Robust imputation procedures in the presence of influential survey data based on adaptative tuning constants

IASS Webinar • 20 October 2026

Sixia Chen[1] joint work with David Haziza[2] and Jia Ning Zhang[2]

[1]Department of Biostatistics and Epidemiology, Hudson College of Public Health, University of Oklahoma Health Campus, Oklahoma City, OK, USA
[2]Department of Mathematics and Statistics, University of Ottawa, Canada

Contact: **sixia-chen@ou.edu**

Outline

1. Background and motivation
2. A motivating example: health-care expenditures
3. Research gap, innovation and aims
4. Set-up: imputation under item nonresponse
5. Two naive robust fixes — and why they fail
6. The conditional bias as a measure of influence
7. Three data-adaptive procedures
   (minimax conditional bias, c_{new}, MSE-selected c)
8. Mean squared error estimation
9. Simulation study
10. Application: 2023 Medical Expenditure Panel Survey
11. Conclusions, practical guidance and future work

Background and motivation

Surveys, nonresponse and imputation

- National statistical offices (Statistics Canada, U.S. Census Bureau, BLS, . . .), health agencies and market-research firms all estimate finite-population totals $t_y = \sum_{i \in U} y_i$ from probability samples.
- **Unit nonresponse**: no information at all $\Rightarrow$ weighting adjustments.
- **Item nonresponse**: some answers missing (income, health expenditures, revenue) $\Rightarrow$ imputation.
- Business and health surveys mostly use *deterministic regression imputation* and release one rectangular "completed" file.
- Under *missing at random* and a correctly specified conditional mean, regression imputation removes the nonresponse bias.

Reference texts

Kim & Shao (2021); Haziza (2009); Särndal, Swensson & Wretman (1992).

But . . .

Bias is only half of the mean squared error. What happens to the *variance* when a handful of respondents are extremely large?

Influential units: a second, distinct problem

Definition

A unit is influential if its inclusion or exclusion has a substantial impact on a point estimator.

- Sources of influence: large residual, extreme leverage ($\mathbf{v}_i$), large sampling weight w_i, or a combination.
- **Not** gross measurement errors (those are fixed at the editing stage): the values are correctly recorded.
- Chambers (1986): representative outliers — a correct value that may well represent similar units among the non-sampled or the non-responding part of the population.
- Typical in business surveys (revenue, employment) and in biomedical surveys (expenditures, utilisation): strongly right-skewed study variables.

Three ways to be influential

large residual	$y_i - \mathbf{v}_i^\top \boldsymbol{\beta}$ huge
high leverage	$\mathbf{v}_i$ far from the bulk
large weight	$w_i = 1/\pi_i$ huge

Classical tools (full response)

Winsorisation and trimming (Kokic & Bell 1994; Lee 1995), weight reduction, *M*-estimation (Huber 1964; Hampel 1974); reviews in Beaumont & Rivest (2009) and Wada (2020).

Why the two problems are different

Item nonresponse

- Primary risk: bias (respondents $\neq$ nonrespondents).
- Cure: a good imputation model under MAR; if the first moment is right, the imputed estimator is unbiased whatever the tails look like.

Influential respondents

- Correctly recorded values do **not** bias the WLS imputed estimator.
- Primary effect: variance inflation; any "cure" that clips, downweights or drops them introduces bias.

Robust imputation is therefore a finite-sample bias–variance trade-off

Accept some bias only when it buys a sufficiently large reduction in variance, so that MSE $= \mathbb{V} + \text{Bias}^2$ goes down.

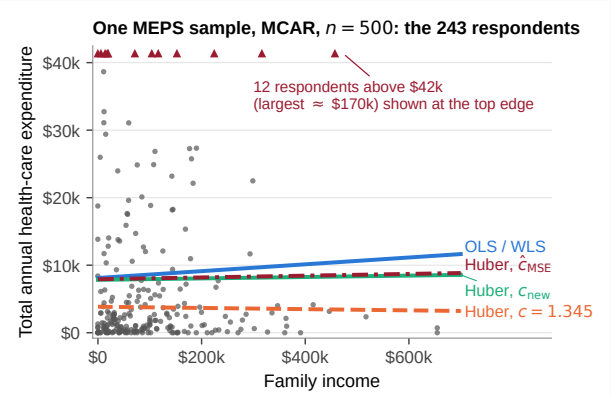
- Flexible/ML imputation (random forests, boosting, penalised regression; Thomas & Rajabi 2021; Chen & Xu 2023; Sun et al. 2023; Sang et al. 2022; Lee & Kim 2022) targets *model misspecification* — a different problem. Without a robust loss or an influence-control mechanism it does nothing against large residuals, leverage or weights.

A motivating example: health-care expenditures

The Medical Expenditure Panel Survey (MEPS)

- Run by the Agency for Healthcare Research and Quality (AHRQ); nationally representative of the U.S. civilian non-institutionalised population; multistage design; collects utilisation, expenditures, insurance, demographics and income.
- Annual expenditures are extraordinarily concentrated (AHRQ Statistical Brief #560, 2022 data):

Group of persons	Share of total spending	Mean spending
Top 1%	21.7%	$147,071
Top 5%	49.7%	$67,321
Top 50%	97.2%	$13,158
Bottom 50%	2.8%	$374
Everyone	100%	$6,765



Redrawn from Figure 1 of the paper (MCAR, $n = 500$): OLS fit and Huber fits with $c = 1.345$, c_{new}, $\hat{c}_{\mathrm{MSE}}$; y-axis cut at \$42k so that the lines can be compared.

Why it matters for imputation: the 1% who hold a fifth of the total are exactly the ones a robust fit discounts.

What one influential respondent does to an estimated total

Back-of-the-envelope (MEPS set-up of Section 10)

- Finite population $N = 18{,}919$, simple random sample $n = 150$: every unit carries $w_i = N/n \approx 126$.
- One respondent with $y_i = \$150{,}000$ contributes $w_i y_i \approx \$18.9$ million *directly* to the estimated total.
- About half the sample is missing: the same unit also shifts the regression line used to impute ≈ 75 nonrespondents — an indirect contribution.

The dilemma

- Keep the unit as is $\Rightarrow$ unbiased, but the estimate swings by tens of millions depending on whether this one person was sampled (**variance**).
- Clip or drop it $\Rightarrow$ stable, but we pretend that the 1% who hold 22% of spending do not exist among nonrespondents and non-sampled units (**bias**).

What we need

A rule that decides *from the data* how much to clip — more when the sample is small and unstable, less when n is large and the bias would dominate.

Research gap, innovation and aims

What practitioners do today

Robust regression at the imputation stage

Replace WLS by an *M*-estimator (Huber, biweight, Andrews; GM-, MM-estimators) with a fixed tuning constant, e.g. Huber $c = 1.345$ (95% efficiency under normal errors). Maronna et al. (2019); Andersen (2008).

Detect and exclude

Flag outliers (studentised residuals, Cook's distance), refit WLS without them, impute from the clean fit.

Robust imputation literature

Little & Smith (1987); Chambers & Ren (2004, reverse calibration); Vanden Branden & Verboven (2009); Templ, Kowarik & Filzmoser (2011); Templ (2023); Templ & Ulmer (2024); Wada (2020).

Data-adaptive robustification

Conditional-bias minimax: Beaumont, Haziza & Ruiz-Gazen (2013) [full response]; Favre-Martinoz et al. (2016) [unit nonresponse]; Chen, Haziza & Michal (2024) [item nonresponse, multiply robust]. Data-driven robustness parameter under MAR: Wang et al. (2024).

Common thread of the two standard fixes

They are designed to describe the *inlier* population (bounded influence, high breakdown). Our target is the total of the *full* finite population — a mixture of inliers and representative outliers.

Research gap

1. **Wrong target.** A fixed tuning constant makes the robust fit track the inlier component. When influential values sit systematically on one side (asymmetric contamination), this produces a persistent bias in the estimated total.
2. **No dependence on the sample size.** The bias of a fixed-c procedure does not vanish as $n \to \infty$, while the variance it saves does; relative efficiency can exceed 100% and *grow without bound*.
3. **Influence is multi-faceted under nonresponse.** Residual, leverage, weight *and* response rate all matter; no single diagnostic covers the sampling *and* the imputation error of an imputed estimator.
4. **Inference.** Mean squared error estimation when the degree of robustification is itself selected from the sample is essentially open.

Guiding principle

The appropriate degree of robustification depends *jointly* on the sample size and on the configuration of the influential observations — so let the data choose it.

Innovation: what is new in this work

Three data-adaptive procedures (two new)

1. **Minimax conditional-bias estimator** $\widehat{t}_{l,\mathrm{CB}}$ — the Beaumont–Haziza–Ruiz-Gazen principle adapted to the WLS imputed estimator.
2. New: an *M*-estimator with a tuning constant c_{new} built from the *asymmetry* of the estimated conditional biases and from $n^{3/2}/N$.
3. New: the tuning constant $\widehat{c}_{\mathrm{MSE}}$ that minimises an estimate of the *design MSE* over a grid.

Supporting contributions

- Analytical characterisation of the bias of fixed-c and "exclude outliers" procedures under mixture contamination.
- Conditional bias of a respondent for the *imputed* estimator: sampling + nonresponse/imputation influence, with a plug-in estimator.
- Design-based MSE estimators: linearisation plug-in and a reweighted pseudo-population bootstrap (RPPB).
- Evidence: 10,000-replicate Monte Carlo over 11 error distributions, plus MEPS 2023.

Research aims

1. Show — empirically and analytically — when and why robust regression imputation with a fixed tuning constant (or outlier exclusion) fails for finite-population totals.
2. Measure the influence of a responding unit on the imputed estimator through its conditional bias, and estimate it from the sample.
3. Build adaptive robust imputation procedures whose degree of robustification responds to the sample size and to the configuration of influential values, and establish their consistency.
4. Propose mean squared error estimators for the adaptive procedures and assess their finite-sample bias.
5. Compare the procedures with WLS, with the fixed-c estimator and with an infeasible "oracle" across symmetric, asymmetric and mixture distributions, and illustrate them on real survey data (MEPS 2023).

Set-up: imputation under item nonresponse

Notation: population, design and nonresponse

Finite population and design

- $U = \{1, \ldots, N\}$; sample S of size n drawn with design $p(S)$.
- $I_i = 1$ if $i \in S$; $\pi_i = \mathbb{P}(I_i = 1)$, $\pi_{ij} = \mathbb{P}(I_i = I_j = 1)$; weights $w_i = 1/\pi_i$.
- Target: $t_y = \sum_{i \in U} y_i$. Full-sample (Horvitz–Thompson) estimator $\widehat{t}_{y,\pi} = \sum_{i \in S} w_i y_i$.

Item nonresponse

- $r_i = 1$ if y_i observed, 0 otherwise; respondents $S_r = \{i \in S : r_i = 1\}$, nonrespondents $S_m = S \setminus S_r$.
- Auxiliary vector $\mathbf{v}_i$ observed for everyone in S.
- **MAR** (Rubin 1976): $p_i \equiv \mathbb{P}(r_i = 1 \mid \mathbf{v}_i, y_i) = \mathbb{P}(r_i = 1 \mid \mathbf{v}_i)$, hence $f(y \mid \mathbf{v}, r = 1) = f(y \mid \mathbf{v}, r = 0)$: the respondents' regression can be transported to the nonrespondents.

Convention used throughout

$\pi_{ii} = \pi_i$, $\pi_{ij} > 0$ for all pairs; $\mathbb{E}_m$, $\mathbb{E}_p$, $\mathbb{E}_q$: expectations with respect to the *imputation model*, the *sampling design* and the *nonresponse mechanism*, each treating everything else as fixed.

Deterministic linear regression imputation

Imputation model

$$y_i = \mathbf{v}_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad \mathbb{E}(\epsilon_i \mid \mathbf{v}_i) = 0, \quad \mathbb{E}(\epsilon_i \epsilon_j \mid \mathbf{v}_i, \mathbf{v}_j) = 0 \ (i \neq j), \quad \mathbb{V}(\epsilon_i \mid \mathbf{v}_i) = \sigma^2 \phi_i,$$

with $\phi_i > 0$ known and $\phi_i = \boldsymbol{\lambda}^\top \mathbf{v}_i$ for some $\boldsymbol{\lambda}$ (e.g. an intercept and $\phi_i = 1$). No distributional assumption on ϵ_i.

Weighted least squares on the respondents

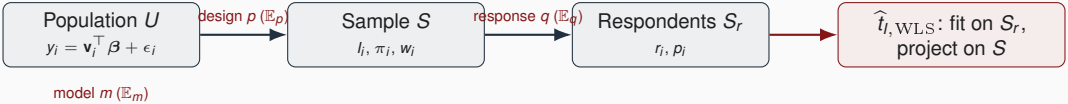
$$\widehat{\mathbf{B}}_{\mathrm{WLS}} = \Big(\sum_{i \in S} w_i r_i \mathbf{v}_i \phi_i^{-1} \mathbf{v}_i^\top \Big)^{-1} \sum_{i \in S} w_i r_i \mathbf{v}_i \phi_i^{-1} y_i$$

imputed value $y_i^* = \mathbf{v}_i^\top \widehat{\mathbf{B}}_{\mathrm{WLS}}$, $i \in S_m$.

