Robust imputation procedures in the presence of influential units in surveys

David Haziza

Department of mathematics and statistics
University of Ottawa

Joint work with

Jia Ning Zhang (University of Ottawa)

and

Sixia Chen (University of Oklahoma)

UNECE 2022

October 3, 2022

Influential units

- In practice, we often face the problem of influential values in the selected sample
- An influential unit is a legitimate unit of the finite population. It is not a measurement error:
  - Gross error;
  - Measurement errors are detected at the editing stage and are treated either manually or by some form of imputation.
- Assumption: Influential units are legitimate observations (not errors)
- Survey statistics are typically sensitive to the presence of influential units

Influential units

- Including or excluding an influential unit in the calculation of survey statistics can have a dramatic impact on their magnitude

  $\longrightarrow$ Their presence in the sample tends to make classical estimators very unstable

  $\longrightarrow$ large variance

- Common issue in business surveys that collect economic variables whose distributions are highly skewed
  - Influential units are often associated with very large values or very large errors
  - Stratum jumpers: may combine a very large value and a large sampling weight

Influential units

- In the presence of influential units, an imputed estimator of a population total:
  - is (approximately) unbiased provided that the imputation model is correctly specified
  - may have a very large variance
- Treatment of influential values: produces stable but biased estimators $\longrightarrow$ trade-off between bias and variance
- Objective: reduce the influence of units that have a large influence
- Our hope: the mean square error of the robust version is smaller than that of the corresponding classical estimator
- How to impute/estimate in the presence of influential units?

The setup

- U: finite population of size N;
- Goal: estimate a population total of a survey variable y:

$$t_y = \sum_{i \in U} y_i$$

- S: sample of size n selected according to a given sampling design $p(S)$;
- I_i: sample selection indicator such that $I_i = 1$ if $i \in S$, and $I_i = 0$, otherwise;
- Design-unbiased (or p-unbiased) estimator of t_y:

$$\widehat{t}_{HT} = \sum_{i \in S} d_i y_i$$

  - $d_i = 1/\pi_i$: design weight attached to unit i;
  - π_i: first-order inclusion probability attached to unit i

The setup

- The survey variable Y is prone to missing values.
- Let r_i be the response indicator such that

$$r_i = \begin{cases} 1, & \text{if } y_i \text{ is observed,} \\ 0, & \text{if } y_i \text{ is missing.} \end{cases}$$

- Set of respondents: $S_r = \{i \in S; r_i = 1\}$.
- Set of nonrespondents: $S_m = \{i \in S; r_i = 0\}$.
- Imputed estimator of t_y:

$$\widehat{t}_I = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i y_i^*,$$

where y_i^* is the imputed value for the missing y_i.

Deterministic linear regression imputation

- x: vector of fully observed variables
- Imputation model

$$y_i = \mathrm{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

such that

$$\mathbb{E}(\epsilon_i \mid \mathrm{x}_i) = 0, \mathbb{E}(\epsilon_i \epsilon_j \mid \mathrm{x}_i, \mathrm{x}_j) = 0, i \neq j \text{ and } \mathbb{V}(\epsilon_i \mid \mathrm{x}_i) = \sigma^2 \phi_i$$

with $\phi_i > 0$ (known)

- Estimator of $\boldsymbol{\beta}$ based on the responding units:

$$\widehat{\mathrm{B}}_{\mathrm{WLS}} = \left(\sum_{i \in S_r} d_i \mathrm{x}_i \phi_i^{-1} \mathrm{x}_i^\top \right)^{-1} \sum_{i \in S_r} d_i \mathrm{x}_i \phi_i^{-1} y_i$$

- Imputed value: $y_i^* = \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{WLS}}$

Imputed estimator

- Estimator of t_y after deterministic linear regression imputation:

$$\widehat{t}_{I,WLS} = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{WLS}}$$

- If the first moment of the imputation model is correctly specified, we have

$$\mathbb{E}_m \mathbb{E}_p \mathbb{E}_q (\widehat{t}_{I,WLS} - t_y) = 0.$$

- That is, the estimator $\widehat{t}_{I,WLS}$ is *mpq*-unbiased for t_y.
- However, $\widehat{t}_{I,WLS}$ may be inefficient in the presence of influential units.

