Statistical inference in the presence of machine learning based imputation for categorical variables

William David-Eddy Niamessou

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the degree of Master of Science Mathematics and Statistics*

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

© William David-Eddy Niamessou, Ottawa, Canada, 2026

*The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics.

Abstract

In survey sampling, item nonresponse occurs when a sampled unit participates in a survey but does not answer one or more questions. Different imputation methods can be used to address this issue. In this thesis, we first review imputation methods based on parametric models. We then address cases where such models are difficult to specify because the relationship between the survey variable and the auxiliary variables is more complex. In such cases, we consider machine learning methods as an alternative to parametric models. We review existing results on the Augmented Inverse Probability Weighting (AIPW) estimator for continuous variables and the double/debiased machine learning framework, including cross-fitting. The main contribution of this thesis is the use of rejective stochastic imputation for categorical variables within this framework. This procedure produces categorical imputed values while keeping the estimates computed from the completed data file close to the cross-fitted AIPW estimates. Simulation studies are conducted to evaluate the performance of the proposed procedure.

Résumé

En théorie de l'échantillonnage, la non-réponse partielle se produit lorsqu'une unité échantillonnée participe à une enquête, mais ne répond pas à une ou plusieurs questions. Différentes méthodes d'imputation peuvent être utilisées pour remédier à ce problème. Dans cette thèse, nous passons d'abord en revue les méthodes d'imputation fondées sur des modèles paramétriques. Nous abordons ensuite les cas où de tels modèles sont difficiles à spécifier parce que la relation entre la variable d'enquête et les variables auxiliaires est plus complexe. Dans ces cas, nous considérons les méthodes d'apprentissage automatique comme une alternative aux modèles paramétriques. Nous passons en revue les résultats existants sur l'estimateur augmenté par pondération inverse des probabilités (AIPW) pour les variables continues et sur le cadre du double apprentissage automatique débiaisé, y compris le cross-fitting. La principale contribution de cette thèse est l'utilisation de l'imputation stochastique réjective pour les variables catégorielles dans ce cadre. Cette procédure produit des valeurs imputées catégorielles tout en maintenant les estimations calculées à partir du fichier de données complété proches des estimations AIPW obtenues par cross-fitting. Des études de simulation sont menées afin d'évaluer les performances de la procédure proposée.

Acknowledgements

I would like to express my deepest gratitude to my family—my mother, my father, and my brothers—as well as to all my loved ones. This thesis would not have been possible without your love, support, and encouragement.

I would also like to sincerely thank my supervisor, Professor David Haziza, for his patience, advice, and guidance throughout this work.

I dedicate this thesis to all of you.

Contents

List of Figures viii

List of Tables ix

1 Introduction 1

2 Survey sampling and nonresponse 4
- 2.1 Introduction to survey sampling . . . 4
- 2.2 Sampling design . . . 6
- 2.3 The Horvitz-Thompson estimator and its properties . . . 8
- 2.4 Nonresponse in surveys . . . 12
  - 2.4.1 Unit vs. item nonresponse . . . 12
  - 2.4.2 Nonresponse mechanisms: MCAR, MAR, and NMAR . . . 13
- 2.5 Imputation methods . . . 16
  - 2.5.1 Deterministic imputation . . . 19
  - 2.5.2 Random imputation . . . 22
  - 2.5.3 Bias of the imputed estimator . . . 24
- 2.6 Variance of the imputed estimator . . . 27
  - 2.6.1 The two-phase approach . . . 28
  - 2.6.2 The reverse approach . . . 29
- 2.7 The Oracle estimator and its properties . . . 30
  - 2.7.1 Variance of the Oracle estimator . . . 32
  - 2.7.2 Root-n consistency of the Oracle estimator . . . 36
- 2.8 The linear regression imputed estimator and its properties . . . 38
  - 2.8.1 Variance of the linear regression imputed estimator . . . 38
  - 2.8.2 Root-n consistency of the linear regression imputed estimator . . . 43

2.9 Simulation study . . . 44
2.9.1 Simulation design . . . 44
2.9.2 Monte Carlo measures . . . 46
2.9.3 Results . . . 47

