Agnostic Model-Assisted Estimation with Machine Learning for Survey Data

David Haziza

Department of Mathematics and Statistics, University of Ottawa

Joint work with

Ziming An (University of Ottawa)

Mehdi Dagdoug (McGill University)

Yves Tillé (Université de Neuchâtel)

Seminar of the International Association of Survey Statisticians

September 17, 2026

Outline

1. The Horvitz–Thompson estimator
2. Model-assisted estimation based on linear regression
3. Model-assisted estimation based on machine learning methods
4. Final remarks

Basic setup

- Consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ of N units.
- Goal: Estimate

$$\mu \;=\; \frac{1}{N} \sum_{k \in U} y_k$$

- A sample S of size n is selected according to a sampling design.
- The vector $\mathbf{I} = (I_1, \ldots, I_k, \ldots, I_N)^\top$ contains the sample selection indicators:
  - $I_k = 1$ if $k \in S$;
  - $I_k = 0$ otherwise.
- The first-order inclusion probabilities are defined as

$$\pi_k \;=\; \mathbb{P}(k \in S) > 0 \quad \text{for all } k \in U$$

- The second-order inclusion probabilities are defined as

$$\pi_{k\ell} \;=\; \mathbb{P}(k, \ell \in S) > 0, \qquad k \neq \ell \in U.$$

The Horvitz–Thompson estimator

- An estimator of μ is the Horvitz–Thompson estimator

$$\widehat{\mu}_{\mathrm{HT}} = \frac{1}{N} \sum_{k \in S} w_k y_k, \qquad w_k := \pi_k^{-1}.$$

- The error can be written as

$$\widehat{\mu}_{\mathrm{HT}} - \mu = \frac{1}{N} \sum_{k \in U} \left(\frac{I_k}{\pi_k} - 1 \right) y_k.$$

- Design-unbiasedness:

$$\mathrm{Bias}_d(\widehat{\mu}_{\mathrm{HT}}) \equiv \mathbb{E}_d(\widehat{\mu}_{\mathrm{HT}} - \mu) = 0.$$

- Under mild regularity conditions,

$$\widehat{\mu}_{\mathrm{HT}} - \mu = O_p\left(n^{-1/2}\right),$$

$\Longrightarrow \widehat{\mu}_{\mathrm{HT}}$ is $\sqrt{n}$-consistent.

Variance estimation

- The design variance of the Horvitz–Thompson estimator is

$$\mathbb{V}_d(\widehat{\mu}_{\mathrm{HT}}) = \frac{1}{N^2} \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}, \qquad \Delta_{k\ell} := \pi_{k\ell} - \pi_k \pi_\ell.$$

- A design-unbiased variance estimator is

$$\widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{HT}}) = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell}.$$

- Under mild regularity conditions,

$$\frac{\widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{HT}})}{\mathbb{V}_d(\widehat{\mu}_{\mathrm{HT}})} \xrightarrow{p} 1.$$

$\Longrightarrow \widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{HT}})$ is design-consistent.

Central limit theorem

- Under suitable conditions,

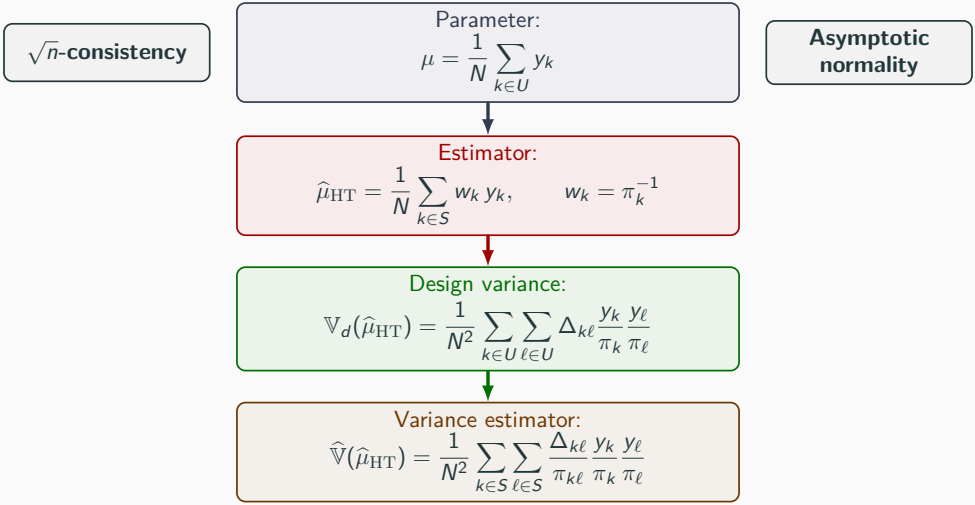
$$\frac{\widehat{\mu}_{\mathrm{HT}} - \mu}{\sqrt{\widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{HT}})}} \xrightarrow{d} \mathcal{N}(0,1),$$

- Central limit theorems for Horvitz–Thompson estimators and related classes of estimators have been established by Hájek (1960), Hájek (1964), Krewski and Rao (1981), Rosén (1997) and Chen and Rao (2007), among others.

- An approximate 95% confidence interval for μ is

$$\widehat{\mu}_{\mathrm{HT}} \pm 1.96 \sqrt{\widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{HT}})}.$$

The chain process: the Horvitz–Thompson estimator



Early work in model-assisted estimation

- Early developments in model-assisted estimation focused primarily on linear working models and regression-type estimators: e.g., Cassel et al. (1977), Särndal (1980), Robinson and Särndal (1983), and Särndal et al. (1992).

- Let
$$\mathbf{x}_k = (1, x_{k1}, \ldots, x_{kp})^\top$$
denote a vector of auxiliary variables for unit k.

- At the estimation stage, we assume that:
  - $\mathbf{x}_k$ is observed for each sampled unit $k \in S$;
  - the population total $\mathbf{t}_\mathbf{x} = \sum_{k \in U} \mathbf{x}_k$ is known.

A linear working model

- Consider the linear working model

$$y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \qquad \mathbb{E}(\varepsilon_k \mid \mathbf{x}_k) = 0.$$

- The model is used as a working device to exploit the association between Y and the auxiliary variables.
- The finite-population values are treated as fixed, and randomness continues to arise from the sampling design.
- The working model need not be correctly specified for the resulting estimator to retain its design-based interpretation.

Model-assisted principle

The linear model is used to improve efficiency, while statistical validity is derived from the sampling design.

A hypothetical population-level fit

- Suppose, hypothetically, that the linear model could be fitted using all the units in the finite population.
- The resulting finite-population least-squares coefficient would be

$$\mathbf{B}_U = \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in U} \mathbf{x}_k y_k.$$

- The corresponding hypothetical fitted values are

$$\widetilde{y}_k = \mathbf{x}_k^\top \mathbf{B}_U, \qquad k \in U.$$

- These fitted values are fixed with respect to the sampling design, but they are infeasible because y_k is not observed for nonsampled units.

The difference estimator

- Using the hypothetical population-level fitted values, define the difference estimator:

$$\widehat{\mu}_{\text{diff}} = \frac{1}{N}\left\{\sum_{k\in U}\mathbf{x}_k^\top \mathbf{B}_U + \sum_{k\in S} w_k \underbrace{\left(y_k - \mathbf{x}_k^\top \mathbf{B}_U\right)}_{\widetilde{e}_k}\right\}.$$

- The error is

$$\widehat{\mu}_{\text{diff}} - \mu = \frac{1}{N}\sum_{k\in U}\left(\frac{I_k}{\pi_k} - 1\right)\widetilde{e}_k, \quad \widetilde{e}_k = y_k - \mathbf{x}_k^\top \mathbf{B}_U$$

- Since $\mathbf{B}_U$ and the residuals $\widetilde{e}_k$ are fixed with respect to the sampling design,

$$\mathbb{E}_d(\widehat{\mu}_{\text{diff}} - \mu) = 0.$$

- Under suitable conditions,

$$\widehat{\mu}_{\text{diff}} - \mu = O_{\mathbb{P}}(n^{-1/2})$$

$\implies \widehat{\mu}_{\text{diff}}$ is $\sqrt{n}$-consistent.