Imputed estimator and its projection form

$$\begin{aligned} \widehat{t}_{I,\mathrm{WLS}} &= \sum_{i \in S} w_i r_i y_i + \sum_{i \in S} w_i (1 - r_i) \mathbf{v}_i^\top \widehat{\mathbf{B}}_{\mathrm{WLS}} \\ &= \sum_{i \in S} w_i \mathbf{v}_i^\top \widehat{\mathbf{B}}_{\mathrm{WLS}} \quad \text{(because } \phi_i = \boldsymbol{\lambda}^\top \mathbf{v}_i\text{)}. \end{aligned}$$

Three sources of randomness and the *mpq* framework



Unbiasedness

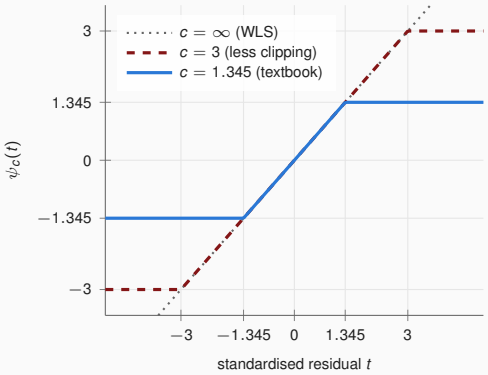
If the first moment of the model is correct, $\mathbb{E}_m\mathbb{E}_p\mathbb{E}_q\left(\widehat{t}_{I,\mathrm{WLS}} - t_y\right) = 0$ (*mpq*-unbiased).

The catch

Correctly recorded influential respondents leave this equality intact — they do not bias $\widehat{t}_{I,\mathrm{WLS}}$ — but can make its variance enormous.

- Every bias/MSE statement in this talk is about one (or a composition) of these three expectations; the MSE-selection approach is purely *design-based* ($\mathbb{E}_p$ given the finite population and the response indicators).

The Huber ψ function and its tuning constant



M-estimation with a Huber score

$\psi_c(t) = t$ if $|t| \le c$ and $\psi_c(t) = c\ \mathrm{sgn}(t)$ otherwise. A residual beyond c scale units is *clipped*: it keeps a vote but not a megaphone.

- Small c: strong resistance, lower variance, but the fit follows the *bulk* of the data.
- Large c: the fit approaches WLS; $c \to \infty$ gives WLS exactly.
- $c = 1.345$ is the textbook value: $\approx 95\%$ efficiency relative to least squares under *normal* errors — a statement about estimating $\boldsymbol{\beta}$, not about estimating t_y.

Two naive robust fixes — and why they fail

Two natural ways to robustify regression imputation

Naive fix 1: robust regression with a fixed c

$\widehat{\mathbf{B}}_R(c)$ solves

$$\sum_{i \in S} w_i r_i \, \psi_c\Big(\frac{y_i - \mathbf{v}_i^\top \boldsymbol{\beta}}{\sqrt{\phi_i}\,\widehat{\sigma}}\Big) \frac{\mathbf{v}_i}{\sqrt{\phi_i}} = \mathbf{0},$$

$\widehat{\sigma}$ a robust scale estimate; then

$$\widehat{t}_{l,R}(c) = \sum_{i \in S} w_i r_i y_i + \sum_{i \in S} w_i (1 - r_i) \mathbf{v}_i^\top \widehat{\mathbf{B}}_R(c).$$

Usual choice: $c = 1.345$. Other ψ's (biweight, Andrews), GM- and MM-estimators are possible.

Naive fix 2: detect, exclude, refit

$a_i = 1$ if unit i is kept, $a_i = 0$ if flagged as an outlier (studentised residuals, Cook's distance, ...):

$$\widehat{\mathbf{B}}^*_{\mathrm{WLS}} = \Big(\sum_{i \in S} w_i r_i a_i \mathbf{v}_i \phi_i^{-1} \mathbf{v}_i^\top\Big)^{-1} \sum_{i \in S} w_i r_i a_i \mathbf{v}_i \phi_i^{-1} y_i,$$

$$\widehat{t}^*_{l,\mathrm{WLS}} = \sum_{i \in S} w_i r_i y_i + \sum_{i \in S} w_i (1 - r_i) \mathbf{v}_i^\top \widehat{\mathbf{B}}^*_{\mathrm{WLS}}.$$

Question

Do $\widehat{t}_{l,R}(c)$ and $\widehat{t}^*_{l,\mathrm{WLS}}$ perform well in the presence of influential values?

A small simulation: two mixture populations

- $R = 10{,}000$ replications; $N = 10{,}000$; $V \sim \mathrm{Gamma}(\text{shape } 1, \text{scale } 10)$.
- $Y_i \sim 0.95\,\mathcal{N}(\mu_{1i}, \sigma^2_{1i}) + 0.05\,\mathcal{N}(\mu_{2i}, \sigma^2_{2i})$ — 5% outliers.
- **Asymmetric outliers**: $\mu_{1i} = 1{,}000 + 5v_i$, $\mu_{2i} = 9{,}000 + 20v_i$, $\sigma^2_{1i} = 19{,}600v_i$, $\sigma^2_{2i} = 640{,}000v_i$.
- **Symmetric outliers**: $\mu_{1i} = \mu_{2i} = 100 + 8v_i$, $\sigma^2_{1i} = 64v_i$, $\sigma^2_{2i} = 14{,}400v_i$.
- SRSWOR, $n \in \{100, 200\}$; uniform nonresponse $p_i = 0.5$.
- Estimators: WLS; Huber with $c \in \{0.1, 1.345, 2.5\}$; exclusion by studentised residuals ($> 2, 2.5, 3$) and by Cook's distance ($> 4/(n-3)$).

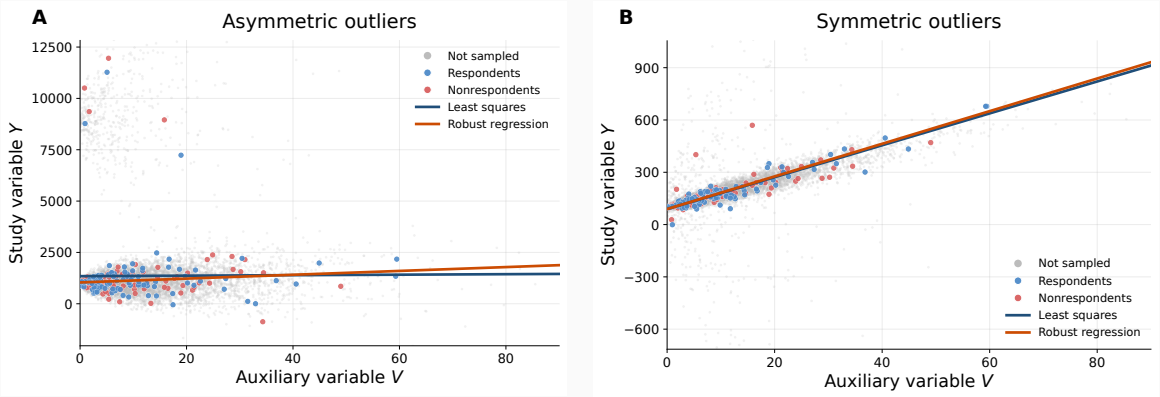
Performance measures (used throughout)

$$\mathrm{RB}_{MC}(\hat{t}_I) = \frac{1}{R}\sum_{r=1}^{R} \frac{\hat{t}_I^{(r)} - t_y}{t_y} \times 100,$$

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

with $\mathrm{MSE}_{MC}(\hat{t}_I) = \frac{1}{R}\sum_r (\hat{t}_I^{(r)} - t_y)^2$. $\mathrm{RE} < 100$: better than WLS; $\mathrm{RE} > 100$: worse.

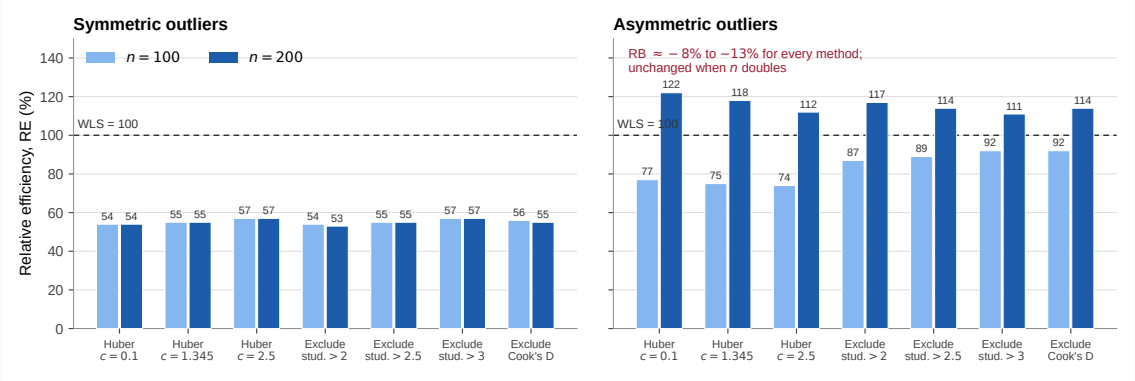
What the two populations look like



One realisation with $n = 100$. Grey: non-sampled; blue: respondents; red: nonrespondents. Lines fitted on the respondents only.

- Asymmetric: the robust line (orange) sits below the least-squares line (blue) — it ignores the upper component that *also* exists among nonrespondents and non-sampled units.
- Symmetric: the two lines are indistinguishable.

Results: relative efficiency and bias of the naive fixes



Symmetric outliers

RB ≈ 0 for every method; RE between 53 and 57: robustifying cuts the MSE almost in half.

Asymmetric outliers

RB from -7.7% to -13%, constant in n; RE rises from 74–92 ($n = 100$) to 111–122 ($n = 200$): worse than WLS.

Analytical investigation: a mixture model with a shifted outlier component

Model

$$Y_i = \Delta_i(\mu_{1i} + \epsilon_{1i}) + (1 - \Delta_i)(\mu_{2i} + \epsilon_{2i}), \qquad \mu_{ki} = \mathbf{v}_i^\top \boldsymbol{\beta}_k, \qquad \mathbb{P}(\Delta_i = 1 \mid \mathbf{v}_i) = \rho \text{ (inlier)},$$

with $\mathbb{E}_m(\epsilon_{ki} \mid \mathbf{v}_i, \Delta_i) = 0$ and $\mathbb{V}_m(\epsilon_{ki} \mid \mathbf{v}_i, \Delta_i) = \sigma_k^2 \phi_i$. Asymmetric case: $\mu_{2i} \gg \mu_{1i}$, $\sigma_2^2 \gg \sigma_1^2$.

- Conditional mean is still linear: $\mathbb{E}_m(Y_i \mid \mathbf{v}_i) = \mathbf{v}_i^\top \boldsymbol{\beta}_{\text{mix}}$ with $\boldsymbol{\beta}_{\text{mix}} = \rho\boldsymbol{\beta}_1 + (1 - \rho)\boldsymbol{\beta}_2$, and $\widehat{\mathbf{B}}_{\text{WLS}} \xrightarrow{p} \boldsymbol{\beta}_{\text{mix}}$.
- Let $\boldsymbol{\beta}_M(c)$ be the probability limit of $\widehat{\mathbf{B}}_R(c)$. Since $\widehat{t}_{l,\text{WLS}}$ is *mpq*-unbiased,

$$\text{Bias}\{\widehat{t}_{l,R}(c)\} \approx \mathbb{E}_p\mathbb{E}_q\Big[\sum_{i\in S} w_i(1 - r_i)\mathbf{v}_i^\top \times \{\boldsymbol{\beta}_M(c) - \boldsymbol{\beta}_{\text{mix}}\}\Big].$$

Well-separated components, fixed c

The robust fit is driven by the inliers: $\boldsymbol{\beta}_M(c) \approx \boldsymbol{\beta}_1$, hence

$$\text{Bias}\{\widehat{t}_{l,R}(c)\} \approx -(1 - \rho)\sum_{i\in U}(1 - p_i)(\mu_{2i} - \mu_{1i}),$$

and under uniform response $p_i = p$:
$-(1 - \rho)(1 - p)\sum_{i\in U}(\mu_{2i} - \mu_{1i})$.

Interpretation, and the symmetric case

Three consequences

1. The first-order bias depends on ρ, p_i and $\mu_{2i} - \mu_{1i}$ — not on n.
2. Increasing n reduces the variance but not the bias: $\mathrm{RE} \to \infty$ as $n \to \infty$.
3. Smaller $c \Rightarrow$ the fit follows the inliers more closely $\Rightarrow$ larger bias.

Same story for "detect and exclude"

If the rule removes mostly outlier-component units, $\boldsymbol{\beta}_A \approx \boldsymbol{\beta}_1$ and $\mathrm{Bias}(\widehat{t}^{\,*}_{l,\mathrm{WLS}}) \approx -(1-\rho)\sum_{i\in U}(1-p_i)(\mu_{2i} - \mu_{1i})$.

Common-location mixture: no bias

$Y_i = \mu_i + \Delta_i\epsilon_{1i} + (1-\Delta_i)\epsilon_{2i}$, $\mu_i = \mathbf{v}_i^\top\boldsymbol{\beta}$, with both error laws symmetric about zero. Then $Y_i - \mathbf{v}_i^\top\boldsymbol{\beta}$ is symmetric, ψ_c is odd, so

$$\mathbb{E}_m\left[\psi_c\left\{\frac{Y_i - \mathbf{v}_i^\top\boldsymbol{\beta}}{\sigma\phi_i^{1/2}}\right\}\middle|\mathbf{v}_i\right] = 0 \;\Rightarrow\; \boldsymbol{\beta}_M(c) = \boldsymbol{\beta},$$

and the first-order bias vanishes for *every* c.