Two methods commonly used in practice

- Robust regression: Replace the estimator $\widehat{\mathrm{B}}_{\mathrm{WLS}}$ by a robust version $\widehat{\mathrm{B}}_{\mathrm{R}}(c)$; for instance an *M*-estimator based on the Huber function;
  $\longrightarrow$ $\widehat{\mathrm{B}}_{\mathrm{R}}(c)$ is solution of

$$\sum_{i \in S_r} \psi_c \left(\frac{y_i - \mathrm{x}_i^\top \boldsymbol{\beta}}{\sqrt{\phi_i}\widehat{\sigma}} \right) \frac{\mathrm{x}_i}{\sqrt{\phi_i}} = 0,$$

  where $\psi_c(\cdot)$ is the so-called Huber function and c is a tuning constant.
- Typically, the value is set to 1.345 (as in classical statistics)
- Imputed value: $y_i^* = \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{R}}(1.345)$
- Other ψ-functions: Biweight, Andrew, etc.
- Other estimators: GM, MM, LTS estimators, etc.
- Objective of robust regression : describe the behavior of the inliers (the non-outliers)

Huber function

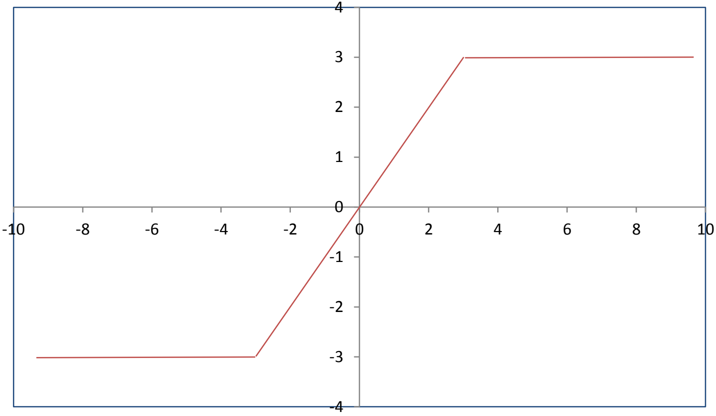


Figure 1: Huber function with $c = 3$

Two methods commonly used in practice

- Excluding outliers: Identify the influential units (usually by an outlier detection method), remove these units and obtain a predicted value obtained by fitting the customary linear regression model
- Imputed value: $y_i^* = \mathrm{x}_i^\top \widehat{\mathrm{B}}^*_{\mathrm{WLS}}$, where

$$\widehat{\mathrm{B}}^*_{\mathrm{WLS}} = \left(\sum_{i \in S_r} \omega_i \mathrm{x}_i \phi_i^{-1} \mathrm{x}_i^\top \right)^{-1} \sum_{i \in S_r} \omega_i \mathrm{x}_i \phi_i^{-1} y_i,$$

  where $\omega_i = d_i$ if i is not discarded and $\omega_i = 0$ if i is discarded.
- Underlying assumption: the discarded respondent y-values are unique; i.e., they do not represent similar non-respondents $\longrightarrow$ nonrepresentative respondents

A simulation study

Are these methods satisfactory?

- We repeated 10, 000 iterations of the following process:
  (1) A population U of size $N = 10,000$ was generated, with one survey variable Y and one covariate X using a mixture of normal distribution with a proportion of outliers equal to 5%;
  (2) A sample S of size $n = 100; 200; 500$ was selected from U according to simple random sampling without replacement;
  (3) Nonresponse to Y was generated according to a uniform nonresponse mechanism with $p_i = 50\%$ for all i;
  (4) Missing values were imputed using 3 imputation procedures.