Variance

- Its design variance is

$$\mathbb{V}_d(\widehat{\mu}_{\text{diff}}) = \frac{1}{N^2} \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{\widetilde{e}_k}{\pi_k} \frac{\widetilde{e}_\ell}{\pi_\ell}.$$

- Unlike the Horvitz–Thompson estimator, the difference estimator is driven by the residuals

$$\widetilde{e}_k = y_k - \mathbf{x}_k^\top \mathbf{B}_U,$$

  rather than by the original values y_k.

- If Y and $\mathbf{x}$ are strongly linearly related, the residuals $\widetilde{e}_k$ are expected to be substantially less variable than the y_k.

- We may therefore expect substantial efficiency gains over the Horvitz–Thompson estimator.

Variance estimation

- A design-unbiased variance estimator is

$$\widehat{\mathbb{V}}(\widehat{\mu}_{\text{diff}}) = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widetilde{e}_k}{\pi_k} \frac{\widetilde{e}_\ell}{\pi_\ell}, \qquad \widetilde{e}_k = y_k - \mathbf{x}_k^\top \mathbf{B}_U.$$

- This is the usual Horvitz–Thompson variance estimator applied to the population residuals. Under mild regularity conditions,

$$\frac{\widehat{\mathbb{V}}(\widehat{\mu}_{\text{diff}})}{\mathbb{V}_d(\widehat{\mu}_{\text{diff}})} \xrightarrow{\mathbb{P}} 1.$$

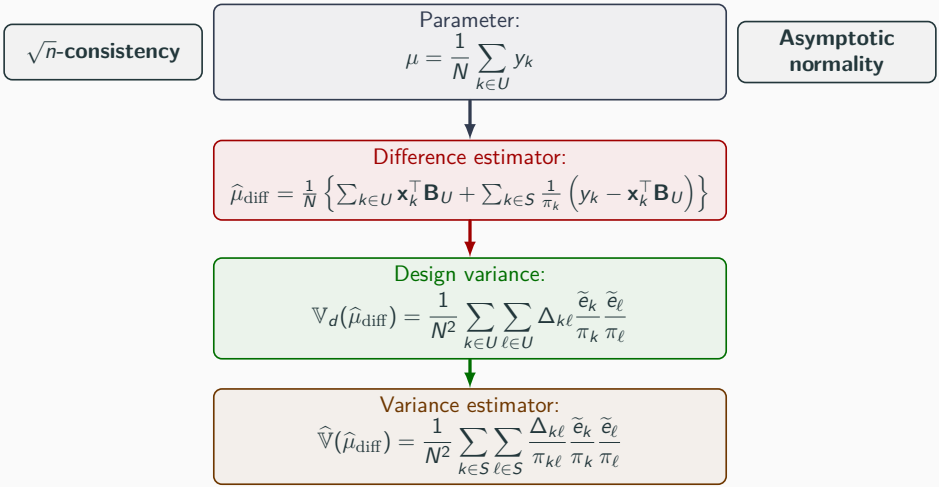
- Classical design-based central limit theorems therefore yield

$$\frac{\widehat{\mu}_{\text{diff}} - \mu}{\sqrt{\widehat{\mathbb{V}}(\widehat{\mu}_{\text{diff}})}} \xrightarrow{d} \mathcal{N}(0,1).$$

- An approximate 95% confidence interval for μ is

$$\widehat{\mu}_{\text{diff}} \pm 1.96 \sqrt{\widehat{\mathbb{V}}(\widehat{\mu}_{\text{diff}})}.$$

The chain process: the difference estimator



From hypothetical fit to sample fit

- The population-level coefficient $\mathbf{B}_U$ is infeasible because it depends on the population values y_k for all $k \in U$.
- In practice, we fit the linear model using the sample only.
- The resulting weighted least-squares coefficient is

$$\widehat{\mathbf{B}} = \left(\sum_{k \in S} w_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S} w_k \mathbf{x}_k y_k = \left(\sum_{k \in U} w_k I_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in U} w_k I_k \mathbf{x}_k y_k.$$

- The corresponding sample-based fitted values are

$$\widehat{y}_k = \mathbf{x}_k^\top \widehat{\mathbf{B}}$$

The generalized regression estimator

- Replacing the hypothetical population fit by the sample fit yields the generalized regression estimator (GREG):

$$\widehat{\mu}_{\mathrm{GREG}} = \frac{1}{N}\left\{\sum_{k\in U}\mathbf{x}_k^\top\widehat{\mathbf{B}} + \sum_{k\in S} w_k \underbrace{\left(y_k - \mathbf{x}_k^\top\widehat{\mathbf{B}}\right)}_{e_k}\right\}.$$

- Unlike the population coefficient $\mathbf{B}_U$, $\widehat{\mathbf{B}}$ depends on the selected sample $\Longrightarrow$ the fitted values $\mathbf{x}_k^\top\widehat{\mathbf{B}}$ and the residuals

$$e_k = y_k - \mathbf{x}_k^\top\widehat{\mathbf{B}}$$

  are random with respect to the sampling design.

- As a result, the GREG estimator is generally not exactly design-unbiased. More on this soon...

- Its design properties are established through its asymptotic equivalence to the difference estimator.

GREG versus the difference estimator

- Let $\widehat{\mathbf{t}}_{\mathbf{x},\mathrm{HT}} = \sum_{k \in S} w_k \mathbf{x}_k$. We have the exact decomposition

$$\widehat{\mu}_{\mathrm{GREG}} = \widehat{\mu}_{\mathrm{diff}} + \underbrace{\frac{1}{N} \left(\mathbf{t}_{\mathbf{x}} - \widehat{\mathbf{t}}_{\mathbf{x},\mathrm{HT}}\right)^{\top} \left(\widehat{\mathbf{B}} - \mathbf{B}_U\right)}_{R_n}.$$

- The remainder term R_n is the price paid for estimating $\mathbf{B}_U$ from the sample.
- Under standard regularity conditions, with a fixed number of auxiliary variables,

$$\frac{1}{N} \left(\widehat{\mathbf{t}}_{\mathbf{x},\mathrm{HT}} - \mathbf{t}_{\mathbf{x}}\right) = O_{\mathbb{P}}(n^{-1/2}),$$

$$\widehat{\mathbf{B}} - \mathbf{B}_U = o_{\mathbb{P}}(1).$$

- Hence, $R_n = o_{\mathbb{P}}(n^{-1/2})$, so

$$\widehat{\mu}_{\mathrm{GREG}} = \widehat{\mu}_{\mathrm{diff}} + o_{\mathbb{P}}(n^{-1/2}).$$

- Estimating $\mathbf{B}_U$ therefore has no first-order effect: the two estimators are asymptotically equivalent.