Take-away

A fixed c is harmless when inliers and outliers share the same conditional location; it is harmful when they do not. Real skewed data are usually of the second kind.

A weight-based look: who represents the nonrespondents?

Both estimators are linear in the observed values

$$\widehat{t}_{I,\mathrm{WLS}} = \sum_{i\in S} \widetilde{w}_i\, r_i y_i, \qquad \widetilde{w}_i = w_i\Big\{1 + \sum_{j\in S} w_j(1-r_j)\mathbf{v}_j^\top \Big(\sum_{\ell\in S} w_\ell r_\ell \mathbf{v}_\ell \phi_\ell^{-1}\mathbf{v}_\ell^\top\Big)^{-1}\phi_i^{-1}\mathbf{v}_i\Big\},$$

and $\widehat{t}^{\,*}_{I,\mathrm{WLS}} = \sum_{i\in S} \widetilde{w}_i^{*}\, r_i y_i$ with a_ℓ and a_i inserted in the sums.

- $\widetilde{w}_i = w_i \times$ (design weight) $\times$ (imputation-induced adjustment): how much respondent *i* contributes *indirectly* through the imputed values.
- The adjustment can be negative, so it is not literally a "number of nonrespondents represented".

Excluding unit *i* ($a_i = 0$) gives $\widetilde{w}_i^{*} = w_i$

Its value then enters only through $w_i y_i$: the procedure declares the unit non-representative of the missing units — reasonable for a unique outlier, questionable for a representative one (Chambers 1986).

The conditional bias as a measure of influence

Conditional bias of a responding unit

Definition (Beaumont, Haziza & Ruiz-Gazen 2013, extended to the imputed estimator)

For a responding unit i,

$$B_i = \mathbb{E}_m \mathbb{E}_p \mathbb{E}_q \big(\widehat{t}_{I,\mathrm{WLS}} - t_y \;\big|\; Y_i = y_i,\; I_i = 1,\; r_i = 1 \big).$$

- The expected estimation error *given that* unit i was sampled, responded, and showed the value y_i.
- $|B_i|$ large $\Rightarrow$ the presence of unit i moves the imputed estimator a lot $\Rightarrow$ influential.
- Expressed in the units of t_y; integrates *all* sources of influence: sampling weight, residual, leverage, response rate.

Why "bias"?

Averaged over all values y_i and over the sampling/response of unit i, the conditional biases average to the unconditional bias, which is zero. A unit with a large $|B_i|$ is one whose realisation *pulls* the estimator away from the target.

Lineage

Full response: Beaumont et al. (2013); unit nonresponse: Favre-Martinoz et al. (2016); item nonresponse, multiply robust imputation: Chen et al. (2024).

A first-order expression

Taylor linearisation (Supplement, Appendix A)

$$B_i \;\approx\; \underbrace{(w_i - 1)y_i + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \mathbf{v}_j^\top \beta}_{\text{sampling influence (Beaumont et al. 2013)}} \;+\; \underbrace{w_i \, \mathbf{C} \, \mathbf{v}_i \phi_i^{-1} \left(y_i - \mathbf{v}_i^\top \beta \right)}_{\text{nonresponse / imputation influence}},$$

$$\mathbf{C} = \Big\{ \sum_{j \in U} (1 - p_j) \mathbf{v}_j^\top \Big\} \Big\{ \sum_{j \in U} p_j \mathbf{v}_j \phi_j^{-1} \mathbf{v}_j^\top \Big\}^{-1} \quad (\text{a } 1 \times \dim \mathbf{v} \text{ row vector}).$$

- First term: conditional bias of the sampling error $\widehat{t}_{y,\pi} - t_y$ — the model expectation of the usual design conditional bias of the HT estimator.
- Second term: first-order contribution of unit *i* to $\widehat{t}_{l,\mathrm{WLS}} - \widehat{t}_{y,\pi}$ — through its *weighted residual* and its *leverage* in the respondents' fit, scaled by how much mass is missing $(1 - p_j)$.

Key steps of the derivation

Linearise $\widehat{\mathbf{B}}_{\mathrm{WLS}}$; write $\widehat{t}_{l,\mathrm{WLS}} - t_y = \sum_{k \in S} w_k \xi_k - t_y + o_p(N/\sqrt{n})$; use $\mathbb{E}_q(r_j \mid r_i = 1) = p_j$, $\mathbb{E}_p(I_j \mid I_i = 1) = \pi_{ij}/\pi_i$ and $\mathbb{E}_m(Y_j - \mathbf{v}_j^\top \beta \mid Y_i = y_i) = 0$ for $j \neq i$.

Estimating the conditional bias from the sample

Plug-in estimator

Completed values $\widetilde{y}_j = r_j y_j + (1 - r_j)\mathbf{v}_j^\top \widehat{\mathbf{B}}_{\mathrm{WLS}}$, $j \in S$, and

$$\widehat{B}_i = \sum_{j \in S} \Big(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}} \Big) \widetilde{y}_j + w_i \, \widehat{\mathbf{C}} \, \mathbf{v}_i \phi_i^{-1} \big(y_i - \mathbf{v}_i^\top \widehat{\mathbf{B}}_{\mathrm{WLS}} \big), \qquad i \in S_r,$$

$$\widehat{\mathbf{C}} = \Big\{ \sum_{j \in S} w_j (1 - r_j) \mathbf{v}_j^\top \Big\} \Big\{ \sum_{j \in S} w_j r_j \mathbf{v}_j \phi_j^{-1} \mathbf{v}_j^\top \Big\}^{-1}.$$

- First term: a *conditional* Horvitz–Thompson estimator (the coefficient uses $\mathbb{P}(I_j = 1 \mid I_i = 1) = \pi_{ij}/\pi_i$), applied to the completed data.
- Second term: replace population quantities by their weighted sample analogues.
- Only needs π_i, π_{ij}, the WLS fit and the response indicators — all available in any production system.

Reading the conditional bias: SRSWOR and simple linear regression

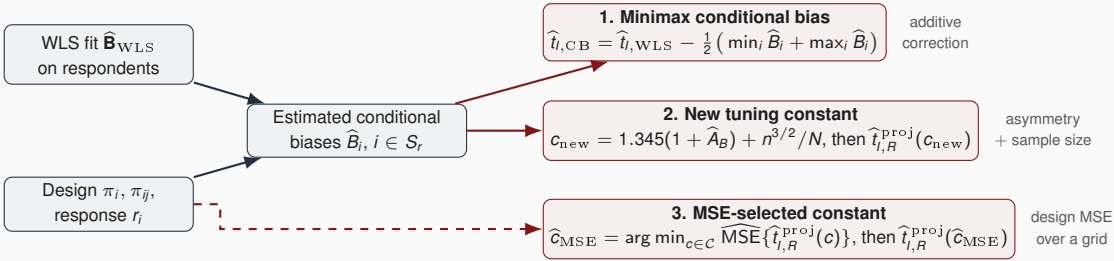
With $\mathbf{v}_i = (1, v_i)^\top$, $\phi_i = 1$, $\widehat{p} = n_r/n$, $\overline{v}$ (sample mean), $\overline{v}_r$, s^2_{vr} (respondent mean and variance) and $\overline{y}_{l,\mathrm{WLS}} = \widehat{t}_{l,\mathrm{WLS}}/N$:

$$\widehat{B}_i \approx \underbrace{\left(\frac{N}{n} - 1\right)(y_i - \overline{y}_{l,\mathrm{WLS}})}_{\text{sampling}} + \underbrace{\left(\frac{N}{n}\right)\frac{1}{\widehat{p}}\left\{(1-\widehat{p}) + \frac{(v_i - \overline{v}_r)(\overline{v} - \overline{v}_r)}{s^2_{vr}}\right\}(y_i - \widehat{B}_{0,\mathrm{WLS}} - \widehat{B}_{1,\mathrm{WLS}} v_i)}_{\text{nonresponse and imputation}}$$

- **Sampling**: large when the sampling fraction is small or y_i is far from the imputed mean.
- **Nonresponse**: amplified by a low response rate $\widehat{p}$ and a large regression residual.
- **Leverage**: through $(v_i - \overline{v}_r)(\overline{v} - \overline{v}_r)$ — an extreme v_i matters when the respondents' mean of V differs from the full-sample mean; if $\overline{v} = \overline{v}_r$ leverage alone is not enough.
- One number summarises sampling, nonresponse, leverage and residual influence.

Three data-adaptive procedures

Overview: three ways to let the data choose the degree of robustness



Quantities appearing in the boxes (defined in detail on the next slides)

- $\widehat{A}_B = \left|\frac{1}{2}\left(\min_{i \in S_r} \widehat{Z}_i + \max_{i \in S_r} \widehat{Z}_i\right)\right|$ with $\widehat{Z}_i = (\widehat{B}_i - \overline{B}_r)/s_B$ the standardised conditional biases: how one-sided the influence is.
- $\widehat{t}_{l,R}^{\,\mathrm{proj}}(c) = \sum_{i \in S} w_i \mathbf{v}_i^\top \widehat{\mathbf{B}}_R(c)$, the robust *projection* estimator built on the Huber M-estimator $\widehat{\mathbf{B}}_R(c)$ (compare $\widehat{t}_{l,\mathrm{WLS}} = \sum_{i \in S} w_i \mathbf{v}_i^\top \widehat{\mathbf{B}}_{\mathrm{WLS}}$).
- Approaches 1–2 use the $\widehat{B}_i$; approach 3 uses the design (π_{ij}) and $\widehat{t}_{l,\mathrm{WLS}}$ as a reference. All three revert towards WLS as the sample becomes informative.

Approach 1: the minimax conditional-bias estimator

Idea (Beaumont et al. 2013; Chen et al. 2024)

Adjust the WLS imputed estimator additively,

$$\widehat{t}_{I,\mathrm{CB}}(\lambda) = \widehat{t}_{I,\mathrm{WLS}} + \lambda,$$

so that the conditional bias of every respondent, $\widehat{B}_i + \lambda$, is as small as possible in the worst case:

$$\widehat{\lambda}_{\mathrm{CB}} = \arg\min_{\lambda \in \mathbb{R}} \max_{i \in S_r} \left| \widehat{B}_i + \lambda \right|.$$

Solution: centre the range of the $\widehat{B}_i$

Only the extremes matter, so minimise $\max\{|\widehat{B}_{\min} + \lambda|, |\widehat{B}_{\max} + \lambda|\}$:

$$\widehat{\lambda}_{\mathrm{CB}} = -\tfrac{1}{2}\left(\widehat{B}_{\min} + \widehat{B}_{\max}\right),$$

$$\widehat{t}_{I,\mathrm{CB}} = \widehat{t}_{I,\mathrm{WLS}} - \tfrac{1}{2}\left(\min_{i \in S_r} \widehat{B}_i + \max_{i \in S_r} \widehat{B}_i\right)$$

- If the extremes are balanced ($\widehat{B}_{\min} \approx -\widehat{B}_{\max}$) the correction is ≈ 0: nothing happens unless influence is *one-sided*.
- Equivalent tuning-constant view: $\lambda(c) = \sum_{i \in S_r} \{\psi_c^B(\widehat{B}_i) - \widehat{B}_i\}$ with a two-sided Huber ψ_c^B applied to the $\widehat{B}_i$; the cut-off c_{CB} solving $\lambda(c_{\mathrm{CB}}) = \widehat{\lambda}_{\mathrm{CB}}$ exists (take the largest if not unique) but never needs to be computed.

Approach 1: properties, pros and cons

Consistency

Suppose $\widehat{t}_{l,\mathrm{WLS}} - t_y = O_p(N/\sqrt{n})$ and $\max_{i \in s_r} |\widehat{B}_i| = O_p(N/n)$. Then $\widehat{\lambda}_{\mathrm{CB}} = O_p(N/n) = o_p(N/\sqrt{n})$, so

$$\frac{\sqrt{n}}{N}\left(\widehat{t}_{l,\mathrm{CB}} - \widehat{t}_{l,\mathrm{WLS}}\right) \xrightarrow{p} 0 :$$

$\widehat{t}_{l,\mathrm{CB}}$ inherits the consistency of $\widehat{t}_{l,\mathrm{WLS}}$ (first-order asymptotic equivalence).

Finite samples

The adjustment can be appreciable when one or a few respondents have large $|\widehat{B}_i|$: fewer wild swings, a better bias–variance trade-off.

Pros

- Closed form; no tuning constant, grid or iteration.
- Uses the unit's *full* influence (weight, residual, leverage, response rate).
- Strongest bias control among the adaptive methods.

Cons

- Sample-level correction: not naturally an observed-plus-imputed file (fixable by reverse calibration).
- Corrects only *one-sided* influence; modest variance reduction under symmetric heavy tails.
- Relies on the first-order approximation behind $\widehat{B}_i$.

Approach 2: a new adaptive tuning constant c_{new}

Standardise the conditional biases

$$\widehat{Z}_i = \frac{\widehat{B}_i - \overline{B}_r}{s_B}, \qquad i \in S_r,$$

$\overline{B}_r = \frac{1}{n_r}\sum_{i\in S_r}\widehat{B}_i$, $s_B^2 = \frac{1}{n_r - 1}\sum_{i\in S_r}(\widehat{B}_i - \overline{B}_r)^2$.