A simulation study: Point estimators

We computed three types of imputed estimators:

- Non-robust estimator:

$$\widehat{t}_{I,WLS} = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{WLS}}$$

- Based on robust regression:

$$\widehat{t}_I(c) = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{R}}(c)$$

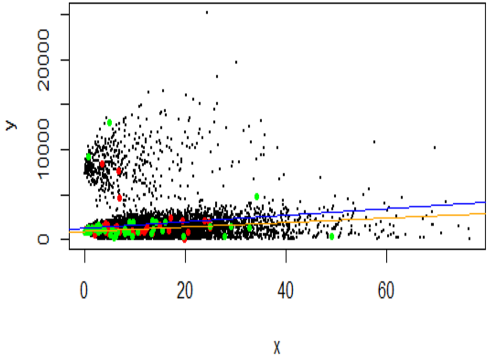
  We used the Huber function with $c = 0.1; 1.345; 2.5$.

- Excluding the outliers:

$$\widehat{t}^*_{I,WLS} = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}^*_{\mathrm{WLS}}$$

  We used the Cook distance with threshold $c = 4/(n-3)$ and studentized residuals with $c = 2; 2.5; 3$.

A simulation study: Asymmetric outliers



$n = 100$

- Respondent
- Nonrespondent
- Nonsampled unit

------- Least squares regression
------- Robust regression

A simulation study: Results

- Monte carlo percent relative bias :

$$\mathrm{RB}(\widehat{t}_I) = \frac{\mathbb{E}_{MC}\left(\widehat{t}_I - t_y\right)}{t_y} \times 100$$

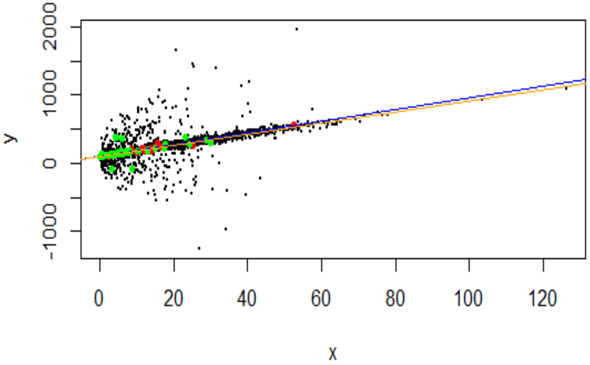
- Relative efficiency:

$$\mathrm{RE} = 100 \times \frac{\mathrm{MSE}_{MC}(\widehat{t}_I)}{\mathrm{MSE}_{MC}(\widehat{t}_{I,WLS})}$$

	WLS	Robust regression			WLS (Exclude outliers)			
n		$c = 0.1$	$c = 1.345$	$c = 2.5$	Studentized $c = 2$	Studentized $c = 2.5$	Studentized $c = 3$	Cook distance
100	-0.0 (100)	-11.5 (78)	-10.7 (73)	-9.7 (70)	-9.3 (82)	-8.3 (84)	-7.5 (86)	-7.5 (87)
200	-0.2 (100)	-11.6 (128)	-10.8 (116)	-9.5 (102)	-9.1 (113)	-7.9 (111)	-6.9 (109)	-7.1 (110)
500	-0.2 (100)	-11.6 (260)	-10.8 (230)	-9.4 (190)	-8.5 (189)	-7.1 (166)	-6.0 (149)	-6.2 (156)

Table 1: Monte Carlo percent relative bias and Monte Carlo relative efficiency of several estimators

A simulation study: Symmetric outliers



n =100

- Respondent
- Nonrespondent
- Nonsampled unit

------- Least squares regression

------- Robust regression

A simulation study

- Monte carlo percent relative bias :

$$\mathrm{RB}(\widehat{t}_I) = \frac{\mathbb{E}_{MC}\left(\widehat{t}_I - t_y\right)}{t_y} \times 100$$