Variance estimation for the GREG estimator

- An estimator of the variance of $\widehat{\mu}_{\mathrm{GREG}}$ is

$$\widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{GREG}}) = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{e_k}{\pi_k} \frac{e_\ell}{\pi_\ell}, \quad e_k = y_k - \mathbf{x}_k^\top \widehat{\mathbf{B}}$$

- Under suitable regularity conditions,

$$\frac{\widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{GREG}})}{\mathbb{V}_d(\widehat{\mu}_{\mathrm{diff}})} \xrightarrow{\mathbb{P}} 1.$$

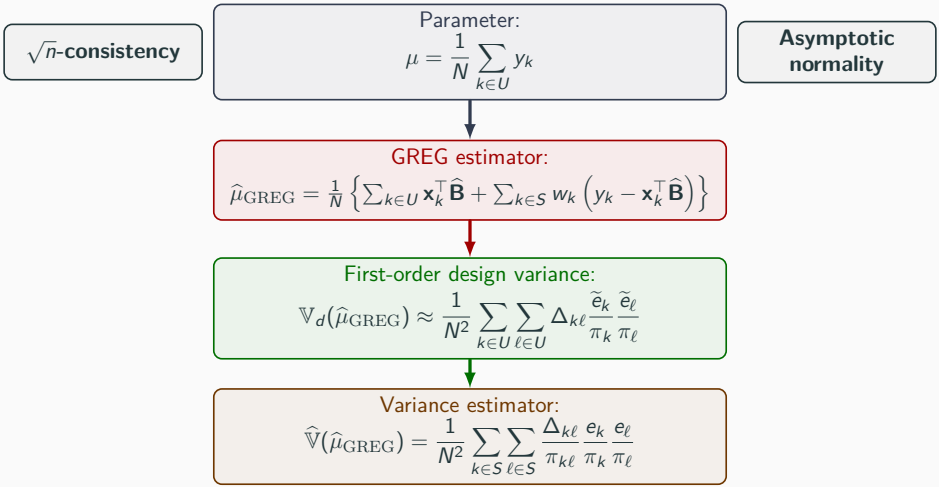
- The GREG estimator inherits the asymptotic normality of the difference estimator:

$$\frac{\widehat{\mu}_{\mathrm{GREG}} - \mu}{\sqrt{\widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{GREG}})}} \xrightarrow{d} \mathcal{N}(0, 1).$$

- An approximate 95% confidence interval for μ is

$$\widehat{\mu}_{\mathrm{GREG}} \pm 1.96 \sqrt{\widehat{\mathbb{V}}(\widehat{\mu}_{\mathrm{GREG}})}.$$

The chain process: the GREG estimator



From linear models to flexible prediction

- In the linear case, fitted values are linear in $\mathbf{x}_k$.
- Therefore, their population total can be recovered from the known auxiliary total $\mathbf{t}_\mathbf{x}$.
- This is why the classical GREG estimator does not require complete unit-level auxiliary information.
- With nonparametric or machine-learning methods, fitted values are no longer linear in $\mathbf{x}_k$.

Key consequence

To evaluate the prediction for every population unit, we generally need

$$\{\mathbf{x}_k : k \in U\}.$$

That is, complete unit-level auxiliary information is required.

Model-assisted estimation beyond linear models

- Many flexible prediction methods have been studied in the model-assisted framework:

- local polynomial regression;
- B-splines;
- penalized splines;
- additive models;
- neural networks.
- k-nearest neighbors;
- regression trees;
- random forests;

Breidt and Opsomer (2000); Goga (2005); Breidt et al. (2005); Montanari and Ranalli (2005); Opsomer et al. (2007); Baffetta et al. (2009); Wang and Wang (2011); Breidt and Opsomer (2017); McConville and Toth (2019); Dagdoug et al. (2023).

Toward agnostic model-assisted estimation

- Existing results are often developed method by method.
- Each prediction method typically requires its own assumptions:
  - smoothness or bandwidth conditions;
  - restrictions on knots or basis dimension;
  - neighborhood-size conditions;
  - assumptions on tree partitions or algorithmic stability.
- Sanguiao Sande and Zhang (2021) propose a subsampling Rao–Blackwell approach that accommodates linear or nonlinear prediction rules and yields exactly design-unbiased estimators.

An alternative approach

In this talk, we present an alternative approach based on cross-fitting. We establish general design-based results under high-level conditions on the sampling design and prediction errors, treating the prediction algorithm as a black box.

Oracle model-assisted estimator

- To describe the relationship between Y and the auxiliary variables, consider the working model
$$y_k = m(\mathbf{x}_k) + \varepsilon_k, \qquad \mathbb{E}(\varepsilon_k \mid \mathbf{x}_k) = 0.$$
- Let $\widetilde{m}(\cdot)$ be a target prediction rule that is fixed with respect to the sampling design. It need not coincide with $m(\cdot)$.
- We construct the oracle (difference) model-assisted estimator
$$\widehat{\mu}_{\mathrm{ma}}(\widetilde{m}, \pi) = \frac{1}{N}\left[\sum_{k \in U} \widetilde{m}(\mathbf{x}_k) + \sum_{k \in S} w_k \underbrace{\{y_k - \widetilde{m}(\mathbf{x}_k)\}}_{\widetilde{e}_k}\right].$$
- Its estimation error is
$$\widehat{\mu}_{\mathrm{ma}}(\widetilde{m}, \pi) - \mu = \frac{1}{N}\sum_{k \in U}\left(\frac{I_k}{\pi_k} - 1\right)\widetilde{e}_k.$$
- Design-unbiased: $\mathbb{E}_d\{\widehat{\mu}_{\mathrm{ma}}(\widetilde{m}, \pi) - \mu\} = 0.$

What is the oracle rule $\widetilde{m}$?

- Let $\widehat{m}(\cdot)$ be a prediction rule trained on the observed sample. We assume that its predictions approach a deterministic target $\widetilde{m}(\cdot)$ that is fixed under the sampling design:

$$\frac{1}{N}\sum_{k\in U}\mathbb{E}_d\Big[\{\widehat{m}(\mathbf{x}_k)-\widetilde{m}(\mathbf{x}_k)\}^2\Big]\longrightarrow 0.$$

- With machine learning, the architecture, complexity or regularization may vary with the training-sample size.

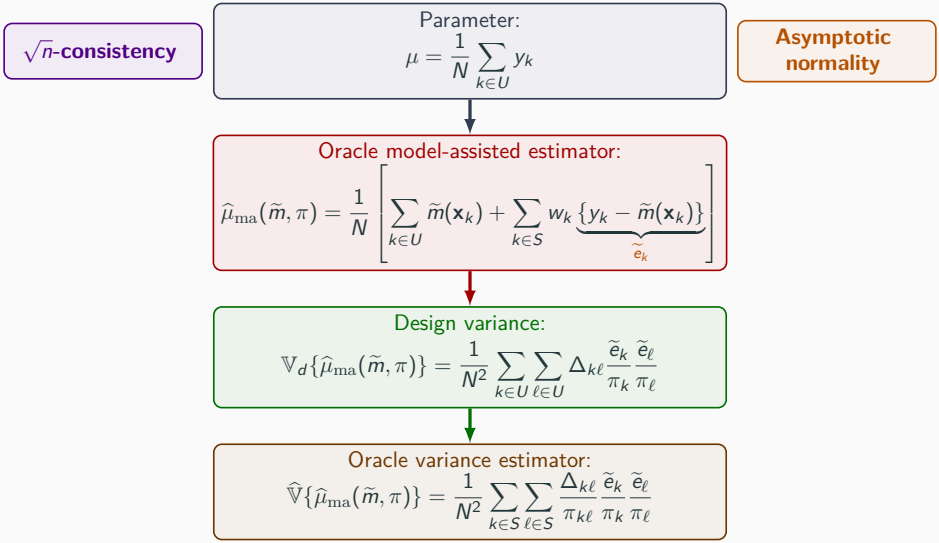
- The target $\widetilde{m}$ need not be the prediction rule fitted on the entire population, nor the regression function m of the working model.