Measure their asymmetry

$$\widehat{A}_B = \left| \frac{\min_{i\in S_r}\widehat{Z}_i + \max_{i\in S_r}\widehat{Z}_i}{2} \right|$$

= absolute midpoint of the range of the $\widehat{Z}_i$: how unbalanced the most negative and most positive influences are around zero. Invariant to location and scale of the $\widehat{B}_i$.

The tuning constant

$$c_{\text{new}} = 1.345\big(1 + \widehat{A}_B\big) + \frac{n^{3/2}}{N}$$

By construction $c_{\text{new}} \geq 1.345$: it only ever *relaxes* the textbook clipping.

The estimator (projection form)

$$\widehat{t}_{l,R}^{\,\text{proj}}(c_{\text{new}}) = \sum_{i\in S} w_i \mathbf{v}_i^\top \widehat{\mathbf{B}}_R(c_{\text{new}}),$$

with $\widehat{\mathbf{B}}_R(c_{\text{new}})$ the Huber M-estimator at $c = c_{\text{new}}$.

Approach 2: rationale and large-sample behaviour

Component 1: $1.345(1+\widehat{A}_B)$ **— shape of the influence**

- Balanced extremes ($\min \widehat{Z}_i + \max \widehat{Z}_i \approx 0$) and small $n^{3/2}/N \Rightarrow c_{\mathrm{new}} \approx 1.345$: the usual degree of robustness.
- One-sided influence $\Rightarrow \widehat{A}_B$ appreciable $\Rightarrow$ larger $c \Rightarrow$ less clipping, exactly when clipping a group of representative outliers would bias the total.

Component 2: $n^{3/2}/N$ **— sample size**

If $n^{3/2}/N \to \infty$ (e.g. $n/N \to f \in (0,1]$) then $c_{\mathrm{new}} \to \infty$ and, provided $\|\widehat{\mathbf{B}}_R(c_{\mathrm{new}}) - \widehat{\mathbf{B}}_{\mathrm{WLS}}\| = o_p(1)$,

$$\tfrac{1}{N}\{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}}) - \widehat{t}_{l,\mathrm{WLS}}\} = \Big\{\tfrac{1}{N}\textstyle\sum_{i\in S} w_i \mathbf{v}_i^{\top}\Big\}\{\widehat{\mathbf{B}}_R(c_{\mathrm{new}}) - \widehat{\mathbf{B}}_{\mathrm{WLS}}\} = o_p(1),$$

so the estimator inherits the consistency of $\widehat{t}_{l,\mathrm{WLS}}$.

What the two components buy

- Symmetric heavy tails: c_{new} stays moderate (medians 2–3 in the simulations) $\Rightarrow$ real variance reduction.
- Asymmetric contamination: c_{new} grows with the imbalance and with $n \Rightarrow$ bias fades as the sample becomes informative.

Practical note

For sample sizes met in practice c_{new} remains moderate, so large residuals *are* still downweighted: a finite-sample gain, obtained with standard weighted robust-regression software.

Projection form versus observed-plus-imputed form

Observed-plus-imputed (conventional)

$$\widehat{t}_{l,R}^{\,\mathrm{imp}}(c_{\mathrm{new}}) = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \mathbf{v}_i^{\top} \widehat{\mathbf{B}}_R(c_{\mathrm{new}})$$

Robustification affects only the predictions for nonrespondents; the direct contributions $w_i y_i$ of influential respondents are untouched — they keep their full impact on the total.

Projection (used in approaches 2 and 3)

$$\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}}) = \sum_{i \in S} w_i \mathbf{v}_i^{\top} \widehat{\mathbf{B}}_R(c_{\mathrm{new}})$$

All respondent information enters through the robust fit: large residuals are treated consistently for the observed *and* the missing part. It is a robust analogue of $\widehat{t}_{l,\mathrm{WLS}} = \sum_{i \in S} w_i \mathbf{v}_i^{\top} \widehat{\mathbf{B}}_{\mathrm{WLS}}$, not a rewriting of $\widehat{t}_{l,R}^{\,\mathrm{imp}}(c_{\mathrm{new}})$.

- Consequence: $\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}})$ is not a conventional imputed estimator in which observed values are retained unchanged.
- For data dissemination, a completed file reproducing $\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}})$ exactly can be built by *reverse calibration* (Chambers & Ren 2004) — Section 11.

Approach 2: pros and cons

Pros

- Explicit formula; computed *before* fitting; one robust regression with a user-specified constant.
- Adapts to both the *shape* of influence and the *sample size*; consistent under $n^{3/2}/N \to \infty$.
- Best or near-best relative efficiency under heavy right skew (Pareto, Fréchet), close to the infeasible oracle.
- Never over-robustifies ($c_{\mathrm{new}} \geq 1.345$).

Cons

- Cannot go below 1.345: some efficiency left on the table under symmetric heavy tails where a smaller c would be better.
- Bias under strong asymmetric contamination is reduced, not eliminated, at small n (e.g. -5% to -7% for $n \leq 100$ under Pareto/Fréchet).
- The $n^{3/2}/N$ schedule is a pragmatic choice; projection form requires reverse calibration for file release.
- MSE estimation must account for the data-driven c_{new} (Section 8).

Approach 3: choose c by minimising an estimated design MSE

Framework: design-based, conditional on $\mathcal{F}_U = \sigma\{(y_i, \mathbf{v}_i, r_i) : i \in U\}$

Only $I_1, \ldots, I_N$ are random ($\mathbb{E}_p$, $\mathbb{V}_p$); c and σ are treated as fixed ($\hat{\sigma}$ plugged in afterwards);
$\mathrm{MSE}_p\{\widehat{t}_{I,R}^{\,\mathrm{proj}}(c)\} = \mathbb{V}_p\{\widehat{t}_{I,R}^{\,\mathrm{proj}}(c)\} + \big[\mathbb{E}_p\{\widehat{t}_{I,R}^{\,\mathrm{proj}}(c)\} - t_y\big]^2$.

Linearisation of the robust projection estimator

Score $\widehat{\mathbf{U}}_N(\boldsymbol{\beta}; c) = \frac{1}{N}\sum_{i\in S} w_i r_i \psi_c\big(\frac{y_i - \mathbf{v}_i^\top \boldsymbol{\beta}}{\sigma \phi_i^{1/2}}\big)\frac{\mathbf{v}_i}{\phi_i^{1/2}}$; $\boldsymbol{\beta}_c^*$ = design probability limit of $\widehat{\mathbf{B}}_R(c)$; $\overline{\mathbf{v}}_U = \frac{1}{N}\sum_{k\in U}\mathbf{v}_k$. Then

$\widehat{t}_{I,R}^{\,\mathrm{proj}}(c) = \sum_{i\in S} w_i \eta_i(c) + o_p(N/\sqrt{n})$ with the linearised variable and Jacobian

$$\eta_i(c) = \mathbf{v}_i^\top \boldsymbol{\beta}_c^* - \overline{\mathbf{v}}_U^\top \mathbf{J}_c^{-1}\, r_i\, \psi_c\Big(\frac{y_i - \mathbf{v}_i^\top \boldsymbol{\beta}_c^*}{\sigma\phi_i^{1/2}}\Big)\frac{\mathbf{v}_i}{\phi_i^{1/2}},$$

$$\mathbf{J}_c = \frac{\partial\, \mathbb{E}_p\{\widehat{\mathbf{U}}_N(\boldsymbol{\beta}; c)\}}{\partial \boldsymbol{\beta}^\top}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_c^*} = -\frac{1}{N}\sum_{i\in U} r_i \mathbb{1}\left\{\left|\frac{y_i - \mathbf{v}_i^\top \boldsymbol{\beta}_c^*}{\sigma\phi_i^{1/2}}\right| \le c\right\}\frac{\mathbf{v}_i\mathbf{v}_i^\top}{\sigma\phi_i} \qquad \text{(Huber score)},$$

since $\psi_c'(t) = \mathbb{1}\{|t| \le c\}$ (no standardised residual exactly at $\pm c$): only the *un-clipped* respondents contribute to $\mathbf{J}_c$.

Design MSE approximation

$$\mathrm{MSE}_p\{\widehat{t}_{I,R}^{\,\mathrm{proj}}(c)\} \approx \Big\{\sum_{i\in U}\eta_i(c) - t_y\Big\}^2 + \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\frac{\eta_i(c)}{\pi_i}\frac{\eta_j(c)}{\pi_j}.$$

Approach 3: the MSE estimator

Variance part (plug-in: $\widehat{\mathbf{B}}_R(c)$, $\widehat{\sigma}$, $\widehat{\mathbf{J}}_c$, $\widehat{\bar{\mathbf{v}}}_U = N^{-1}\sum_{k\in S} w_k \mathbf{v}_k$)

$$\widehat{\mathbb{V}}\{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c)\} = \sum_{i\in S}\sum_{j\in S} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \frac{\widehat{\eta}_i(c)}{\pi_i} \frac{\widehat{\eta}_j(c)}{\pi_j}.$$

Squared-bias part: use $\widehat{t}_{l,\mathrm{WLS}}$ as an (approximately design-unbiased) reference

$D(c) = \widehat{t}_{l,R}^{\,\mathrm{proj}}(c) - \widehat{t}_{l,\mathrm{WLS}}$ satisfies $\mathbb{E}_p(D^2) = \mathbb{V}_p(D) + \{\mathbb{E}_p(D)\}^2$, so a non-negative estimator of the squared design bias is $\max\left[0,\ D(c)^2 - \widehat{\mathbb{V}}\{D(c)\}\right]$, with

$$\widehat{\mathbb{V}}\{D(c)\} = \sum_{i\in S}\sum_{j\in S} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}} \frac{\widehat{\eta}_i(c)-\widehat{\tau}_i}{\pi_i} \frac{\widehat{\eta}_j(c)-\widehat{\tau}_j}{\pi_j}, \quad \widehat{\tau}_i = r_i y_i + (1-r_i)\mathbf{v}_i^\top \widehat{\mathbf{B}}_{\mathrm{WLS}} + \widehat{\mathbf{A}} r_i \mathbf{v}_i \phi_i^{-1}(y_i - \mathbf{v}_i^\top \widehat{\mathbf{B}}_{\mathrm{WLS}}).$$

$$\widehat{\mathrm{MSE}}\{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c)\} = \widehat{\mathbb{V}}\{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c)\} + \max\left[0, \{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c) - \widehat{t}_{l,\mathrm{WLS}}\}^2 - \widehat{\mathbb{V}}\{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c) - \widehat{t}_{l,\mathrm{WLS}}\}\right]$$

Approach 3: selection over a grid, expected behaviour, pros and cons

Selected tuning constant and estimator

$$\widehat{c}_{\mathrm{MSE}} = \arg\min_{c \in \mathcal{C}} \widehat{\mathrm{MSE}}\{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c)\},$$

$$\widehat{t}_{l,R}^{\,\mathrm{proj}}(\widehat{c}_{\mathrm{MSE}}) = \textstyle\sum_{i \in S} w_i \mathbf{v}_i^{\top} \widehat{\mathbf{B}}_R(\widehat{c}_{\mathrm{MSE}}),$$

$\mathcal{C}$ a prespecified finite grid (ties broken by a prespecified rule); in the simulations $\widehat{c}_{\mathrm{MSE}}$ is written c^*.

Expected behaviour

Under asymmetric contamination the squared bias of a small c is persistent while the variance saved shrinks with $n \Rightarrow$ if the grid contains large values, $\widehat{c}_{\mathrm{MSE}}$ increases with n and the estimator moves towards WLS. Under symmetric heavy tails a finite c stays optimal.

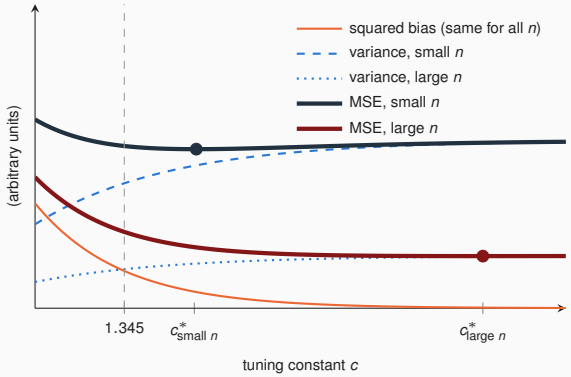
Pros

- Targets the right criterion (MSE of the total) directly; no floor at 1.345.
- Best feasible method under symmetric heavy tails (Student, Laplace).
- Comes with an MSE estimate for free.

Cons

- One robust fit per grid point (moderate cost; needs π_{ij}).
- Relies on $\widehat{t}_{l,\mathrm{WLS}}$ as a design-unbiased reference and on first-order approximations; the selection step adds variability that the plug-in MSE ignores.
- Mixed results under strong mixture contamination at small n.

Why the optimal degree of robustness moves with n (schematic)



Curves are stylised (bias$^2 \propto e^{-c/1.3}$, variance $\propto \{0.5 + 0.5(1 - e^{-c/2})\}/n$); they are not simulation output.

Illustrative, asymmetric contamination

- Squared bias *decreases* in c (less clipping) and does not depend on n.
- Variance *increases* in c (towards the WLS variance) and scales like $1/n$.
- The MSE minimiser moves to the right as n grows: adaptive $c \approx$ "robust when small, WLS when large".
- At $c = 1.345$ the large-n MSE is far above its minimum — the RE > 100 phenomenon.