- Relative efficiency:

$$\mathrm{RE} = 100 \times \frac{\mathrm{MSE}_{MC}(\widehat{t}_I)}{\mathrm{MSE}_{MC}(\widehat{t}_{I,WLS})}$$

	WLS	Robust regression			WLS (Exclude outliers)			
n		$c = 0.1$	$c = 1.345$	$c = 2.5$	Studentized $c = 2$	Studentized $c = 2.5$	Studentized $c = 3$	Cook distance
100	-0.1 (100)	-0.1 (57)	-0.1 (57)	-0.1 (58)	-0.1 (57)	-0.1 (58)	-0.1 (60)	-0.1 (59)
200	-0.1 (100)	-0.0 (57)	-0.0 (57)	-0.0 (58)	-0.0 (57)	-0.0 (58)	-0.0 (59)	-0.0 (58)
500	-0.0 (100)	-0.0 (57)	-0.0 (57)	-0.0 (58)	-0.0 (57)	-0.0 (58)	-0.0 (59)	-0.0 (58)

Table 2: Monte Carlo percent relative bias and Monte Carlo relative efficiency of several estimators

Are these methods satisfactory?

- In the case of symmetric outliers, robust regression and weighted least squares regression after removing outliers, behave very well in terms of bias and efficiency;
- In the case of asymmetric outliers:
  - Robust regression and weighted least squares regression may work well in some scenarios but they tend to breakdown as the sample size increases
  - Why? Because the tuning constant c (e.g., $c = 1.345$) was fixed $\longrightarrow$ not adaptative
- c should be adaptative $\longrightarrow$ c increases as n increases
- At least two criteria: Determine the value of c that minimizes
  - the estimated mean square error of the robust estimator: complex without simplifying assumptions
  - the maximum estimated conditional bias of the robust estimator; Beaumont et al. (2013); Chen et al. (2022)

Influence of a unit

- How measure the influence (or impact) of a unit?
- We measure the influence of $i \in S_r$ (respondent) using the concept of conditional bias:

$$B_i = \mathbb{E}_m \mathbb{E}_p \mathbb{E}_q \left(\widehat{t}_{I,WLS} - t_y \mid Y_i = y_i, I_i = 1, r_i = 1 \right).$$

- After some algebra, we obtain

$$B_i \approx \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j + d_i \left(\sum_{\ell \in U} (1 - p_\ell) \mathrm{x}_\ell^\top \right) \left(\sum_{\ell \in U} p_\ell \mathrm{x}_\ell \phi_\ell^{-1} \mathrm{x}_\ell^\top \right)^{-1} \mathrm{x}_i \phi_i^{-1} (y_i - \mathrm{x}_i^\top \mathrm{B})$$

- First term on the right hand-side: influence of unit i on the sampling error
- Second term on the right hand-side: influence of unit i on the nonresponse error
- B_i: unknown $\longrightarrow$ It must be estimated

Influence of a unit

- Special case: simple random sampling without replacement and simple linear regression imputation (i.e., $\mathrm{x}_i = (1, x_i)^\top$ and $\phi_i = 1$):

$$\widehat{B}_i \approx \left(\frac{N}{n} - 1\right)(y_i - \overline{y}_I) + \frac{1}{\widehat{p}}\left\{(1 - \widehat{p}) + \frac{(x_i - \overline{x})(\overline{x} - \overline{x}_r)}{s_{xr}^2}\right\}\left(y_i - \widehat{B}_{0,WLS} - \widehat{B}_{1,WLS} x_i\right),$$

  where

$$\overline{y}_I = \widehat{t}_I / N, \quad \widehat{p} = n_r / n, \quad s_{xr}^2 = (n_r - 1)^{-1} \sum_{i \in S_r} (x_i - \overline{x}_r)^2$$

