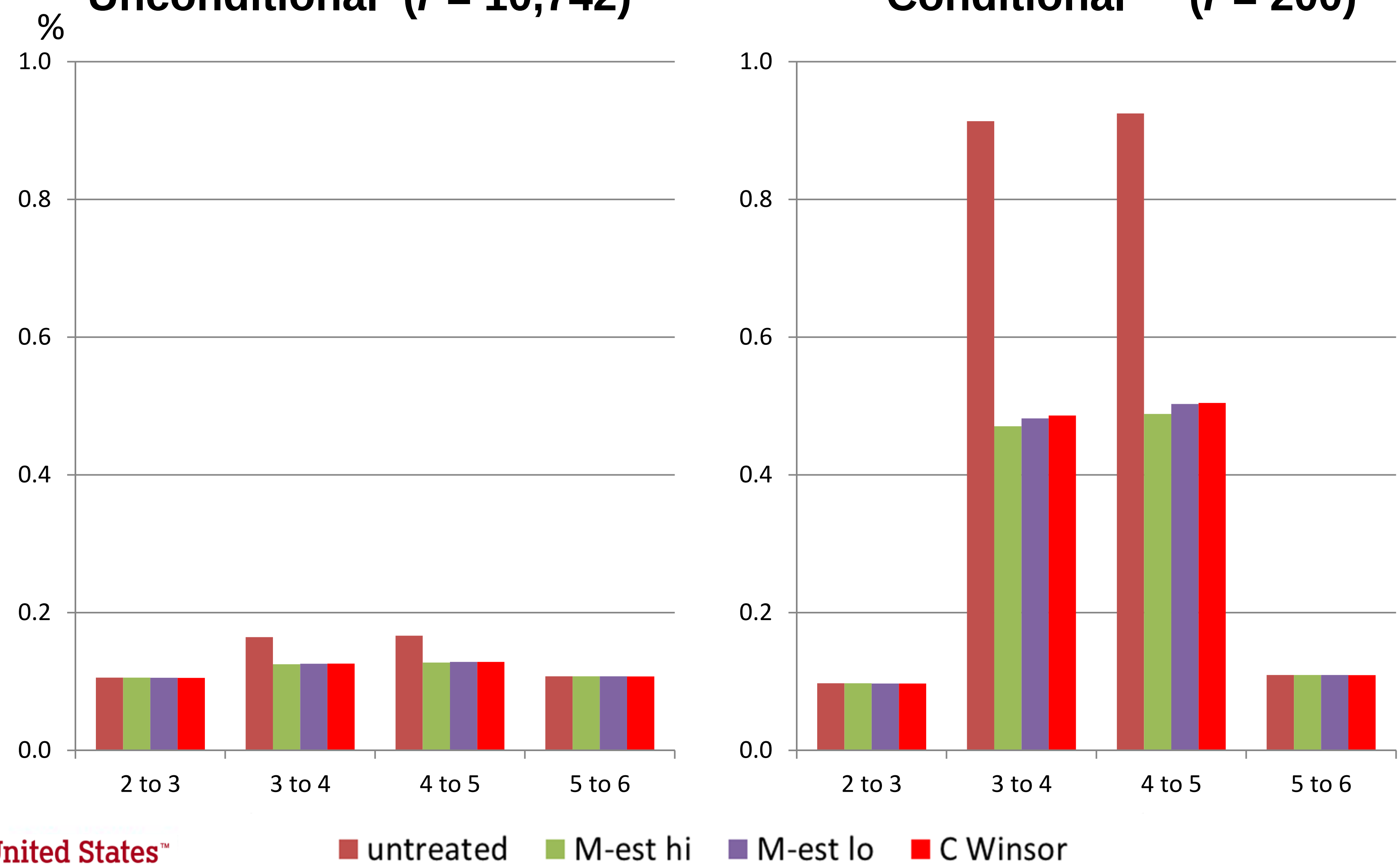
Industry 1; Huber II (percentages)



RRMSE for Month-to-Month Change

Industry 1, Huber II (percentages)

Unconditional (*r* = 10,742)   Conditional (*r* = 200)



Summary

Conditional & Unconditional Analyses

- All methods offer improved results over no treatment when sample contains influential value
- Relative bias
  - Least biased: M-estimation with high φ
  - Similar performance: M-estimation with low φ and Clark Winsorization
- RRMSE
  - Comparable for all methods
    - bias/variance trade-off more apparent for total sales

Summary of methods

Clark Winsorization

- Displays effective performance
  - Trims 0.5% of observations when no infl. value present
  - Detects infl. value when present
- Ease of implementation
  - No input parameters
- No advance knowledge of population required
  - Not flexible
- No convergence problems since algorithm is not iterative

M-Estimation

- Input parameters affect performance
  - Low φ: trims 0.5% when no infl. value present
  - High φ: effective, no trimming
- Flexibility of application needed for wide range of populations
- Some prior knowledge of population required
  - Data for setting parameters
- May have convergence problems when using low initial φ

Next steps

- Chose to focus on M-estimation (high φ)
  - Full endorsement of program managers
- Research underway on how to set initial φ in ongoing monthly surveys
- Issues include
  - Accommodating seasonal effects
  - Data-driven method for selection of initial φ
  - Accounting for changing economy

Contact

Mary.H.Mulry@Census.Gov

Broderick.E.Oliver@Census.Gov

Stephen.Kaputa@Census.Gov