The three approaches side by side

	1. Minimax CB $\widehat{t}_{l,\mathrm{CB}}$	**2.** c_{new} $\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}})$	**3.** $\widehat{c}_{\mathrm{MSE}}$ $\widehat{t}_{l,R}^{\,\mathrm{proj}}(\widehat{c}_{\mathrm{MSE}})$
What adapts	an additive correction $\widehat{\lambda}_{\mathrm{CB}}$	the Huber constant, from asymmetry of $\widehat{B}_i$ and $n^{3/2}/N$	the Huber constant, from a design-MSE estimate
Inputs	$\widehat{B}_i$	$\widehat{B}_i$, n, N	π_{ij}, grid $\mathcal{C}$, $\widehat{t}_{l,\mathrm{WLS}}$
Form	WLS + correction	projection	projection
Range of robustness	correction $\to 0$ if balanced	$c \geq 1.345$	any $c \in \mathcal{C}$
Computation	closed form	one robust fit	one robust fit per grid point
Bias control	strongest	good; improves with n	good; improves with n
Variance reduction	modest	large under right skew	large under symmetric heavy tails
Large-sample target	$\widehat{t}_{l,\mathrm{WLS}}$	$\widehat{t}_{l,\mathrm{WLS}}$ (if $n^{3/2}/N \to \infty$)	$\widehat{t}_{l,\mathrm{WLS}}$ (if grid contains large c)

No procedure dominates

They offer different compromises between bias control and variance reduction; the simulations quantify them.

Mean squared error estimation

MSE estimation when the tuning constant is data-driven

Plug-in linearisation estimator for $\widehat{t}_{l,R}^{\,\mathrm{proj}}(c^*)$

Evaluate the fixed-c MSE estimator at the selected constant:

$$\widehat{\mathbb{V}}\{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c^*)\} + \max\left[0, D(c^*)^2 - \widehat{\mathbb{V}}\{D(c^*)\}\right],$$

$$D(c^*) = \widehat{t}_{l,R}^{\,\mathrm{proj}}(c^*) - \widehat{t}_{l,\mathrm{WLS}}.$$

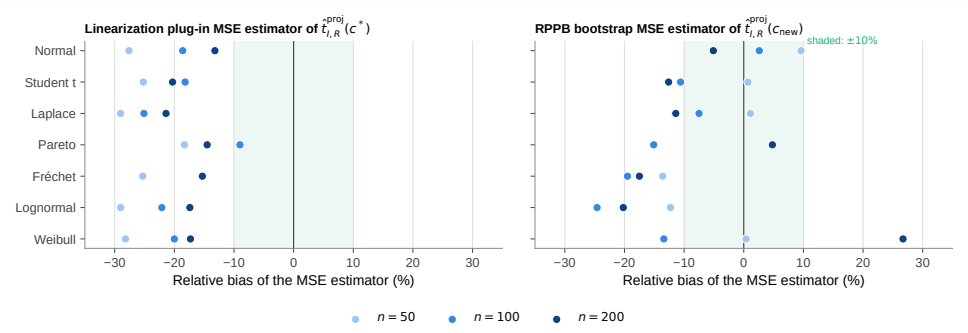
Limitation

The linearisation treats c^* as fixed; c^* varies from sample to sample, so the estimator ignores the *selection variability* and any selection effect. Expect under-estimation.

Reweighted pseudo-population bootstrap (RPPB) for $\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}})$

- Pseudo-population bootstrap (Mashreghi, Haziza & Léger 2016): replicate sample units into a synthetic population, resample with the original design, redo *everything* — including c_{new} — in each replicate.
- New twist: replication weights $w_i^* \propto d_i\, h_i(c_{\mathrm{new}})$ with Huber weights $h_i = \psi_{c_{\mathrm{new}}}(\widetilde{e}_i)/\widetilde{e}_i \leq 1$, so a downweighted influential respondent is *not* mechanically copied N/n times; the removed mass is redistributed ($\widehat{\gamma}$) so that $\sum_i w_i^* = N$.
- Bootstrap-assisted MSE: $\widehat{V}_{B,R}^* + \max[0, \{\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}}) - \widehat{t}_{l,\mathrm{WLS}}\}^2 - \widehat{V}_{B,\Delta}^*]$.

How well do the MSE estimators work? (Monte Carlo relative bias)



Plug-in at c^*

Negative bias everywhere: -9% to -29%. Consistent with ignoring the variability of the selected constant.

RPPB for c_{new}

Good under symmetric errors (within $\pm 13\%$, often within $\pm 5\%$); adequate but uneven under asymmetric errors (-25% to $+27\%$). Room for improvement — an open problem.

Simulation study

Design of the Monte Carlo study

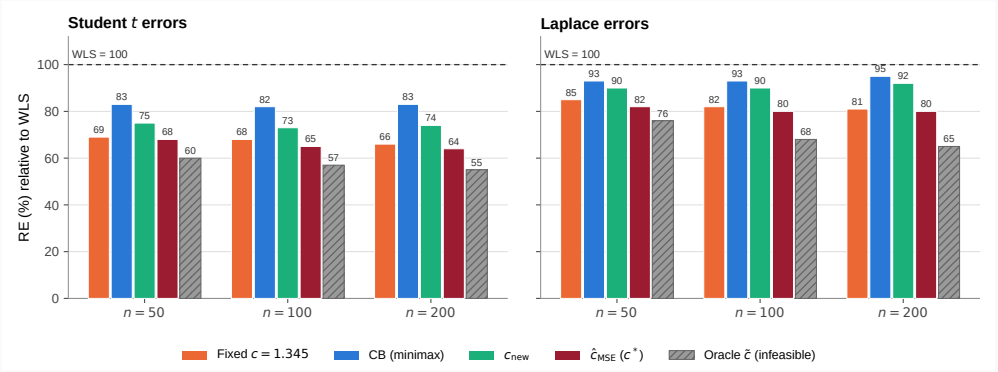
Populations ($R = 10{,}000$ replications, $N = 5{,}000$)

- $V \sim \mathrm{Gamma}(\text{shape } 5, \text{scale } 10)$;
  $y_i \mid v_i \sim \mathcal{D}(\mu_i; \sigma^2)$, $\mu_i = \beta_0 + \beta_1 v_i$,
  $(\beta_0, \beta_1, \sigma) = (50, 12, 600)$.
- Seven error laws with identical first two moments: Normal, Student t, Laplace (symmetric); Pareto, Fréchet, Lognormal, Weibull (asymmetric).
- Mixtures with $\varepsilon \in \{1\%, 3\%\}$ outliers, $\Delta_i \sim \mathrm{Bernoulli}(\varepsilon)$ independent of V: normal: $\mu_{1i} = 150 + 20 v_i$, $\mu_{2i} = 2000 + 85 v_i$, $\sigma_1 = 400$, $\sigma_2 = 2000$; lognormal: $\mu_{1i} = 150 + 8 v_i$, $\mu_{2i} = 1200 + 60 v_i$, $\sigma_1 = 150$, $\sigma_2 = 1500$.

Sampling, nonresponse, estimators

- SRSWOR, $n \in \{50, 100, 200\}$.
- MAR response probability $p_i = 0.1 + 0.9 \dfrac{\exp(4 - 0.09 v_i)}{1 + \exp(4 - 0.09 v_i)}$: $\approx 50\%$ overall, lower for large v_i.
- Compared: $\widehat{t}_{I,\mathrm{WLS}}$; fixed $\widehat{t}_{I,R}(1.345)$; $\widehat{t}_{I,\mathrm{CB}}$; $\widehat{t}^{\,\mathrm{proj}}_{I,R}(c_{\mathrm{new}})$; $\widehat{t}^{\,\mathrm{proj}}_{I,R}(c^*)$ ($c^* = \widehat{c}_{\mathrm{MSE}}$); and the *infeasible oracle* $\widehat{t}_{I,R}(\widetilde{c})$ with $\widetilde{c}$ minimising the Monte Carlo MSE.
- Measures: RB and RE (WLS $= 100$), as before.

Symmetric errors: relative efficiency (RB ≈ 0 for all estimators)



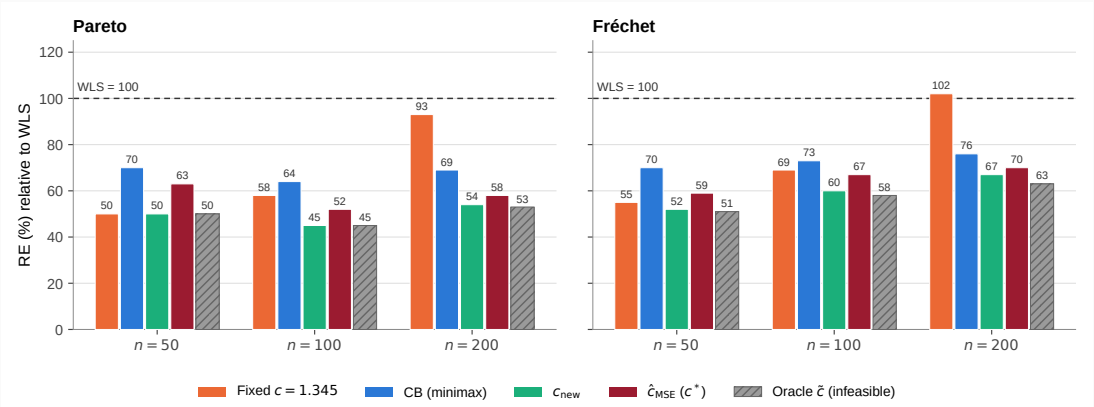
Normal errors (not shown)

RE between 100 and 106 for every robust procedure: a *modest insurance premium*; the CB estimator is closest to WLS (100–101).

Student *t* and Laplace

All robust procedures reduce MSE. Largest feasible gains for c^* (RE 64–68 Student, 80–82 Laplace), then fixed c, then c_{new}, then CB; the oracle reaches 55–76.

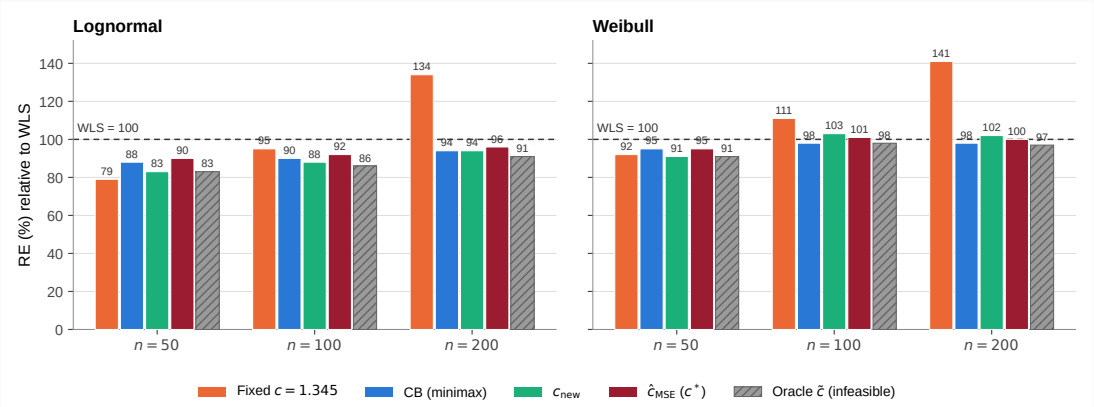
Asymmetric errors (I): relative efficiency under heavy right skew



Bottom line

Fixed $c = 1.345$ deteriorates as n grows (RE 93 and 102 at $n = 200$). c_{new} is best or near-best (Pareto 45–54, Fréchet 52–67), essentially at the oracle; c^* is close behind; CB gains less (64–76) but never loses.

Asymmetric errors (II): relative efficiency under moderate skew

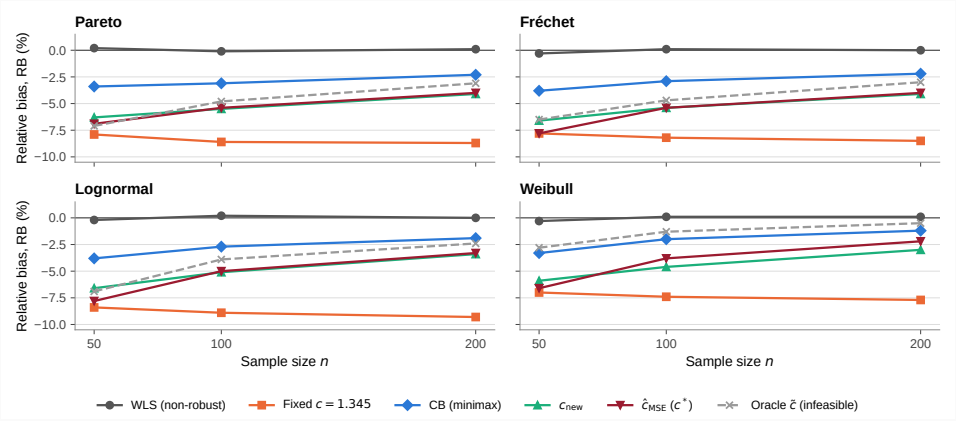


Bottom line

Lognormal: all adaptive procedures beat WLS at every n (83–96), while fixed c reaches 134 at $n = 200$.

Weibull: the adaptive estimators are broadly comparable to WLS (91–103), while fixed c deteriorates to 141.