Linear regression: a familiar example

$$\widehat{m}(\mathbf{x}_k)=\mathbf{x}_k^\top\widehat{\mathbf{B}},\qquad \widetilde{m}(\mathbf{x}_k)=\mathbf{x}_k^\top\mathbf{B}_U.$$

In this case, the target coincides with the population regression fit introduced earlier.

The chain process: the oracle model-assisted estimator



Feasible model-assisted estimator

- In practice, the target prediction rule $\widetilde{m}(\cdot)$ is unknown.
- We replace $\widetilde{m}(\cdot)$ by a prediction rule $\widehat{m}(\cdot)$ trained on the observed sample.
- The plug-in model-assisted estimator of μ is

$$\widehat{\mu}_{\mathrm{ma}}(\widehat{m}, \pi) = \frac{1}{N}\left[\sum_{k \in U} \widehat{m}(\mathbf{x}_k) + \sum_{k \in S} w_k \underbrace{\{y_k - \widehat{m}(\mathbf{x}_k)\}}_{e_k}\right].$$

Is this plug-in step harmless?

Not in general. Why? Because the same sample is used twice:

- to construct the prediction rule $\widehat{m}(\cdot)$;
- to form the residual correction term.

This creates a data-reuse problem.

What is the issue?

To make the dependence explicit, write the fitted prediction as $\widehat{m}(\mathbf{x}_k; \mathbf{I})$. Then, the error is

$$\widehat{\mu}_{\mathrm{ma}}(\widehat{m}, \pi) - \mu = \frac{1}{N} \sum_{k \in U} \left(\frac{I_k}{\pi_k} - 1 \right) \{ y_k - \widehat{m}(\mathbf{x}_k; \mathbf{I}) \}.$$

If $\widehat{m}$ were fixed under the design, the design expectation would be zero. Here, however, the fitted prediction depends on the same sampling indicators that appear in the estimator.

Loss of exact design-unbiasedness

$$\mathrm{Bias}_d \left\{ \widehat{\mu}_{\mathrm{ma}}(\widehat{m}, \pi) \right\} = -\frac{1}{N} \sum_{k \in U} \mathrm{Cov}_d \left(\frac{I_k}{\pi_k}, \widehat{m}(\mathbf{x}_k; \mathbf{I}) \right).$$

Interpretation

The design bias is driven by the dependence between the fitted predictions and the sampling indicators. It is not automatically negligible when highly adaptive prediction methods are used.

Another view of the design bias

Design bias

The design bias can be written as

$$\begin{aligned}
\text{Bias}_d\left\{\widehat{\mu}_{\text{ma}}(\widehat{m},\pi)\right\} &= -\frac{1}{N}\sum_{k\in U}\text{Cov}_d\left(\frac{I_k}{\pi_k},\widehat{m}(\mathbf{x}_k;\mathbf{I})\right)\\
&= -\frac{1}{N}\sum_{k\in U}(1-\pi_k)\underbrace{\left[\mathbb{E}_d\{\widehat{m}(\mathbf{x}_k;\mathbf{I})\mid I_k=1\}-\mathbb{E}_d\{\widehat{m}(\mathbf{x}_k;\mathbf{I})\mid I_k=0\}\right]}_{\Delta_k}\\
&= -\frac{1}{N}\sum_{k\in U}(1-\pi_k)\Delta_k
\end{aligned}$$

Interpretation

Δ_k measures how much the average prediction for unit k changes according to whether unit k is included in the sample $\Longrightarrow$ can be viewed as a design-based measure of prediction instability.

Another view of the design bias

- Algorithmic stability describes the sensitivity of a fitted rule to changes in its training sample; Bousquet and Elisseeff (2002), Kearns and Ron (1999).
- Here, Δ_k compares the average predictions conditional on $I_k = 1$ and $I_k = 0$. A small $|\Delta_k|$ means that these conditional mean predictions are close.
- Large values of $|\Delta_k|$ do not necessarily imply a large overall bias: positive and negative contributions may cancel across population units.
- Large values of $|\Delta_k|$ can produce a significant design bias when their contributions tend to have the same sign.

Fixed-size designs

Conditioning on $I_k = 1$ rather than $I_k = 0$ also changes the distribution of the other sampling indicators. Thus, Δ_k reflects changes in the rest of the training sample as well as the inclusion of unit k itself.

Impact on variance estimation

- The same data-reuse problem also affects variance estimation.
- A natural plug-in variance estimator is

$$\widehat{\mathbb{V}}\left\{\widehat{\mu}_{\mathrm{ma}}(\widehat{m},\pi)\right\} = \frac{1}{N^2}\sum_{k\in S}\sum_{\ell\in S}\frac{\Delta_{k\ell}}{\pi_{k\ell}}\frac{e_k}{\pi_k}\frac{e_\ell}{\pi_\ell} = \frac{1}{N^2}\sum_{k\in U}\sum_{\ell\in U}\frac{I_k I_\ell \Delta_{k\ell}}{\pi_{k\ell}}\frac{y_k-\widehat{m}(\mathbf{x}_k;\mathbf{I})}{\pi_k}\frac{y_\ell-\widehat{m}(\mathbf{x}_\ell;\mathbf{I})}{\pi_\ell}.$$

- The same sample determines both the inclusion indicators and the fitted residuals.
- Overfitting can produce artificially small training residuals, leading to substantial underestimation of the design variance.

Good news

The model-assisted estimator is Neyman orthogonal, so one key ingredient is already present. We need a second ingredient: cross-fitting.

Why cross-fitting?

Main idea

Cross-fitting aims to break the dependence created by data reuse by separating training from evaluation.

We partition the population U, and therefore the observed sample S, into M folds.

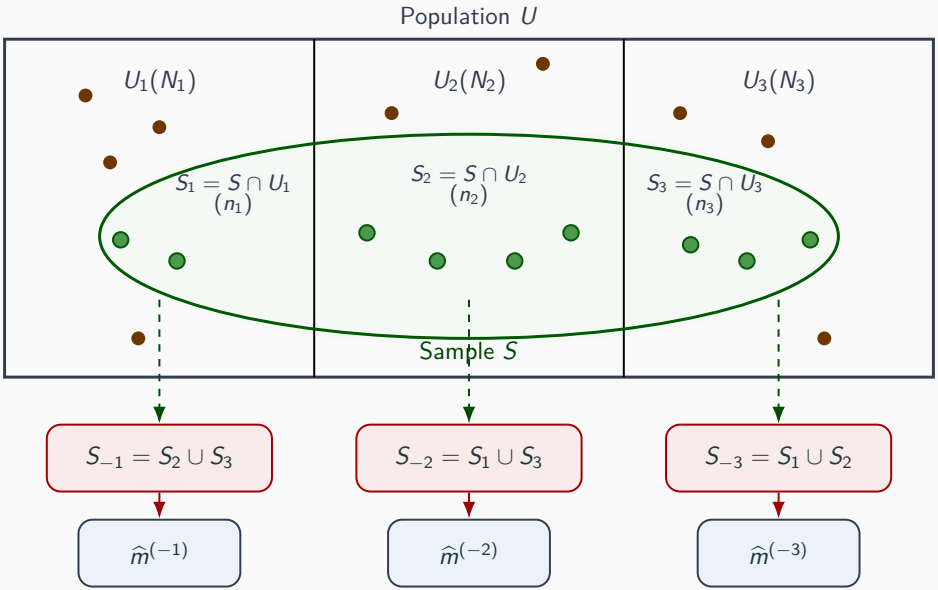
For each fold U_j, the prediction rule is trained using the sampled units outside that fold and is then evaluated on units inside U_j.