3 Inference of doubly robust estimators based on machine learning methods 51
3.1 The imputed estimator under machine learning methods . . . 51
3.1.1 Overview of machine learning methods . . . 51
3.1.2 Rate of convergence of the imputed estimator . . . 56
3.2 The double/debiased machine learning framework . . . 63
3.2.1 The Oracle AIPW estimator . . . 64
3.2.2 Double robustness . . . 66
3.2.3 Variance of the Oracle AIPW estimator . . . 66
3.2.4 The feasible AIPW estimator . . . 68
3.2.5 Cross-fitting and the product-rate condition . . . 70
3.3 Simulation study . . . 73
3.3.1 Simulation setup . . . 73
3.3.2 Slow convergence of the imputed estimator . . . 75
3.3.3 Behaviour of the cross-fitted AIPW estimator . . . 77

4 Imputation of categorical variables 82
4.1 Imputation for a binary variable . . . 84
4.1.1 Oracle imputation for a binary variable . . . 85
4.1.2 Feasible imputation based on logistic regression . . . 90
4.2 Direct stochastic versions of the cross-fitted AIPW estimator . . . 96
4.3 Rejective stochastic imputation . . . 100
4.3.1 The AIPW target for the nonrespondents . . . 100
4.3.2 Can the AIPW target always be reached exactly? . . . 101
4.3.3 The rejective imputation procedure . . . 103
4.3.4 Choice of the tolerance . . . 107
4.3.5 Extension to a categorical variable . . . 107
4.4 Simulation study . . . 113
4.4.1 Common simulation design . . . 113
4.4.2 Simulation design for the binary variable . . . 115
4.4.3 Simulation design for the four-category variable . . . 116
4.4.4 Machine learning methods . . . 118

4.4.5 Simulation results 118

5 Concluding remarks and future research **122**

References **127**

List of Figures

3.1 Distribution of the errors $\hat{\mu}_I - \mu$ 79
3.2 Distribution of the errors $\sqrt{n}(\hat{\mu}_I - \mu)$ 80

List of Tables

2.1 Monte Carlo percent relative bias and coverage rates of the variance estimators. 48
2.2 Relative contributions of the variance components to the total variance estimator. 50

3.1 Monte Carlo results for the linear variable y_1. 78
3.2 Monte Carlo results for the nonlinear variable y_2. 81

4.1 Performance of the rejective stochastic imputed estimator for the binary variable 120
4.2 Performance of the rejective stochastic imputed estimator for the four categories 121

1

Introduction

Item nonresponse is an important challenge in survey sampling. It occurs when a sampled unit participates in a survey but does not provide an answer to one or more questions. This situation affects the estimation of finite population parameters and may lead to invalid statistical inference. In general, missing data can be replaced by plausible values constructed from the available information. Such a procedure is referred to as imputation. Different imputation methods can be used depending on the nature of the survey variable and the assumptions made about the relationship between the variable of interest and the available auxiliary variables.

One of the most commonly used approaches to imputation is based on parametric models. When the assumed model is correctly specified, parametric imputation can provide estimators with desirable statistical properties, such as unbiasedness and consistency, and their variance can be estimated. However, in practice, the relationship between the survey variable and the auxiliary variables may be complex and difficult to specify using a parametric model. In this thesis, we consider cases where it is difficult to specify this relationship using a parametric model. This motivates the use of machine learning methods, which offer greater flexibility and can provide some robustness to model misspecification. They therefore represent an alternative to parametric models. However, good predictive performance does not necessarily translate into valid statistical inference for finite population

parameters. In particular, estimators based on machine learning methods may converge at slower rates than parametric estimators, which can complicate the establishment of desirable inferential properties and the construction of valid variance estimators and confidence intervals.

The challenges associated with machine learning methods can be addressed using the double/debiased machine learning framework. This framework can be used with doubly robust estimators, which combine an outcome regression function with a response propensity function and offer protection against misspecification of either one of the two models. In this thesis, we consider the Augmented Inverse Probability Weighting (AIPW) estimator as a doubly robust estimator. We first review its Oracle version and use it as a benchmark for studying feasible estimators. The double/debiased machine learning framework, together with cross-fitting, is then used to reduce the effects of nuisance function estimation and the dependence resulting from using the same observations to estimate these functions and construct the final estimator. Under suitable conditions, this leads to asymptotically valid inference.