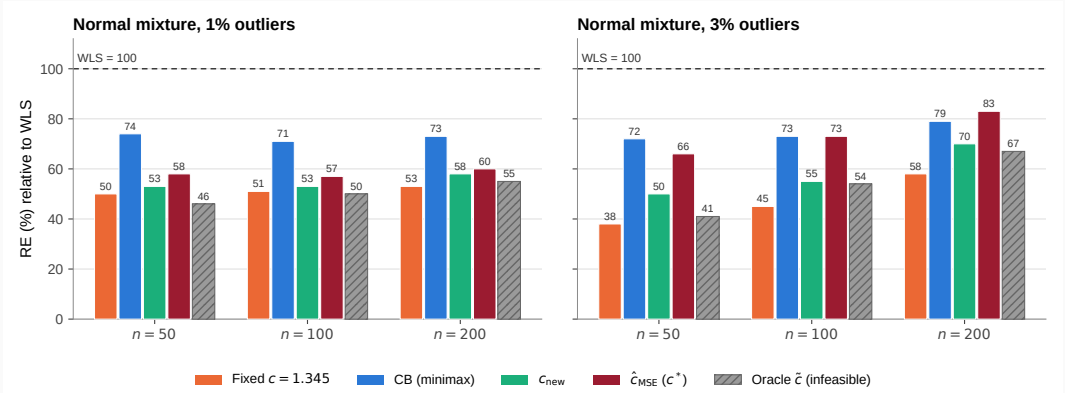
Asymmetric errors: relative bias shrinks with n — except for fixed c



Reading the lines

Fixed c: flat at -7% to -9%. c_{new} and c^*: from -6% to -8% at $n = 50$ down to -2% to -4% at $n = 200$.
CB: smallest bias throughout (-1.2% to -3.8%).

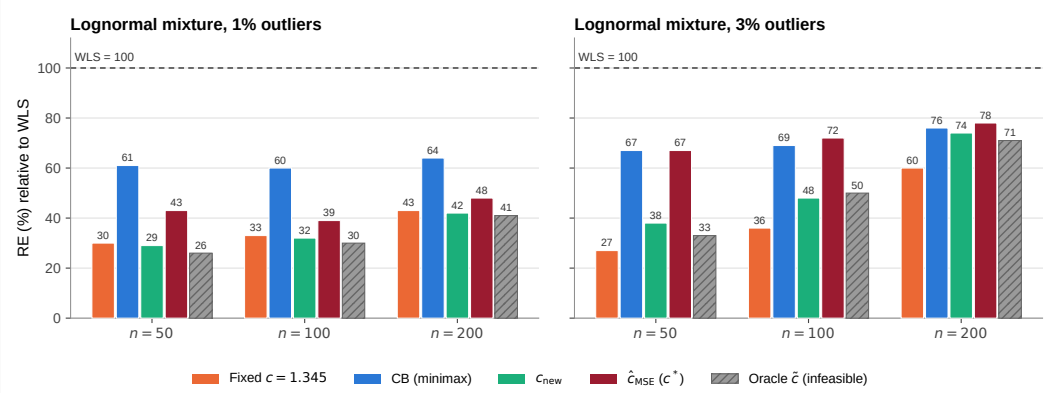
Mixture distributions (I): normal mixtures, 1% and 3% outliers



Bottom line

Every robust procedure reduces MSE substantially (RE 38–83). Fixed c has the smallest MSE in most normal-mixture settings; CB is the least biased (-1.3% to -3.2%) but gains the least; c_{new} and c^* sit in between, with biases of -2% to -8%.

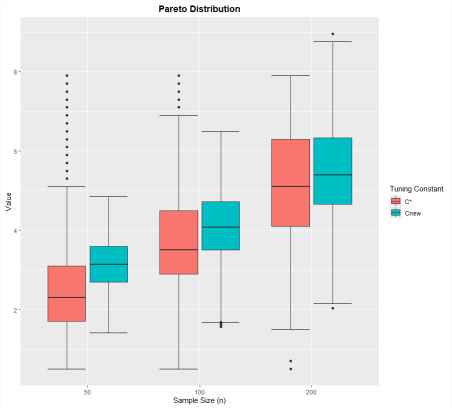
Mixture distributions (II): lognormal mixtures, 1% and 3% outliers

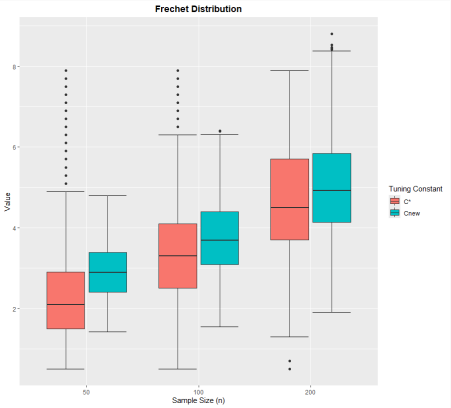


Bottom line

RE 27–78: the largest gains of the whole study. c_{new} is slightly better than fixed c with 1% outliers, fixed c slightly better with 3%; CB again has the smallest bias (-2% to -5%) and the most modest gains. No single procedure dominates across the mixture settings.

Behaviour of the adaptive tuning constants (I): right-skewed errors

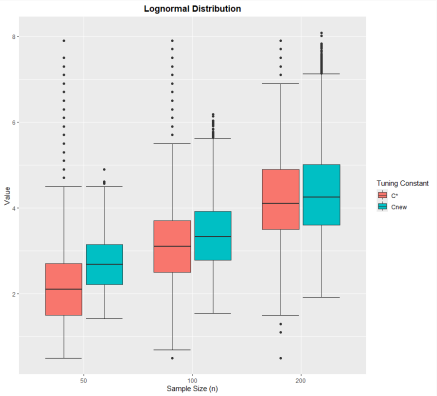


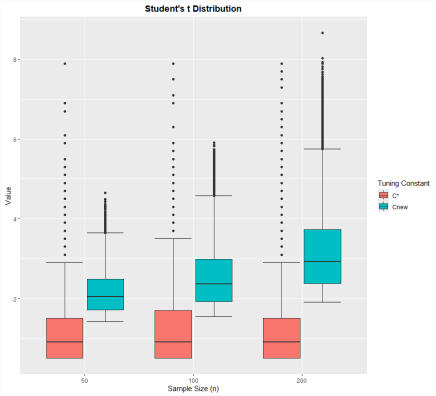


Empirical distribution of c^* (red) and c_{new} (teal) over 10,000 replications; textbook value 1.345.

- Both constants increase with n (medians $\approx$ 2–3 at $n = 50$, ≈ 5 at $n = 200$) and exceed 1.345 in most samples: the data ask for less clipping as the sample grows.

Behaviour of the adaptive tuning constants (II): lognormal versus Student t





- Lognormal: same pattern as Pareto and Fréchet.
- **Student** t (symmetric, heavy-tailed): c^* stays *below* 1.345 (median ≈ 1) at every n — stronger clipping is optimal and c^* finds it — whereas c_{new} sits at $\approx$ 2–3 because of its floor: exactly the efficiency gap seen in the symmetric results.

Simulation summary

Setting	$\widehat{t}_{l,\mathrm{WLS}}$	Fixed c	CB	c_{new}	c^*
Normal	ref.	$\approx$ ref.	$\approx$ ref.	$\approx$ ref.	$+3$–6%
Student t / Laplace	ref.	good	modest	good	best
Pareto / Fréchet	ref.	fails for large n	good	best ($\approx$ oracle)	very good
Lognormal / Weibull	ref.	fails for large n	$\approx$ ref., least bias	$\approx$ ref.	$\approx$ ref.
Mixtures (1–3%)	ref.	best or close	least bias	good	good
Bias	0	constant in n	smallest	$\downarrow$ in n	$\downarrow$ in n

Three messages

1. A fixed tuning constant is a bet on symmetry; it loses badly, and increasingly with n, under right skew.
2. Adaptive constants trade a little efficiency under symmetry for protection under asymmetry: c_{new} for skewed data, c^* for symmetric heavy tails, CB when bias control is the priority.
3. The gains are finite-sample gains; all adaptive estimators approach WLS as the sample becomes informative.

Application: 2023 Medical Expenditure Panel Survey

Data, variables and design of the illustration

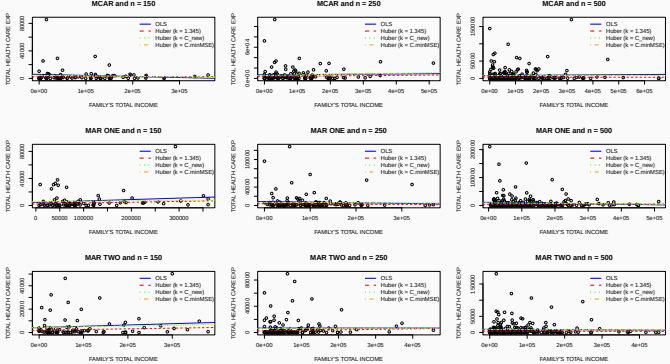
Data

- MEPS 2023 Full-Year Consolidated File (HC-251), AHRQ.
- After removing records with missing values on the variables of interest: $N = 18{,}919$ individuals, treated as the finite population (so the true total is known).
- Study variable y: total annual health-care expenditure, `TOTEXP23`.
- Auxiliary variable x: total family income, `FAMINC23`.
- Both highly skewed, with influential observations.

Mimicking an incomplete survey

- SRSWOR samples of size $n \in \{150, 250, 500\}$.
- Three response mechanisms, each $\approx 50\%$ response: **MCAR** $p_i = 0.5$; **MAR ONE** (low income responds less) $p_i = \dfrac{\exp\{-3 + |x_i|^{1/10}\}}{1 + \exp\{-3 + |x_i|^{1/10}\}}$; **MAR TWO** (high income responds less) $p_i = \dfrac{\exp\{3 - |x_i|^{1/10}\}}{1 + \exp\{3 - |x_i|^{1/10}\}}$.
- Estimators: WLS; fixed $c = 1.345$ (observed-plus-imputed); c_{new}; CB; $\widehat{c}_{\mathrm{MSE}}$.
- Criterion: relative error $\mathrm{RelErr}(\widehat{T}) = (\widehat{T} - T)/T \times 100\%$ on the *selected* samples.

Respondents' data and fitted lines



Total annual health-care expenditure versus family income for the respondents, by response mechanism (rows) and sample size (columns). Solid blue: OLS; dashed/dotted: Huber fits with $c = 1.345$, c_{new} and $\hat{c}_{\mathrm{MSE}}$.

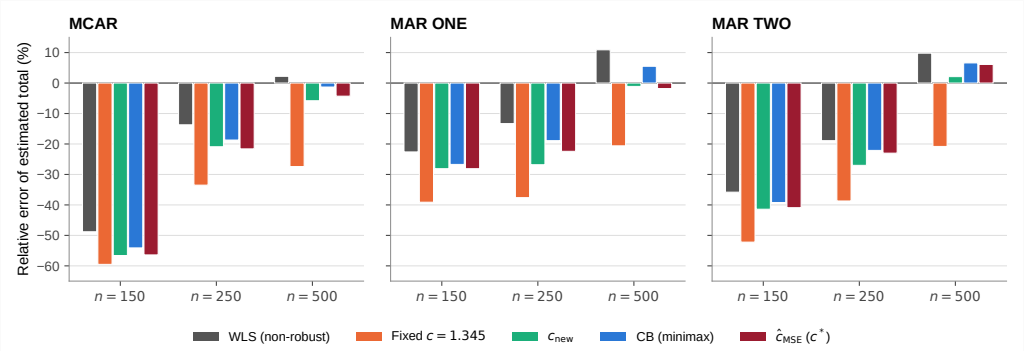
- Pronounced right skew and several influential observations in every panel; the OLS line is the one that moves, the Huber lines are stable — but stability is not the same as accuracy for the total.

Relative error (%) of the estimated total of TOTEXP23

Scenario	n	WLS	Fixed-c	c_{new}	CB	$\widehat{c}_{\text{MSE}}$
MCAR	150	−48.8	−59.5	−56.6	−54.1	−56.4
	250	−13.7	−33.5	−20.9	−18.7	−21.6
	500	2.2	−27.4	−5.8	**−1.3**	−4.3
MAR ONE	150	−22.6	−39.1	−28.1	−26.7	−28.1
	250	−13.3	−37.6	−26.8	−18.9	−22.4
	500	10.9	−20.6	**−1.1**	5.5	−1.8
MAR TWO	150	−35.8	−52.2	−41.4	−39.2	−40.9
	250	−18.9	−38.7	−27.0	−22.1	−23.0
	500	9.8	−20.8	**2.1**	6.6	6.1

- In all nine scenarios the fixed-c estimator has the largest absolute error (-21% to -60%): systematic under-estimation of total spending.
- The adaptive procedures substantially reduce this error; for $n = 500$ one of them is the most accurate under each mechanism (bold). For $n = 150$ and 250, WLS remains closest in these particular samples.

The same results as a picture



Reading the chart — and a caveat

Orange (fixed c) is always the longest negative bar; the adaptive estimators (green, blue, crimson) shrink towards zero as n grows and are closest to the truth at $n = 500$. The table is based on *single* sample realisations: it shows the direction of the effects on real data, not a performance ranking — that is what the Monte Carlo study is for.

Conclusions, practical guidance and future work

Conclusions

1. Influential respondents are a variance problem for regression imputation; robustification converts it into a bias problem. The right amount of robustness is a finite-sample bias–variance trade-off that depends on n and on how one-sided the influence is.
2. With a fixed tuning constant, the bias is persistent under asymmetric contamination (mixture-model formula), so the relative efficiency against WLS *increases without bound* with n — confirmed in simulations and on MEPS.
3. The conditional bias of a respondent — sampling plus nonresponse/imputation influence — is an estimable, interpretable, one-number measure of influence for imputed estimators.
4. Three data-adaptive procedures (minimax CB; c_{new}; $\widehat{c}_{\mathrm{MSE}}$), two of them new, give substantial MSE reductions under heavy-tailed, asymmetric and mixture distributions, with only modest losses under normal errors; each inherits the consistency of WLS.
5. The procedures offer different compromises: CB for bias control, c_{new} for right-skewed data, $\widehat{c}_{\mathrm{MSE}}$ for symmetric heavy tails. No procedure dominates uniformly.