Honest predictions

The outcome of a unit is not used to train the rule that predicts that unit.

This produces honest predictions and honest residuals.

Cross-fitting: an illustration with $M = 3$ folds



The conditional-independence property

For a uniform partition independent of S, define the fold indicators

$$\boldsymbol{\delta}_k = (\delta_{k,1}, \ldots, \delta_{k,M})^\top, \qquad \delta_{k,j} = \mathbf{1}\{k \in U_j\},$$
$$\boldsymbol{\delta} = (\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_N^\top)^\top.$$

We condition on the partition $\boldsymbol{\delta}$ and the fold sample counts

$$\mathbf{n} = (n_1, \ldots, n_M)^\top, \qquad n_j = |S \cap U_j|.$$

For which sampling designs?

Our result establishes conditional independence across folds for exponential sampling designs with symmetric support.

This includes SRSWOR, Poisson sampling, and conditional Poisson sampling.

See Tillé (2006) for the definitions of this class of designs and symmetric support.

Illustration: SRSWOR

Conditional on $(\mathbf{n}, \boldsymbol{\delta})$, the design is stratified SRSWOR, with the folds as strata.

Each S_j is an SRSWOR sample of size n_j from U_j, independently across folds.

Consequence for the predictions

Since $\widehat{m}^{(-j)}$ is trained only on S_{-j},

$$\widehat{m}^{(-j)}(\mathbf{x}_k) \perp\!\!\!\perp I_k \mid (\mathbf{n}, \boldsymbol{\delta}), \qquad k \in U_j.$$

We have broken the dependence between the prediction for unit k and its sampling indicator, conditionally on $(\mathbf{n}, \boldsymbol{\delta})$.

Conditional inclusion probabilities

Under the conditional design, the first-order inclusion probabilities are

$$\pi_k^{(\mathbf{n},\boldsymbol{\delta})} = \mathbb{P}(I_k = 1 \mid \mathbf{n}, \boldsymbol{\delta}).$$

- Under SRSWOR,
$$\pi_k^{(\mathbf{n},\boldsymbol{\delta})} = \frac{n_j}{N_j}, \qquad k \in U_j.$$
- Under conditional Poisson sampling, they are computable numerically. This design maximizes entropy for fixed sample size and initial inclusion probabilities.

Conjecture for high-entropy designs

We conjecture that the conditional inclusion probabilities under high-entropy fixed-size designs are well approximated in large samples by those under conditional Poisson sampling, using the same initial π_k's and the same $(\mathbf{n}, \boldsymbol{\delta})$.

Designs of interest include Rao–Sampford and randomized systematic PPS sampling.

Cross-fitted model-assisted estimator

- Let $j(k)$ denote the fold containing unit k. The prediction $\widehat{m}^{(-j(k))}(\mathbf{x}_k)$ uses only observations outside that fold.

- The cross-fitted model-assisted estimator is

$$\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right) = \frac{1}{N}\left[\sum_{k\in U} \widehat{m}^{(-j(k))}(\mathbf{x}_k) + \sum_{k\in S} \frac{1}{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}} \underbrace{\left\{y_k - \widehat{m}^{(-j(k))}(\mathbf{x}_k)\right\}}_{\widehat{e}_k^{\mathrm{cf}}}\right].$$

- The error is

$$\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right) - \mu = \frac{1}{N}\sum_{k\in U}\left(\frac{I_k}{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}} - 1\right) \underbrace{\left\{y_k - \widehat{m}^{(-j(k))}(\mathbf{x}_k)\right\}}_{\widehat{e}_k^{\mathrm{cf}}}.$$

Exact design-unbiasedness

Assume conditional independence across folds and $\pi_k^{(\mathbf{n},\boldsymbol{\delta})} > 0$ for every $k \in U$, almost surely.

Conditional independence

$$\widehat{m}^{(-j(k))}(\mathbf{x}_k) \perp\!\!\!\perp I_k \mid (\mathbf{n}, \boldsymbol{\delta}).$$

Since $\mathbb{E}_d\{I_k/\pi_k^{(\mathbf{n},\boldsymbol{\delta})} - 1 \mid \mathbf{n}, \boldsymbol{\delta}\} = 0$, the conditional bias is

$$\mathbb{E}_d\left[\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right) - \mu \,\middle|\, \mathbf{n}, \boldsymbol{\delta}\right] = -\frac{1}{N}\sum_{k\in U} \mathrm{Cov}_d\left(\frac{I_k}{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}}, \widehat{m}^{(-j(k))}(\mathbf{x}_k) \,\middle|\, \mathbf{n}, \boldsymbol{\delta}\right) = 0.$$

Exact design-unbiasedness

Averaging over $(\mathbf{n}, \boldsymbol{\delta})$ gives

$$\mathbb{E}_d\left[\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right)\right] - \mu = 0.$$

Main assumption on the learner

The oracle uses the same conditional weights and replaces every cross-fitted prediction by a fixed target rule $\widetilde{m}$.

Finite-population prediction convergence

$$\frac{1}{N}\sum_{k\in U}\mathbb{E}_d\left[\left\{\widehat{m}^{(-j(k))}(\mathbf{x}_k)-\widetilde{m}(\mathbf{x}_k)\right\}^2\right]\longrightarrow 0.$$

Oracle equivalence

Under suitable regularity conditions,

$$\widehat{\mu}^{\mathrm{cf}}_{\mathrm{ma}}\left(\widehat{m},\pi^{(\mathbf{n},\boldsymbol{\delta})}\right)=\widehat{\mu}^{\mathrm{cf}}_{\mathrm{ma}}\left(\widetilde{m},\pi^{(\mathbf{n},\boldsymbol{\delta})}\right)+o_{\mathbb{P}}(n^{-1/2}).$$

Interpretation

The target $\widetilde{m}$ need not equal the true regression function. Also, no specific rate of prediction convergence is required.

Cross-fitted variance estimator

Conditionally on $(\mathbf{n}, \boldsymbol{\delta})$, define

$$\pi_{k\ell}^{(\mathbf{n},\boldsymbol{\delta})} = \mathbb{P}\left(I_k = 1, I_\ell = 1 \mid \mathbf{n}, \boldsymbol{\delta}\right),$$
$$\Delta_{k\ell}^{(\mathbf{n},\boldsymbol{\delta})} = \pi_{k\ell}^{(\mathbf{n},\boldsymbol{\delta})} - \pi_k^{(\mathbf{n},\boldsymbol{\delta})}\pi_\ell^{(\mathbf{n},\boldsymbol{\delta})}.$$

Using the cross-fitted residuals $\widehat{e}_k^{\mathrm{cf}} = y_k - \widehat{m}^{(-j(k))}(\mathbf{x}_k)$, we obtain

Variance estimator

$$\widehat{\mathbb{V}}_{\mathrm{cf}}^{(\mathbf{n},\boldsymbol{\delta})} = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}^{(\mathbf{n},\boldsymbol{\delta})}}{\pi_{k\ell}^{(\mathbf{n},\boldsymbol{\delta})}} \frac{\widehat{e}_k^{\mathrm{cf}}}{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}} \frac{\widehat{e}_\ell^{\mathrm{cf}}}{\pi_\ell^{(\mathbf{n},\boldsymbol{\delta})}}.$$

Variance target

Under additional regularity conditions, this consistently estimates the design variance of the oracle estimator based on conditional weights. It is generally not exactly design-unbiased.