- Responding unit i has a large influence if
  - The sampling fraction n/N is small;
  - Its y-value is far from the overall estimated mean $\overline{y}_I$;
  - The response rate is low;
  - Its x-value is far from the overall estimated mean $\overline{x} \longrightarrow$ high leverage point;
  - It has a large vertical residual, $y_i - \widehat{B}_{0,WLS} - \widehat{B}_{1,WLS} x_i$

First proposal

- Following Beaumont et al. (2013), we consider a robust version of

$$\widehat{t}_{I,WLS} = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{WLS}$$

based on the concept of conditional bias:

$$\widehat{t}_{I,CB}(c) \;=\; \widehat{t}_{I,WLS} - \sum_{i \in S_r} \widehat{B}_i + \sum_{i \in S_r} \psi_c \left\{ \widehat{B}_i \right\} \equiv \widehat{t}_{I,WLS} + \Delta(c),$$

where $\psi_c(\cdot)$ denotes the Huber function.

- Proposal: select the value of c that minimizes

$$\max_{i \in S_r} \left| \widehat{B}_i^R \right|,$$

where $\widehat{B}_i^R$ is the conditional bias (influence) of unit i on the robust estimator $\widehat{t}_{I,CB}(c)$.

First proposal

- Resulting estimator:

$$\widehat{t}_{I,CB}(c_{opt}) \;=\; \widehat{t}_{I,WLS} - \frac{1}{2}\left[\min_{i\in S_r}\left\{\widehat{B}_i\right\} + \max_{i\in S_r}\left\{\widehat{B}_i\right\}\right]$$

- The value c_{opt} is obtained by solving

$$\Delta(c) = -\frac{1}{2}\left[\min_{i\in S_r}\left\{\widehat{B}_i\right\} + \max_{i\in S_r}\left\{\widehat{B}_i\right\}\right]$$

- There always exists a solution to the previous equation but the solution may not be unique; see Beaumont et al. (2013) and Favre Martinoz et al. (2015).

- c_{opt} increases as n increases $\longrightarrow$ $\widehat{t}_{I,CB}(c_{opt})$ is a consistent estimator of t_y; see Chen et al. (2022).

Second proposal

- Idea: Propose an adaptative tuning constant c, c_{new}, and use robust regression (based on Huber function say) with this constant.
- Let $\widehat{\mathrm{B}}_{\mathrm{R}}(c_{\text{new}})$ be the solution of

$$\sum_{i \in S_r} \psi_{c_{\text{new}}}\left(\frac{y_i - \mathrm{x}_i^\top \boldsymbol{\beta}}{\widehat{\sigma}\sqrt{\phi_i}}\right)\frac{\mathrm{x}_i}{\sqrt{\phi_i}} = 0,$$

  where $\psi(\cdot)$ is the Huber function.
- Should we use the following estimator?

$$\widehat{t}_{I,R}(c_{\text{new}}) = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{R}}(c_{\text{new}})$$

- May not be a good idea because we are only "taking care" of the missing values. However, some respondents may also be influential

Second proposal

- If $\phi_i = \boldsymbol{\lambda}^\top \mathrm{x}_i$, then

$$\widehat{t}_{I,WLS} = \sum_{i \in S} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{WLS}}$$

$\longrightarrow$ Projection form.

- Proposal:

$$\widehat{t}_{I,R}(c_{\mathrm{new}}) = \sum_{i \in S} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{R}}(c_{\mathrm{new}}),$$

where

$$c_{\mathrm{new}} = 1.345 \left\{ 1 + \frac{\left| \min_{i \in S_r} \left\{ \widehat{B}_i^* \right\} + \max_{i \in S_r} \left\{ \widehat{B}_i^* \right\} \right|}{2} \right\} + \frac{n}{N} \sqrt{n},$$

where $\widehat{B}_i^*$ denotes the standardized version of $\widehat{B}_i$.