A workflow

1. Fit WLS on the respondents; compute $\widehat{B}_i$, $i \in S_r$ (needs π_i, π_{ij}).
2. Inspect the standardised $\widehat{Z}_i$: balanced extremes $\Rightarrow$ symmetric problem; one-sided extremes $\Rightarrow$ any fixed c will bias the total.
3. Right-skewed study variable (expenditures, revenue): use c_{new} — one weighted Huber fit with $k = c_{\mathrm{new}}$ in standard software.
4. Symmetric heavy tails, or when π_{ij} and computing time are available: $\widehat{c}_{\mathrm{MSE}}$ over a grid, with its MSE estimate.
5. Bias is the main concern (official totals, benchmarks): $\widehat{t}_{l,\mathrm{CB}}$, a closed-form correction of $\widehat{t}_{l,\mathrm{WLS}}$.

Releasing a rectangular file: reverse calibration (Chambers & Ren 2004)

Target $\widehat{t}_{\mathcal{R}}$ (any of the three); preliminary imputations $y_i^{(0)}$ (WLS for CB, robust fitted values for the projection estimators). Choose final y_{iF}^* close to $y_i^{(0)}$ subject to $\sum_{i\in S_r} w_i y_i + \sum_{i\in S_m} w_i y_{iF}^* = \widehat{t}_{\mathcal{R}}$:

$$y_{iF}^* = y_i^{(0)} + \frac{q_i w_i}{\sum_{j\in S_m} q_j w_j^2}\left\{\widehat{t}_{\mathcal{R}} - \widehat{t}_l^{(0)}\right\}, \quad i \in S_m.$$

Observed values unchanged; the usual weighted estimator on the completed file reproduces $\widehat{t}_{\mathcal{R}}$ exactly. Cost of the procedures: one WLS plus one robust fit (CB, c_{new}) or one robust fit per grid point ($\widehat{c}_{\mathrm{MSE}}$).

Limitations and future research directions

Limitations of the present work

- Linear regression imputation; influence mainly through large standardised residuals.
- SRSWOR in the simulations and the application; RPPB presented for SRSWOR.
- MSE estimation does not yet fully account for the sample-to-sample variability of the selected constant (plug-in under-estimates by 9–29%).
- MEPS illustration based on single realised samples.

Future research

- Generalised linear and nonparametric / machine-learning imputation with an adaptive robust loss.
- Explicit control of influence from extreme *leverage* and large *sampling weights* (GM-type weights built from $\widehat{B}_i$).
- Unequal-probability and multistage designs; calibration weights; multiple imputation.
- MSE estimation that incorporates the selection step (bootstrap of the grid search; selective-inference ideas).
- Alternative schedules to $n^{3/2}/N$; joint selection of c and the scale; scalability.

Thank you!

Questions?

Sixia Chen

Department of Biostatistics and Epidemiology, Hudson College of Public Health
University of Oklahoma Health Campus, Oklahoma City, OK, USA

sixia-chen@ou.edu

Joint work with David Haziza and Jia Ning Zhang (University of Ottawa)

Paper: *Robust imputation procedures in the presence of influential survey data based on adaptative tuning constants*

References i

Agency for Healthcare Research and Quality. *Medical Expenditure Panel Survey, 2023 Full-Year Consolidated Data File (HC-251).* Rockville, MD.

Andersen, R. (2008). *Modern Methods for Robust Regression*. Sage.

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika* 100, 555–569.

Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers in survey data. In *Handbook of Statistics, Vol. 29A: Sample Surveys: Design, Methods and Applications* (D. Pfeffermann and C. R. Rao, eds.), 247–279. Elsevier.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association* 81, 1063–1069.

Chambers, R. L. and Ren, R. (2004). Outlier robust imputation of survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 3336–3344.

Chatterjee, S. and Hadi, A. S. (1988). Impact of simultaneous omission of a variable and an observation on a linear regression equation. *Computational Statistics & Data Analysis* 6, 129–144.

Chen, S. and Xu, C. (2023). Handling high-dimensional data with missing values by modern machine learning techniques. *Journal of Applied Statistics* 50, 786–804.

Chen, S., Haziza, D. and Michal, V. (2024). Efficient multiply robust imputation in the presence of influential units in surveys. *Canadian Journal of Statistics* 52, 829–854.

Chen, S., Haziza, D. and Zhang, J. N. Robust imputation procedures in the presence of influential survey data based on adaptative tuning constants. Manuscript and Supplementary Material.

Favre-Martinoz, C., Haziza, D. and Beaumont, J.-F. (2016). Robust inference in two-phase sampling designs with application to unit nonresponse. *Scandinavian Journal of Statistics* 43, 1019–1034.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69, 383–393.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Vol. 29A* (D. Pfeffermann and C. R. Rao, eds.), 215–246. Elsevier.

References ii

Hernandez-Viver, A. and Mitchell, E. M. (2025). Concentration of healthcare expenditures and selected characteristics of people with high expenses, United States civilian noninstitutionalized population, 2018–2022. *MEPS Statistical Brief #560*. Agency for Healthcare Research and Quality, Rockville, MD.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663–685.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73–101.

Kim, J. K. and Shao, J. (2021). *Statistical Methods for Handling Incomplete Data* (2nd ed.). Chapman & Hall/CRC.

Kokic, P. N. and Bell, P. A. (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics* 10, 419–435.

Lee, D. and Kim, J. K. (2022). Semiparametric imputation using conditional Gaussian mixture models under item nonresponse. *Biometrics* 78, 227–237.

Lee, H. (1995). Outliers in business surveys. In *Business Survey Methods* (B. G. Cox et al., eds.), 503–526. Wiley.

Little, R. J. A. and Smith, P. J. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association* 82, 58–68.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2019). *Robust Statistics: Theory and Methods (with R)* (2nd ed.). Wiley.

Mashreghi, Z., Haziza, D. and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys* 10, 1–52.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63, 581–592.

Sang, H., Kim, J. K. and Lee, D. (2022). Semiparametric fractional imputation using Gaussian mixture models for handling multivariate missing data. *Journal of the American Statistical Association* 117, 654–663.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Sun, Y., Li, J., Xu, Y., Zhang, T. and Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications* 227, 120201.

References iii

Templ, M. (2023). Enhancing precision in large-scale data analysis: An innovative robust imputation algorithm for managing outliers and missing values. *Mathematics* 11, 2729.

Templ, M., Kowarik, A. and Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis* 55, 2793–2806.

Templ, M. and Ulmer, M. (2024). The impact of misclassifications and outliers on imputation methods. *Journal of Applied Statistics* 51, 2894–2928.

Thomas, T. and Rajabi, E. (2021). A systematic review of machine learning-based missing value imputation techniques. *Data Technologies and Applications* 55, 558–585.

Vanden Branden, K. and Verboven, S. (2009). Robust data imputation. *Computational Biology and Chemistry* 33, 7–13.

Wada, K. (2020). Outliers in official statistics. *Japanese Journal of Statistics and Data Science* 3, 669–691.

Wang, H., Kim, J. K., Han, J. and Lee, Y. (2024). Robust propensity score weighting estimation under missing at random. *Electronic Journal of Statistics* 18, 2687–2720.

Backup slides

derivations, algorithms and full tables for the discussion

Backup: derivation of the conditional bias (sketch)

1. Linearise the WLS coefficient: $\widehat{\mathbf{B}}_{\mathrm{WLS}} - \boldsymbol{\beta} = \{\sum_{k\in U} p_k \mathbf{v}_k \phi_k^{-1} \mathbf{v}_k^\top\}^{-1} \sum_{k\in S} w_k r_k \mathbf{v}_k \phi_k^{-1} (y_k - \mathbf{v}_k^\top \boldsymbol{\beta}) + o_p(n^{-1/2})$.
2. Substitute: $\widehat{t}_{l,\mathrm{WLS}} - t_y = \sum_{k\in S} w_k \xi_k - t_y + o_p(N/\sqrt{n})$ with $\xi_k = r_k y_k + (1 - r_k)\mathbf{v}_k^\top \boldsymbol{\beta} + \mathbf{C} r_k \mathbf{v}_k \phi_k^{-1} (y_k - \mathbf{v}_k^\top \boldsymbol{\beta})$.
3. Condition on $(Y_i = y_i, I_i = 1, r_i = 1)$ and take $\mathbb{E}_q$: use $\mathbb{E}_q(r_j \mid r_i = 1) = p_j$ $(j \neq i)$, giving $w_i y_i + w_i \mathbf{C} \mathbf{v}_i \phi_i^{-1} (y_i - \mathbf{v}_i^\top \boldsymbol{\beta}) + \sum_{j\neq i} w_j I_j \widetilde{\xi}_j - t_y$, $\widetilde{\xi}_j = y_j + \{\mathbf{C} p_j \mathbf{v}_j \phi_j^{-1} - (1 - p_j)\}(y_j - \mathbf{v}_j^\top \boldsymbol{\beta})$.
4. Take $\mathbb{E}_p$ with $\mathbb{E}_p(I_j \mid I_i = 1) = \pi_{ij}/\pi_i$: the HT part becomes $B_i^{\mathrm{sam}}(\mathbf{y}) = \sum_{j\in U} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_j$ and the rest is a sum of $(y_j - \mathbf{v}_j^\top \boldsymbol{\beta})$, $j \neq i$.
5. Take $\mathbb{E}_m$: $\mathbb{E}_m(Y_j - \mathbf{v}_j^\top \boldsymbol{\beta} \mid Y_i = y_i) = 0$ for $j \neq i$ kills those terms, leaving

$$B_i \approx \mathbb{E}_m\{B_i^{\mathrm{sam}}(\mathbf{Y}) \mid Y_i = y_i\} + w_i \mathbf{C} \mathbf{v}_i \phi_i^{-1} (y_i - \mathbf{v}_i^\top \boldsymbol{\beta}) = (w_i - 1) y_i + \sum_{j\neq i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \mathbf{v}_j^\top \boldsymbol{\beta} + w_i \mathbf{C} \mathbf{v}_i \phi_i^{-1} (y_i - \mathbf{v}_i^\top \boldsymbol{\beta}).$$

6. Estimate with completed values $\widetilde{y}_j$, $\widehat{\mathbf{C}}$, $\widehat{\mathbf{B}}_{\mathrm{WLS}}$ and the conditional HT coefficients $(\pi_{ij} - \pi_i \pi_j)/(\pi_j \pi_{ij})$. Under SRSWOR and simple linear regression the finite-sample factors $n/(n-1)$ and $n_r/(n_r - 1)$ appear; they are dropped in the main text.

Backup: the RPPB algorithm (SRSWOR)

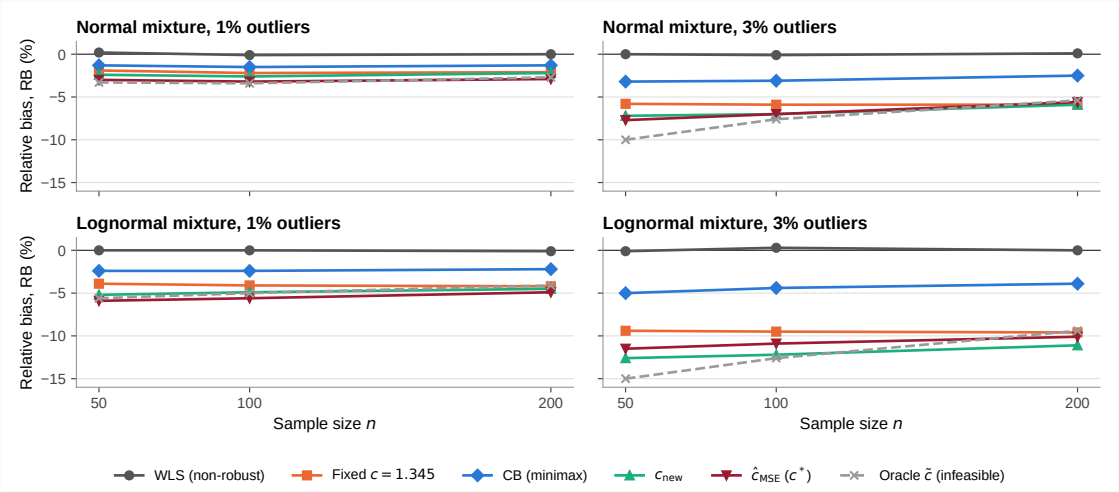
1. On the original respondents compute c_{new} and $\widehat{\mathbf{B}}_R(c_{\mathrm{new}})$; residuals $e_i = y_i - \mathbf{v}_i^\top \widehat{\mathbf{B}}_R(c_{\mathrm{new}})$, scaled $\widetilde{e}_i = e_i / \widehat{s}$ ($\widehat{s}$ = MAD of the WLS residuals); Huber weights $h_i(c_{\mathrm{new}}) = \psi_{c_{\mathrm{new}}}(\widetilde{e}_i)/\widetilde{e}_i$ (1 if $\widetilde{e}_i = 0$); $h_i = 1$ for $i \in S_m$.
2. Initial replication weights $\widetilde{w}_i = d_i h_i(c_{\mathrm{new}})$ for $i \in S_r$, $\widetilde{w}_i = d_i = N/n$ for $i \in S_m$; $\delta_i = \mathbb{1}\{i \in S_r, h_i < 1\}$ flags down-weighted respondents. Final weights $w_i^* = \delta_i \widetilde{w}_i + (1-\delta_i)\widehat{\gamma}\widetilde{w}_i$ with $\widehat{\gamma} = \{N - \sum_i \delta_i \widetilde{w}_i\} / \sum_i (1-\delta_i)\widetilde{w}_i$, so that $\sum_{i \in S} w_i^* = N$.
3. Pseudo-population: $\lfloor w_i^* \rfloor$ copies of unit i plus one more with probability $w_i^* - \lfloor w_i^* \rfloor$ (Poisson sampling); $E_*(\widetilde{N}^* \mid S) = N$. Copies inherit $\mathbf{v}_i$, r_i and (for respondents) y_i; nonrespondent copies stay missing.
4. For $b = 1, \ldots, B$: draw an SRSWOR sample of size n from the pseudo-population; compute $\widehat{t}_{I,\mathrm{WLS}}^{*(b)}$; redo the *complete* adaptive procedure — conditional biases, $\widehat{Z}_i$, $c_{\mathrm{new}}^{*(b)}$, robust fit — to get $\widehat{t}_{I,R}^{*(b)}(c_{\mathrm{new}}^{*(b)})$.
5. $\widehat{V}_{B,R}^* = \frac{1}{B-1} \sum_b \{\widehat{t}_{I,R}^{*(b)} - \widehat{t}_{I,R}^{*(\cdot)}\}^2$; $\Delta^{*(b)} = \widehat{t}_{I,R}^{*(b)} - \widehat{t}_{I,\mathrm{WLS}}^{*(b)}$, $\widehat{V}_{B,\Delta}^* = \frac{1}{B-1} \sum_b \{\Delta^{*(b)} - \Delta^{*(\cdot)}\}^2$;