First-order properties

Under the regularity conditions on the design and the predictions:

Root-n consistency

$$\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right) - \mu = O_{\mathbb{P}}(n^{-1/2}).$$

Variance consistency

Under some regularity conditions

$$n\left|\widehat{\mathbb{V}}_{\mathrm{cf}}^{(\mathbf{n},\boldsymbol{\delta})} - \mathbb{V}_d\left\{\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widetilde{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right)\right\}\right| = o_{\mathbb{P}}(1).$$

Asymptotic normality

If the standardized oracle estimator satisfies a CLT, then

$$\frac{\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right) - \mu}{\sqrt{\widehat{\mathbb{V}}_{\mathrm{cf}}^{(\mathbf{n},\boldsymbol{\delta})}}} \xrightarrow{d} \mathcal{N}(0,1).$$

The chain process: conditional probabilities

$\sqrt{n}$-consistency

Parameter:

$$\mu = \frac{1}{N}\sum_{k\in U} y_k$$

Asymptotic normality

Cross-fitted model-assisted estimator:

$$\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right) = \frac{1}{N}\left[\sum_{k\in U} \widehat{m}^{(-j(k))}(\mathbf{x}_k) + \sum_{k\in S} \frac{1}{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}} \underbrace{\left\{y_k - \widehat{m}^{(-j(k))}(\mathbf{x}_k)\right\}}_{\widehat{e}_k^{\mathrm{cf}}}\right]$$

First-order oracle variance:

$$\mathbb{V}_d\left\{\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widetilde{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right)\right\} = \frac{1}{N^2}\mathbb{E}_d\left[\sum_{k\in U}\sum_{\ell\in U} \frac{\Delta_{k\ell}^{(\mathbf{n},\boldsymbol{\delta})}}{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}\pi_\ell^{(\mathbf{n},\boldsymbol{\delta})}} \widetilde{e}_k \widetilde{e}_\ell\right]$$

Consistent cross-fitted variance estimator:

$$\widehat{\mathbb{V}}_{\mathrm{cf}}^{(\mathbf{n},\boldsymbol{\delta})} = \frac{1}{N^2}\sum_{k\in S}\sum_{\ell\in S} \frac{\Delta_{k\ell}^{(\mathbf{n},\boldsymbol{\delta})}}{\pi_{k\ell}^{(\mathbf{n},\boldsymbol{\delta})}} \frac{\widehat{e}_k^{\mathrm{cf}}}{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}} \frac{\widehat{e}_\ell^{\mathrm{cf}}}{\pi_\ell^{(\mathbf{n},\boldsymbol{\delta})}}$$

Can we use the original weights $w_k = 1/\pi_k$?

Keep the same cross-fitted predictions and use the original design weights:

$$\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}(\widehat{m}, \pi) = \frac{1}{N}\left[\sum_{k \in U} \widehat{m}^{(-j(k))}(\mathbf{x}_k) + \sum_{k \in S} \frac{1}{\pi_k} \underbrace{\left\{y_k - \widehat{m}^{(-j(k))}(\mathbf{x}_k)\right\}}_{\widehat{e}_k^{\mathrm{cf}}}\right].$$

The error is

$$\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}(\widehat{m}, \pi) - \mu = \frac{1}{N} \sum_{k \in U} \left(\frac{I_k}{\pi_k} - 1\right) \left\{y_k - \widehat{m}^{(-j(k))}(\mathbf{x}_k)\right\}.$$

Practical advantage

We use π_k for point estimation and $\pi_k, \pi_{k\ell}$ for variance estimation. No conditional inclusion probabilities need to be computed.

Original weights: what happens to the bias?

- Cross-fitting still makes the prediction for unit k conditionally independent of I_k, given $(\mathbf{n}, \boldsymbol{\delta})$.
- However, the residual correction is no longer conditionally centered:

$$\mathbb{E}_d\left[\frac{I_k}{\pi_k} - 1 \,\middle|\, \mathbf{n}, \boldsymbol{\delta}\right] = \frac{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}}{\pi_k} - 1.$$

- Consequently, exact design-unbiasedness is generally lost.

The remaining bias is negligible at first order

Under suitable design conditions, if the predictions converge and the conditional inclusion probabilities are sufficiently close to the original inclusion probabilities,

$$\mathrm{Bias}_d\left\{\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}(\widehat{m}, \pi)\right\} = o(n^{-1/2}).$$

Original weights: equivalence to their own oracle

In addition to conditional independence and standard design and moment conditions, we require:

- Prediction convergence: cross-fitted predictions approach a fixed rule $\widetilde{m}$ in finite-population mean square.
- Probability proximity: changes in inclusion probabilities are small enough that their interaction with prediction errors is negligible at the root-n scale.

Direct oracle equivalence

$$\widehat{\mu}^{\mathrm{cf}}_{\mathrm{ma}}\left(\widehat{m}, \pi\right) = \widehat{\mu}_{\mathrm{ma}}\left(\widetilde{m}, \pi\right) + o_{\mathbb{P}}(n^{-1/2}).$$

This oracle uses the original probabilities and fixed predictions.

Therefore,

$$\widehat{\mu}^{\mathrm{cf}}_{\mathrm{ma}}(\widehat{m}, \pi) - \mu = O_{\mathbb{P}}(n^{-1/2}).$$

Original weights: variance estimation and inference

An estimator of the variance of $\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}(\widehat{m}, \pi)$ is

$$\widehat{\mathbb{V}}_{\mathrm{cf}}(\widehat{m}, \pi) = \frac{1}{N^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{e}_k^{\mathrm{cf}}}{\pi_k} \frac{\widehat{e}_\ell^{\mathrm{cf}}}{\pi_\ell}, \quad \widehat{e}_k^{\mathrm{cf}} = y_k - \widehat{m}^{(-j(k))}(\mathbf{x}_k)$$

Under some regularity conditions,

$$n \left[\widehat{\mathbb{V}}_{\mathrm{cf}}(\widehat{m}, \pi) - \mathbb{V}_d\{\widehat{\mu}_{\mathrm{ma}}(\widetilde{m}, \pi)\} \right] = o_{\mathbb{P}}(1).$$

Asymptotic normality

If the oracle satisfies a CLT, then

$$\frac{\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}(\widehat{m}, \pi) - \mu}{\sqrt{\widehat{\mathbb{V}}_{\mathrm{cf}}(\widehat{m}, \pi)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Hence, Wald-type confidence intervals are asymptotically valid.

Are the two weighting choices asymptotically equivalent?

Let $\widetilde{e}_k = y_k - \widetilde{m}(\mathbf{x}_k)$ and define

$$B_{\mathbf{n},\boldsymbol{\delta}} = \frac{1}{N} \sum_{k \in U} \left(\frac{\pi_k^{(\mathbf{n},\boldsymbol{\delta})}}{\pi_k} - 1 \right) \widetilde{e}_k.$$

This is the conditional bias of the original-weight oracle, given $(\mathbf{n}, \boldsymbol{\delta})$.

Under the comparison conditions,

$$\widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}(\widehat{m}, \pi) - \widehat{\mu}_{\mathrm{ma}}^{\mathrm{cf}}\left(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})}\right) = B_{\mathbf{n},\boldsymbol{\delta}} + o_{\mathbb{P}}(n^{-1/2}).$$

An additional condition is needed

First-order equivalence holds if and only if

$$\sqrt{n}\, B_{\mathbf{n},\boldsymbol{\delta}} = o_{\mathbb{P}}(1).$$

Under SRSWOR, this holds with an independent uniform population partition and a fixed number of folds.