Second proposal

$$c_{\text{new}} = 1.345\left\{1 + \frac{\left|\min_{i\in S_r}\left\{\widehat{B}_i^*\right\} + \max_{i\in S_r}\left\{\widehat{B}_i^*\right\}\right|}{2}\right\} + \frac{n}{N}\sqrt{n}$$

- If n/N small, the second term on the right hand-side is small $\longrightarrow$ we can omit it:
  - If the distribution has symmetric outliers, then c_{new} will be slightly larger than 1.345.
  - If the distribution has asymmetric outliers (say to the right), then c_{new} will be larger than 1.345.
- If n gets larger, then the second term on the right hand-side gets larger and $\widehat{\mathrm{B}}_{\mathrm{R}}(c_{\text{new}})$ get closer and closer to $\widehat{\mathrm{B}}_{\mathrm{WLS}}$

Simulation study: Set-up

10,000 iterations of the following process:

(1) Generate a population of size $N = 1,000$;

Models used to generate the populations:

$$y_i \mid x_i \sim \mathcal{D}(\mu_i; \sigma^2 \phi_i),$$

- $\mu_i = \beta_0 + \beta_1 x_i$ and $\phi_i = x_i$; $\quad x_i \sim Gamma(1, 10)$;
- $\mathcal{D}$: Normal, Lognormal, Pareto, Frechet, Weibull, Student, mixture of normals, mixture of lognormals.

(2) From the population, select a sample of size $n = 50; 100; 200$ according to simple random sampling without replacement.

(3) In each sample: generate nonresponse to the y-variable according to an uniform nonresponse mechanism with probability 50%.

Simulation study: Point estimators

- In each sample, we computed four estimators of t_y:
  - The non-robust estimator:

$$\widehat{t}_{I,WLS} = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{WLS}$$

  - The naive estimator:

$$\widehat{t}_{I,R}(1.345) = \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_R(1.345)$$

  - The robust estimator based on the conditional bias:

$$\widehat{t}_{I,CB}(c_{opt}) = \widehat{t}_{I,WLS} - \frac{1}{2}\left[\min_{i \in S_r}\left\{\widehat{B}_i\right\} + \max_{i \in S_r}\left\{\widehat{B}_i\right\}\right]$$

  - The robust estimator based on c_{new}:

$$\widehat{t}_{I,R}(c_{\mathrm{new}}) = \sum_{i \in S} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{R}}(c_{\mathrm{new}})$$

Simulation study: Results

	Point estimator	Normal distribution	Lognormal distribution	Pareto distribution
$n = 50$	$\widehat{t}_{I,WLS}$	-0.3 (100)	-0.1 (100)	-0.1 (100)
	$\widehat{t}_{I,R}(1.345)$	-0.4 (101)	-13.5 (73.6)	-8.3 (51)
	$\widehat{t}_{I,CB}(c_{opt})$	-0.8 (100)	-7.2 (77)	-4.9 (56)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-0.2 (101)	-8.7 (73)	-7.0 (38)
$n = 100$	$\widehat{t}_{I,WLS}$	0.0 (100)	-0.5 (100)	-0.0 (100)
	$\widehat{t}_{I,R}(1.345)$	0.0 (102)	-14.6 (101)	-8.6 (59)
	$\widehat{t}_{I,CB}(c_{opt})$	-0.3 (100)	-5.7 (84)	-3.8 (57)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-0.3 (100)	-6.1 (79)	-5.2 (39)
$n = 200$	$\widehat{t}_{I,WLS}$	0.0 (100)	-0.2 (100)	-0.0 (100)
	$\widehat{t}_{I,R}(1.345)$	0.0 (102)	-14.6 (151)	-8.6 (87)
	$\widehat{t}_{I,CB}(c_{opt})$	-0.2 (100)	-3.6 (89)	-2.5 (64)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-0.2 (100)	-2.8 (89)	-3.1 (49)

Table 3: Monte Carlo percent relative bias and relative efficiency of several estimators