$$\widehat{\mathrm{MSE}}_{\mathrm{RPPB}} = \widehat{V}_{B,R}^* + \max\left[0, \{\widehat{t}_{I,R}^{\,\mathrm{proj}}(c_{\mathrm{new}}) - \widehat{t}_{I,\mathrm{WLS}}\}^2 - \widehat{V}_{B,\Delta}^*\right].$$

Why reweight?

Respondents with large scaled residuals have $h_i < 1$ and are represented fewer times: an influential unit is not mechanically replicated N/n times; $\widehat{\gamma}$ redistributes the removed mass. The estimator is bootstrap-*assisted*: variances by bootstrap, squared bias from the robust–WLS discrepancy.

Backup: mixture distributions — relative bias



Backup: Monte Carlo RB and RE — symmetric distributions (Table 3)

	Normal			Student t			Laplace		
Estimator / n	50	100	200	50	100	200	50	100	200
$\widehat{t}_{l,\mathrm{WLS}}$	0.2 (100)	−0.2 (100)	0.1 (100)	0.1 (100)	−0.0 (100)	−0.1 (100)	−0.2 (100)	−0.0 (100)	0.1 (100)
$\widehat{t}_{l,R}(1.345)$	0.2 (103)	−0.2 (103)	0.1 (102)	0.0 (69)	−0.1 (68)	−0.1 (66)	−0.2 (85)	−0.0 (82)	0.1 (81)
$\widehat{t}_{l,\mathrm{CB}}$	−0.1 (101)	−0.4 (100)	−0.1 (100)	−0.1 (83)	−0.2 (82)	−0.2 (83)	−0.4 (93)	−0.2 (93)	−0.1 (95)
$\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}})$	0.4 (102)	−0.1 (101)	0.1 (100)	0.4 (75)	0.3 (73)	0.1 (74)	0.2 (90)	0.3 (90)	0.2 (92)
$\widehat{t}_{l,R}^{\,\mathrm{proj}}(c^*)$	0.1 (106)	−0.3 (105)	0.0 (103)	−0.0 (68)	−0.1 (65)	−0.1 (64)	−0.3 (82)	−0.1 (80)	0.1 (80)
$\widehat{t}_{l,R}(\widetilde{c})$ (oracle)	0.2 (100)	−0.2 (100)	0.1 (100)	0.1 (60)	−0.1 (57)	−0.0 (55)	−0.2 (76)	0.0 (68)	0.2 (65)

Entries: percent relative bias (relative efficiency in parentheses, WLS $=$ 100); $R = 10{,}000$.

Backup: Monte Carlo RB and RE — asymmetric distributions (Table 4)

Estimator	n	Pareto	Fréchet	Lognormal	Weibull
$\widehat{\iota}_{l,\mathrm{WLS}}$	50/100/200	0.2/−0.1/0.1 (100)	−0.3/0.1/0.0 (100)	−0.2/0.2/0.0 (100)	−0.3/0.1/0.1 (100)
$\widehat{\iota}_{l,R}(1.345)$	50	−7.9 (50)	−7.8 (55)	−8.4 (79)	−7.0 (92)
	100	−8.6 (58)	−8.2 (69)	−8.9 (95)	−7.4 (111)
	200	−8.7 (93)	−8.5 (102)	−9.3 (134)	−7.7 (141)
$\widehat{\iota}_{l,\mathrm{CB}}$	50	−3.4 (70)	−3.8 (70)	−3.8 (88)	−3.3 (95)
	100	−3.1 (64)	−2.9 (73)	−2.7 (90)	−2.0 (98)
	200	−2.3 (69)	−2.2 (76)	−1.9 (94)	−1.2 (98)
$\widehat{\iota}_{l,R}^{\mathrm{proj}}(c_{\mathrm{new}})$	50	−6.3 (50)	−6.6 (52)	−6.6 (83)	−5.9 (91)
	100	−5.5 (45)	−5.4 (60)	−5.1 (88)	−4.6 (103)
	200	−4.1 (54)	−4.1 (67)	−3.4 (94)	−3.0 (102)
$\widehat{\iota}_{l,R}^{\mathrm{proj}}(c^*)$	50	−6.9 (63)	−7.8 (59)	−7.8 (90)	−6.6 (95)
	100	−5.4 (52)	−5.4 (67)	−5.0 (92)	−3.8 (101)
	200	−4.0 (58)	−4.0 (70)	−3.3 (96)	−2.2 (100)
$\widehat{\iota}_{l,R}(\widetilde{c})$ (oracle)	50	−7.1 (50)	−6.5 (51)	−6.9 (83)	−2.8 (91)
	100	−4.8 (45)	−4.7 (58)	−3.9 (86)	−1.3 (98)
	200	−3.1 (53)	−3.0 (63)	−2.4 (91)	−0.5 (97)

Entries: percent relative bias (relative efficiency in parentheses, WLS $=$ 100); $R = 10{,}000$.

Backup: Monte Carlo RB and RE — mixture distributions (Table 5)

		Mixture of normals		Mixture of lognormals	
Estimator	n	1% outliers	3% outliers	1% outliers	3% outliers
$\widehat{t}_{l,\mathrm{WLS}}$	50/100/200	0.2/−0.1/0.0 (100)	0.0/−0.1/0.1 (100)	0.0/0.0/−0.1 (100)	−0.1/0.3/0.0 (100)
$\widehat{t}_{l,R}(1.345)$	50	−1.9 (50)	−5.8 (38)	−3.9 (30)	−9.4 (27)
	100	−2.2 (51)	−5.9 (45)	−4.1 (33)	−9.5 (36)
	200	−2.1 (53)	−5.9 (58)	−4.2 (43)	−9.6 (60)
$\widehat{t}_{l,\mathrm{CB}}$	50	−1.3 (74)	−3.2 (72)	−2.4 (61)	−5.0 (67)
	100	−1.5 (71)	−3.1 (73)	−2.4 (60)	−4.4 (69)
	200	−1.3 (73)	−2.5 (79)	−2.2 (64)	−3.9 (76)
$\widehat{t}_{l,R}^{\,\mathrm{proj}}(c_{\mathrm{new}})$	50	−2.4 (53)	−7.2 (50)	−5.2 (29)	−12.6 (38)
	100	−2.6 (53)	−7.0 (55)	−4.9 (32)	−12.2 (48)
	200	−2.2 (58)	−5.9 (70)	−4.5 (42)	−11.1 (74)
$\widehat{t}_{l,R}^{\,\mathrm{proj}}(c^*)$	50	−3.0 (58)	−7.7 (66)	−5.9 (43)	−11.5 (67)
	100	−3.2 (57)	−7.0 (73)	−5.6 (39)	−10.9 (72)
	200	−2.9 (60)	−5.6 (83)	−4.9 (48)	−10.1 (78)
$\widehat{t}_{l,R}(\widetilde{c})$ (oracle)	50	−3.3 (46)	−10.0 (41)	−5.6 (26)	−15.0 (33)
	100	−3.4 (50)	−7.6 (54)	−5.0 (30)	−12.6 (50)
	200	−2.7 (55)	−5.4 (67)	−4.1 (41)	−9.4 (71)

Entries: percent relative bias (relative efficiency in parentheses, WLS = 100); $R = 10{,}000$.

Backup: naive fixes — full results (Section 3 simulation)

		WLS	Huber, fixed c			WLS after excluding outliers			
Outliers	n		$c = 0.1$	$c = 1.345$	$c = 2.5$	stud. > 2	stud. > 2.5	stud. > 3	Cook's D
Symmetric	100	0.0 (100)	0.0 (54)	0.0 (55)	0.0 (57)	0.0 (54)	0.0 (55)	−0.0 (57)	−0.0 (56)
	200	−0.1 (100)	−0.0 (54)	−0.0 (55)	−0.1 (57)	−0.0 (53)	−0.1 (55)	−0.1 (57)	−0.1 (55)
Asymmetric	100	0.0 (100)	−13.0 (77)	−12.7 (75)	−11.9 (74)	−10.7 (87)	−9.6 (89)	−8.6 (92)	−8.6 (92)
	200	−0.0 (100)	−13.0 (122)	−12.7 (118)	−12.0 (112)	−10.2 (117)	−8.9 (114)	−7.7 (111)	−7.9 (114)

Entries: percent relative bias (relative efficiency in parentheses, WLS $= 100$); $N = 10{,}000$, $p_i = 0.5$, $R = 10{,}000$.

Backup: relative bias (%) of the MSE estimators (Supplementary Tables)

	Plug-in $\widehat{\mathrm{MSE}}\{\hat{t}_{l,R}^{\mathrm{proj}}(c^*)\}$			RPPB $\widehat{\mathrm{MSE}}\{\hat{t}_{l,R}^{\mathrm{proj}}(c_{\mathrm{new}})\}$		
Distribution / n	50	100	200	50	100	200
Normal	−27.6	−18.6	−13.2	9.6	2.6	−5.1
Student t	−25.2	−18.2	−20.3	0.7	−10.6	−12.6
Laplace	−29.0	−25.1	−21.4	1.1	−7.5	−11.4
Pareto	−18.3	−9.0	−14.5	−15.1	−15.1	4.8
Fréchet	−25.3	−15.4	−15.3	−13.6	−19.5	−17.5
Lognormal	−29.0	−22.1	−17.4	−12.3	−24.6	−20.2
Weibull	−28.2	−20.0	−17.3	0.4	−13.4	26.7

Percent relative bias of the MSE estimators; $R = 10{,}000$.

Backup: notation

Symbol	Meaning
U, S, N, n	finite population, sample, and their sizes
I_i, π_i, π_{ij}, $w_i = 1/\pi_i$	inclusion indicator, inclusion probabilities, design weight
y_i, $t_y = \sum_{i \in U} y_i$	study variable and its population total
r_i, p_i, S_r, S_m	response indicator and probability; respondents; nonrespondents
$\mathbf{v}_i$, $\boldsymbol{\beta}$, ϕ_i, σ^2	auxiliary vector, regression coefficients, variance weight, scale
$\mathbb{E}_m, \mathbb{E}_p, \mathbb{E}_q$	expectation under the imputation model, the design, the response mechanism
$\widehat{\mathbf{B}}_{\mathrm{WLS}}$, $\widehat{t}_{l,\mathrm{WLS}}$	WLS coefficient and WLS imputed estimator (projection form $\sum_{i \in S} w_i \mathbf{v}_i^\top \widehat{\mathbf{B}}_{\mathrm{WLS}}$)
ψ_c, $\widehat{\mathbf{B}}_R(c)$, $\widehat{t}_{l,R}(c)$, $\widehat{t}_{l,R}^{\,\mathrm{proj}}(c)$	Huber score, robust coefficient, observed-plus-imputed and projection robust estimators
B_i, $\widehat{B}_i$, $\widehat{Z}_i$, $\widehat{A}_B$	conditional bias, its estimator, standardised version, asymmetry measure
$\widehat{t}_{l,\mathrm{CB}}$, c_{new}, $\widehat{c}_{\mathrm{MSE}} = c^*$, $\widetilde{c}$	minimax CB estimator; new adaptive constant; MSE-selected constant; oracle constant
$\eta_i(c)$, τ_i, $\mathbf{J}_c$	linearised variables of the robust and WLS estimators; Jacobian of the score
h_i, δ_i, w_i^*, $\widehat{\gamma}$	Huber weight, down-weighting flag, RPPB replication weight, rescaling factor
RB, RE	Monte Carlo percent relative bias; relative efficiency (MSE ratio to WLS $\times 100$)