Original versus conditional weights: the practical message

Let $\mathbb{V}^{\mathrm{or}}_{\mathrm{orig}}$ and $\mathbb{V}^{\mathrm{or}}_{\mathrm{cond}}$ denote the two unconditional oracle design variances.

First-order efficiency

Under the conditions for the variance comparison,

$$n\left(\mathbb{V}^{\mathrm{or}}_{\mathrm{orig}} - \mathbb{V}^{\mathrm{or}}_{\mathrm{cond}}\right) = n\,\mathbb{E}_d\left(B^2_{\mathbf{n},\boldsymbol{\delta}}\right) + o(1).$$

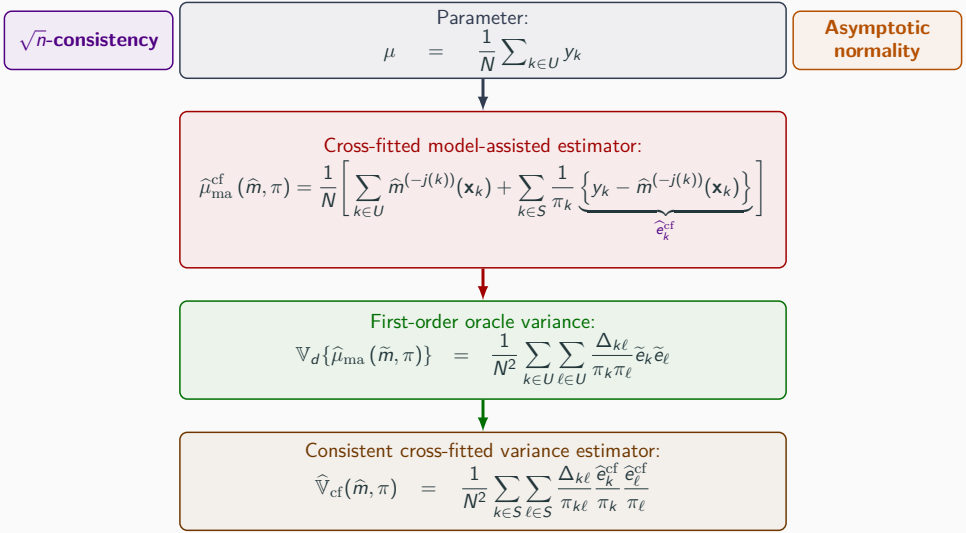
The conditional oracle is therefore at least as efficient at first order.

- Conditional weights: exact design-unbiasedness under conditional independence and positivity, with possible first-order efficiency gains.
- Original weights: simpler implementation, negligible first-order bias, and valid inference.

Take-home message

Both weighting choices support valid first-order inference. This does not require the two estimators to be asymptotically equivalent.

The chain process: original probabilities



Simulation: performance of cross-fitting

- We generated finite populations U of sizes $N \in \{1000, 2500, 5000\}$ with $p = 20$ independent $\mathcal{N}(5, 2)$ covariates.
- Two survey variables were considered:

$$Y_1 = 4X_1 + 3X_2 + 3X_3 + 2X_4 + 5X_5 + 7X_6 + \mathcal{N}(0, 225),$$
$$Y_2 = 10 + 21\mathbb{1}(X_1 > 5) + (X_2 + X_3 - 10)^2 + \mathcal{N}(0, 9).$$

- Samples were selected according to SRSWOR with $n/N = 0.2$.
- Regression function estimators $\widehat{m}$: linear regression, Random forests, k-NN and regression trees based on all 20 predictors.
- We compared 3 different point estimators:
  - Model-assisted without cross-fitting $\widehat{\mu}_{\mathrm{ma}}(\widehat{m}, \pi)$.
  - Cross-fitting ($M = 2$ folds) with conditional inclusion probabilities, $\widehat{\mu}^{\mathrm{cf}}_{\mathrm{ma}}(\widehat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})})$.
  - Cross-fitting ($M = 2$ folds) with unconditional inclusion probabilities, $\widehat{\mu}^{\mathrm{cf}}_{\mathrm{ma}}(\widehat{m}, \pi)$.

Monte-Carlo criteria

We used a Monte Carlo approach with $R = 20,\ 000$ iterations. We computed:

- The Relative Bias (RB) of point estimators:

$$\mathrm{RB}(\widehat{\mu}) = 100\% \times \frac{1}{R} \sum_{r=1}^{R} \frac{\widehat{\mu}^{(r)} - \mu}{\mu},$$

- The Relative Efficiency (RE) with respect to HT:

$$\mathrm{RE}(\widehat{\mu}) = 100\% \times \frac{\mathbb{E}_{MC}\left[(\widehat{\mu} - \mu)^2\right]}{\mathbb{E}_{MC}\left[(\widehat{\mu}_{\mathrm{HT}} - \mu)^2\right]}.$$

- The RB of variance estimators:

$$\text{RB-Var}(\widehat{\mu}) = 100\% \times \frac{\mathbb{E}_{MC}\left[\widehat{\mathbb{V}}(\widehat{\mu})\right] - \mathbb{V}_{MC}(\widehat{\mu})}{\mathbb{V}_{MC}(\widehat{\mu})}.$$

- The Monte-Carlo coverage probabilities, with desired target coverage of 95%.

Results: Relative bias and relative efficiency (1)

Prediction method	Estimator		Linear relationship			Non-linear relationship		
		n	200	500	1000	200	500	1000
Linear Regression	$\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-0.4	-0.2	-0.1
		RE	39	34	34	112	105	103
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})})$	RB	0.0	0.0	0.0	-0.1	0.0	0.0
		RE	44	35	35	133	111	106
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-0.1	0.0	0.0
		RE	44	35	35	133	111	106
Random Forests	$\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-1.1	-0.9	-0.7
		RE	61	50	48	45	34	31
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})})$	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	69	54	52	59	37	25
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	69	54	52	60	37	25

Results: Relative bias and relative efficiency (2)

Prediction method	Estimator		Linear relationship			Non-linear relationship		
		n	200	500	1000	200	500	1000
KNN	$\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-1.7	-1.7	-2.1
		RE	72	65	63	90	135	161
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})})$	RB	0.0	0.0	0.0	-0.1	0.0	0.0
		RE	78	68	66	90	84	81
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-0.1	0.0	0.0
		RE	78	68	66	89	84	81
Regression Trees	$\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-1.4	-1.3	-1.0
		RE	92	78	75	72	66	60
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi^{(\mathbf{n},\boldsymbol{\delta})})$	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	104	84	83	89	58	42
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	104	84	82	89	58	42

Results: Variance estimation and coverage probability (1)

Prediction method	Estimator		Linear Relationship			Non-linear Relationship		
		n	200	500	1000	200	500	1000
Linear Regression	$\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$	RB-Var	-18.2	-7.0	-4.7	-20.3	-8.0	-4.7
		CP	92.1	94.0	94.3	90.9	93.6	94.2
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi^{(\mathbf{n},\delta)})$	RB-Var	0.3	1.0	-0.6	0.6	0.7	-0.7
		CP	94.6	95.0	94.8	94.4	94.8	95.0
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi)$	RB-Var	1.1	1.2	-0.3	1.4	0.9	-0.6
		CP	94.8	95.1	94.7	94.5	94.8	95.0
Random Forests	$\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$	RB-Var	-65.1	-65.8	-66.6	-58.0	-56.4	-60.7
		CP	75.5	74.9	74.5	74.8	68.7	57.7
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi^{(\mathbf{n},\delta)})$	RB-Var	1.4	0.5	-0.8	-2.0	-0.9	-2.8
		CP	94.9	94.9	95.1	94.3	94.4	94.2
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi)$	RB-Var	2.4	0.6	-0.7	-1.6	-0.5	-2.7
		CP	95.0	94.9	95.1	94.3	94.5	94.2