The thesis proceeds as follows:

Chapter 2 reviews the basic concepts and existing methods required throughout the thesis. We first present the survey sampling framework, including probability sampling designs and the Horvitz-Thompson estimator. We then introduce the problem of nonresponse and distinguish between unit and item nonresponse. The main nonresponse mechanisms are discussed, followed by a review of deterministic and random imputation methods. We study the bias and variance of imputed estimators and present two approaches for variance estimation. Finally, we consider the Oracle estimator and the linear regression imputed estimator, review their statistical properties, and conduct a simulation study to evaluate their performance. The chapter therefore establishes the parametric framework that will serve as a basis for the developments in the following chapters.

Chapter 3 also reviews existing methods and results, focusing on inference for imputed estimators when the imputation function is estimated using machine learning methods. We first discuss the limitations of parametric imputation when the relationship between the survey variable and the auxiliary variables is complex. We then introduce several machine learning methods and discuss the convergence rate of the usual imputed estimator. The double/debiased machine learning framework is subsequently introduced through the Oracle AIPW estimator, double robustness, variance estimation, and the cross-fitted AIPW estimator. We review the conditions under which the feasible estimator has the same first-order asymptotic properties as its Oracle counterpart. A simulation study is then

conducted to illustrate the convergence behaviour of the usual imputed estimator and to evaluate the performance of the cross-fitted AIPW estimator.

Chapter 4 presents the main contribution of this thesis: the use of rejective stochastic imputation for categorical variables within the double/debiased machine learning framework. Although the cross-fitted AIPW estimator allows valid inference under suitable conditions, it does not directly produce a completed data file whose weighted mean reproduces the AIPW estimate. Moreover, when the survey variable is categorical, deterministic machine learning predictions may not belong to the support of the original variable. For example, a predicted probability for a binary variable is a continuous quantity and therefore cannot itself be used as an imputed binary value. We first consider the binary case and review Oracle and feasible imputation based on logistic regression. We then examine stochastic versions of the cross-fitted AIPW estimator and develop a rejective stochastic imputation procedure. This procedure produces binary imputed values while keeping the estimate computed from the completed file close to the cross-fitted AIPW estimate. The procedure is subsequently extended to categorical variables with more than two categories. Finally, simulation studies are conducted for both binary and four-category variables to evaluate the performance of the proposed estimators, including their bias, variance estimation, and confidence interval coverage.

Chapter 5 concludes the thesis and summarizes the main results. We discuss the implications of the proposed methodology and outline possible directions for future research.

2

Survey sampling and nonresponse

2.1 Introduction to survey sampling

Survey sampling is a branch of statistics concerned with the study of finite populations. A finite population is a collection of a fixed number of units, such as individuals, households, businesses, or institutions. For each unit in the population, we may be interested in one or several characteristics. For example, we may want to study the income of Canadian households, the number of employees in Canadian businesses, or the proportion of individuals with a particular characteristic. The objective is then to estimate population quantities such as totals, means, proportions, or quantiles.

In principle, these population quantities could be calculated exactly if information were collected from every unit in the population. Such a complete enumeration is called a census. In practice, however, conducting a census is often costly, time-consuming, and difficult to implement, particularly when the population is large. Moreover, collecting information from the entire population does not eliminate other sources of error, such as nonresponse, coverage errors, or measurement errors. For these reasons, statistical studies

are often based on a sample containing only a fraction of the units in the target population.

The purpose of survey sampling is therefore to use the information collected from the sampled units to make reliable inferences about the entire population. An estimator should ideally have little or no bias, meaning that, over repeated samples selected according to the sampling design, it should on average be close to the true population quantity. Its variance should also be reasonably small, since a large variance means that the estimate can change substantially from one possible sample to another. Both bias and variance are therefore important when assessing the quality of a survey estimator.

Survey sampling methods are widely used by national statistical agencies, such as Statistics Canada (StatCan) in Canada and the Institut national de la statistique et des études économiques (INSEE) in France. These agencies conduct surveys on a wide range of economic and social topics and provide information used by governments, businesses, researchers, and the general public. For example, Statistics Canada may use survey data to estimate average household income or to study consumer spending patterns in the Canadian population.