Simulation study: Results

	Point estimator	Frechet distribution	Weibull distribution	Student distribution
$n = 50$	$\widehat{t}_{I,WLS}$	-0.1 (100)	0.0 (100)	0.4 (100)
	$\widehat{t}_{I,R}(1.345)$	-9.2 (52)	-17.0 (87)	0.3 (73)
	$\widehat{t}_{I,CB}(c_{opt})$	-5.4 (57)	-8.1 (86)	0.0 (81)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-7.6 (43)	-9.5 (86)	-0.0 (74)
$n = 100$	$\widehat{t}_{I,WLS}$	0.0 (100)	-0.1 (100)	0.0 (100)
	$\widehat{t}_{I,R}(1.345)$	-9.4 (67)	-17.9 (122)	0.1 (72)
	$\widehat{t}_{I,CB}(c_{opt})$	-4.1 (65)	-5.7 (92)	-0.1 (84)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-5.6 (51)	-5.7 (92)	-0.1 (78)
$n = 200$	$\widehat{t}_{I,WLS}$	0.0 (100)	-0.0 (100)	-0.1 (100)
	$\widehat{t}_{I,R}(1.345)$	-9.7 (93)	-18.5 (192)	0.0 (71)
	$\widehat{t}_{I,CB}(c_{opt})$	-3.0 (69)	-3.6 (95)	-0.2 (87)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-3.4 (54)	-3.6 (95)	-0.0 (89)

Table 4: Monte Carlo percent relative bias and relative efficiency of several estimators

Simulation study: Results

	Point estimator	Mixture normal (0.01)	Mixture normal (0.03)	Mixture normal (0.05)
$n = 50$	$\widehat{t}_{I,WLS}$	0.1 (100)	-0.1 (100)	-.0.5 (100)
	$\widehat{t}_{I,R}(1.345)$	-1.8 (78)	-5.2 (67)	-7.6 (65)
	$\widehat{t}_{I,CB}(c_{opt})$	-1.8 (83)	-3.8 (79)	-4.5 (82)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-2.2 (76)	-6.0 (71)	-8.0 (79)
$n = 100$	$\widehat{t}_{I,WLS}$	0.1 (100)	-0.1 (100)	0.1 (100)
	$\widehat{t}_{I,R}(1.345)$	-1.9 (78)	-5.3 (72)	-8.1 (78)
	$\widehat{t}_{I,CB}(c_{opt})$	-1.5 (85)	-3.1 (86)	-3.8 (91)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-1.7 (79)	-4.6 (79)	-6.3 (89)
$n = 200$	$\widehat{t}_{I,WLS}$	0.0 (100)	0.1 (100)	-0.1 (100)
	$\widehat{t}_{I,R}(1.345)$	-1.9 (82)	-5.2 (85)	-7.7 (101)
	$\widehat{t}_{I,CB}(c_{opt})$	-1.2 (89)	-2.0 (93)	-2.1 (96)
	$\widehat{t}_{I,R}(c_{\text{new}})$	-0.7 (90)	-2.0 (91)	-1.7 (96)

Table 5: Monte Carlo percent relative bias and relative efficiency of several estimators