Results: Variance estimation and coverage probability (2)

Prediction method	Estimator		Linear Relationship			Non-linear Relationship		
		n	200	500	1000	200	500	1000
KNN	$\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$	RB-Var	-31.2	-32.0	-32.8	-28.2	-26.3	-27.1
		CP	89.5	89.3	89.0	84.0	77.8	62.1
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi^{(\mathbf{n},\delta)})$	RB-Var	1.8	0.8	-0.8	-0.6	1.0	-0.5
		CP	95.0	95.1	94.9	94.5	95.0	95.0
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi)$	RB-Var	2.5	0.8	-0.7	-0.3	1.0	-0.5
		CP	95.2	95.0	95.0	94.6	95.0	95.0
Regression Trees	$\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$	RB-Var	-79.5	-80.3	-80.9	-66.9	-66.4	-71.0
		CP	62.4	61.7	61.0	68.3	59.4	44.8
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi^{(\mathbf{n},\delta)})$	RB-Var	-1.3	-1.0	0.3	-2.3	-0.2	-1.8
		CP	94.7	94.9	95.0	94.3	94.6	94.3
	$\hat{\mu}^{\text{cf}}_{\text{ma}}(\hat{m}, \pi)$	RB-Var	-0.7	-1.1	0.3	-2.0	0.1	-1.9
		CP	94.8	94.9	95.0	94.3	94.6	94.3

Beyond the population mean: estimating equations

For a known estimating function ψ, define θ_N through the census estimating equation

$$\Psi_N(\theta) = \frac{1}{N} \sum_{k \in U} \psi(y_k, \theta) = 0.$$

- Population mean:

$$\psi(y, \theta) = y - \theta, \qquad \theta_N = \mu = \frac{1}{N} \sum_{k \in U} y_k.$$

- Distribution function at a fixed t:

$$\psi_t(y, \theta) = \mathbb{1}(y \le t) - \theta, \qquad \theta_N = F_N(t) = \frac{1}{N} \sum_{k \in U} \mathbb{1}(y_k \le t).$$

Earlier model-assisted work

Rao et al. (1990) used auxiliary information to estimate finite-population distribution functions and quantiles through ratio- and difference-type estimators.

Model-assisted estimating equations

Let $\widehat{m}(\mathbf{x}_k)$ predict y_k. We use $\psi(\widehat{m}(\mathbf{x}_k), \theta)$ in the estimating equation.

Model-assisted estimating equation

$$\widehat{\Psi}_{\mathrm{ma}}(\theta) = \frac{1}{N}\left[\sum_{k \in U} \psi(\widehat{m}(\mathbf{x}_k), \theta) + \sum_{k \in S} \frac{1}{\pi_k}\left\{\psi(y_k, \theta) - \psi(\widehat{m}(\mathbf{x}_k), \theta)\right\}\right] = 0.$$

- Cross-fitting: replace $\widehat{m}(\mathbf{x}_k)$ by $\widehat{m}^{(-j(k))}(\mathbf{x}_k)$, trained outside the fold containing unit k. This defines $\widehat{\Psi}^{\mathrm{cf}}_{\mathrm{ma}}(\theta)$.
- Estimate θ_N by solving

$$\widehat{\Psi}^{\mathrm{cf}}_{\mathrm{ma}}(\widehat{\theta}_{\mathrm{cf}}) = 0.$$

Work in progress

Our preliminary simulations for finite-population distribution functions are very encouraging.

A detailed theoretical and empirical study is ongoing.

A final perspective: model-based estimation

A long-standing approach

Model-based, or prediction-based, inference dates back at least to Brewer (1963) and Royall (1970). See also Valliant et al. (2000) and Chambers and Clark (2012).

- Use the observed y_k's for sampled units and predict the unobserved values for nonsampled units.
- These estimators can be highly efficient when the model is appropriate. Valid inference relies on assumptions about the population and the sampling mechanism.

Our perspective

We ask how this approach behaves with modern machine-learning predictions.

The challenge with modern machine learning

Existing nonparametric approaches

Kernel, local polynomial and spline methods have already been studied.

Dorfman (1992); Chambers et al. (1993); Dorfman and Hall (1993); Kuk (1993); Zheng and Little (2003); Rueda and Sánchez-Borrego (2009).

Their theory typically relies on method-specific assumptions.

Lack of Neyman orthogonality

The direct model-based estimator is generally not Neyman orthogonal. Prediction errors can therefore enter at first order.

The challenge with general learners

Without additional structure, prediction errors from slowly converging learners may dominate the $n^{-1/2}$ scale. A general theory for root-n inference therefore requires more than prediction consistency.

Why the model-assisted route?

Neyman orthogonality

The model-assisted correction makes the estimating equation Neyman orthogonal.

Cross-fitting

With the conditioning described earlier, cross-fitting provides the conditional independence needed to control the effect of estimating the prediction rule.

Take-home message

Under the conditions discussed earlier, this combination yields an agnostic first-order theory:

- oracle equivalence and root-n consistency;
- consistent variance estimation and asymptotically valid confidence intervals.

References

Baffetta, F., Fattorini, L., Franceschi, S., and Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment*, 113(3):463–475.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Breidt, F. J., Claeskens, G., and Opsomer, J. D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4):831–846.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1026–1053.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205.

References ii

Brewer, K. R. W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5(3):93–105.

Cassel, C.-M., Särndal, C.-E., and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley, New York.

Chambers, R. L. and Clark, R. G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*, volume 37 of *Oxford Statistical Science Series*. Oxford University Press, Oxford.

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88(421):268–277.

Chen, J. and Rao, J. N. K. (2007). Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17(3):1047–1064.

Dagdoug, M., Goga, C., and Haziza, D. (2023). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542):1234–1251.

References iii

Dorfman, A. H. (1992). Non-parametric regression for estimating totals in finite populations. In *Proceedings of the Section on Survey Research Methods*, pages 622–625. American Statistical Association.

Dorfman, A. H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21(3):1452–1475.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire : une approche non paramétrique par splines de régression. *The Canadian Journal of Statistics*, 33(2):163–180.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 5(3):361–374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523.

Kearns, M. and Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453.

References iv

Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9(5):1010–1019.

Kuk, A. Y. C. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika*, 80(2):385–392.

McConville, K. S. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2):389–413.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472):1429–1442.

Opsomer, J. D., Breidt, F. J., Moisen, G. G., and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*, 102(478):400–409.

Rao, J. N. K., Kovar, J. G., and Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77(2):365–375.

References v

Robinson, P. M. and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā: The Indian Journal of Statistics, Series B*, 45:240–248.

Rosén, B. (1997). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62(2):135–158.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387.

Rueda, M. and Sánchez-Borrego, I. R. (2009). A predictive estimator of finite population mean using nonparametric regression. *Computational Statistics*, 24(1):1–14.

Sanguiao Sande, L. and Zhang, L. C. (2021). Design-unbiased statistical learning in survey sampling. *Sankhya A: The Indian Journal of Statistics*, 83(2):714–744.

Särndal, C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67(3):639–650.

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley Series in Survey Methodology. John Wiley & Sons, New York.

Wang, L. and Wang, S. (2011). Nonparametric additive model-assisted estimation for survey data. *Journal of Multivariate Analysis*, 102(7):1126–1140.

Zheng, H. and Little, R. J. A. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19(2):99–117.