Simulation study: Results

	Point estimator	Mixture lognormal (0.01)	Mixture lognormal (0.03)	Mixture lognormal (0.05)
$n = 50$	$\widehat{t}_{I,WLS}$	0.1 (100)	0.0 (100)	-.0.1 (100)
	$\widehat{t}_{I,R}(1.345)$	-1.6 (55)	-4.0 (48)	-6.1 (51)
	$\widehat{t}_{I,CB}(c_{opt})$	-1.3 (63)	-2.8 (63)	-3.9 (69)
	$\widehat{t}_{I,R}(c_{\mathrm{new}})$	-2.0 (44)	-5.4 (47)	-7.9 (61)
$n = 100$	$\widehat{t}_{I,WLS}$	0.0 (100)	0.0 (100)	0.1 (100)
	$\widehat{t}_{I,R}(1.345)$	-1.8 (59)	-4.1 (58)	-5.0 (63)
	$\widehat{t}_{I,CB}(c_{opt})$	-1.2 (66)	-2.4 (72)	-3.1 (80)
	$\widehat{t}_{I,R}(c_{\mathrm{new}})$	-1.8 (48)	-4.7 (57)	-6.8 (79)
$n = 200$	$\widehat{t}_{I,WLS}$	0.0 (100)	-0.1 (100)	0.0 (100)
	$\widehat{t}_{I,R}(1.345)$	-1.8 (66)	-4.0 (79)	-3.6 (81)
	$\widehat{t}_{I,CB}(c_{opt})$	-0.9 (73)	-1.7 (83)	-2.1 (90)
	$\widehat{t}_{I,R}(c_{\mathrm{new}})$	-1.3 (58)	-3.3 (72)	-4.6 (96)

Table 6: Monte Carlo percent relative bias and relative efficiency of several estimators

Implementation via calibrated imputation

- Both robust estimators

$$\widehat{t}_{I,CB}(c_{opt}) = \widehat{t}_{I,WLS} - \frac{1}{2}\left[\min_{i\in S_r}\left\{\widehat{B}_i\right\} + \max_{i\in S_r}\left\{\widehat{B}_i\right\}\right]$$

and

$$\widehat{t}_{I,R}(c_{\text{new}}) = \sum_{i\in S} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}_{\mathrm{R}}(c_{\text{new}})$$

need to be implemented.

- Estimation of totals: data users simply compute

$$\widehat{t}_I = \sum_{i\in S} d_i \widetilde{y}_i, \quad \widetilde{y}_i = r_i y_i + (1 - r_i) y_i^*$$

- How to implement these estimator? $\longrightarrow$ Calibrated imputation

Implementation via calibrated imputation

- Calibrated robust imputation: e.g., Ren and Chambers (2003), Beaumont (2005) and Chen et al. (2022)
- Illustration for $\widehat{t}_{I,R}(c_{\text{new}})$
- Initial imputed values: $y_i^* = \mathrm{x}_i^\top \widehat{\mathrm{B}}_{WLS}$
- We seek final imputed values, $y_{iF}^*,\ i \in S_m$, that minimize

$$\sum_{i \in S} G(y_{iF}^*/y_i^*),$$

subject to

$$\widehat{t}_{I,F} \equiv \sum_{i \in S_r} d_i y_i + \sum_{i \in S_m} d_i y_{iF}^* = \sum_{i \in S} d_i \mathrm{x}_i^\top \widehat{\mathrm{B}}(c_{\text{new}}),$$

where $G(\cdot)$ is a pseudo-distance function.

Estimation of the mean square error

- Estimator of the mean square error of $\widehat{t}_{I,R}(c_{\mathrm{new}})$:

$$\widehat{\mathrm{MSE}} = \widehat{\mathbb{V}}\left(\widehat{t}_{I,R}(c_{\mathrm{new}})\right) + \max\left\{0, \left(\widehat{t}_{I,R}(c_{\mathrm{new}}) - \widehat{t}_{I,WLS}\right)^2 - \widehat{\mathbb{V}}\left(\widehat{t}_{I,R}(c_{\mathrm{new}}) - \widehat{t}_{I,WLS}\right)\right\}$$

- Obtaining the terms $\widehat{\mathbb{V}}\left(\widehat{t}_{I,R}(c_{\mathrm{new}})\right)$ and $\widehat{\mathbb{V}}\left(\widehat{t}_{I,R}(c_{\mathrm{new}}) - \widehat{t}_{I,WLS}\right)$ may be obtained using a pseudo-population bootstrap procedure, motivated by the reverse approach of Shao and Steel (1999) for variance estimation in the presence of imputed data.

- Future work: Conduct a simulation study to assess the performance of $\widehat{\mathrm{MSE}}$, in terms of bias.

THANK YOU.