We now describe the framework and introduce the notation used throughout the thesis. Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population of size N and y be a variable of interest. We denote by y_k the value of the variable y for the unit k, while $\boldsymbol{y} = (y_1, \ldots, y_k, \ldots, y_N)^\top$ represents the vector of all population values. In survey sampling, the most common finite population parameters of interest are the population total

$$t_y = \sum_{k \in U} y_k,$$

and the population mean

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

To estimate these finite population parameters, it is common practice to select a sample S of size n ($n < N$) from the population U. The sample is selected according to a probability sampling design, which will be formally defined in the next section.

2.2 Sampling design

Before the sample is selected, S is treated as random, while s denotes a particular realization of S. Let Ω denote the collection of all possible samples that can be drawn from the finite population U. A sampling design assigns a probability $p(s) = P(S = s)$ to each $s \in \Omega$, such that

$$p(s) \geq 0, \qquad \forall s \in \Omega,$$

and

$$\sum_{s \in \Omega} p(s) = 1.$$

The sampling design therefore describes the random mechanism used to select the sample from the finite population. A sample can equivalently be represented through the vector of sample selection indicators

$$\mathbf{I} = (I_1, \ldots, I_k, \ldots, I_N)^\top,$$

where

$$I_k = \begin{cases} 1, & \text{if } k \in S, \\ 0, & \text{if } k \notin S. \end{cases}$$

Under the design-based framework, the population values $y_1, \ldots, y_N$ are regarded as fixed, and the randomness comes from the sample selection mechanism, or equivalently from the indicators $I_1, \ldots, I_N$. We denote by $\mathbb{E}_p(\cdot)$, $\mathbb{V}_p(\cdot)$, and $\mathrm{Cov}_p(\cdot)$ the expectation, variance, and covariance with respect to the sampling design.

For each unit $k \in U$, the first-order inclusion probability is defined as the probability that unit k is included in the sample:

$$\pi_k = P(k \in S) = P(I_k = 1) = \mathbb{E}_p(I_k).$$

Since I_k is a Bernoulli random variable with parameter π_k,

$$\mathbb{V}_p(I_k) = \pi_k(1 - \pi_k).$$

Similarly, for two distinct units $k, l \in U$, the second-order inclusion probability is defined

as

$$\pi_{kl} = P(k \in S, l \in S) = P(I_k = 1, I_l = 1) = \mathbb{E}_p(I_k I_l)$$

It represents the probability that units k and l are simultaneously included in the sample. It follows that

$$\mathrm{Cov}_p(I_k, I_l) = \pi_{kl} - \pi_k \pi_l \equiv \Delta_{kl}.$$

By convention,

$$\pi_{kk} = \pi_k,$$

and clearly

$$\pi_{kl} = \pi_{lk}.$$

The sampling weight associated with unit k is defined as

$$w_k = \frac{1}{\pi_k}.$$

We now present three common examples of probability sampling designs.

Example 1 (Simple random sampling without replacement (SRSWOR))**.** *Simple random sampling without replacement is one of the simplest probability sampling designs. A sample of fixed size n is selected from the population in such a way that every subset of n units has the same probability of being selected. Therefore,*

$$p(s) = \begin{cases} \dfrac{1}{\binom{N}{n}}, & \text{if } |s| = n, \\ 0, & \text{otherwise.} \end{cases}$$

There are $\binom{N}{n}$ possible samples of size n. Under SRSWOR, all units have the same first-order inclusion probability:

$$\pi_k = \frac{n}{N}, \qquad k \in U.$$

For two distinct units k and l, the second-order inclusion probability is

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)}, \qquad k \neq l.$$

SRSWOR is referred to as a fixed-size sampling design because the number of sampled units is equal to n for every possible sample.

Example 2 (Inclusion probabilities proportional to size (IPPS)). *Suppose that an auxiliary variable x, often referred to as a size variable, is known for every unit in the population before sampling. Let*

$$t_x = \sum_{k \in U} x_k$$

denote its population total. For a fixed expected sample size n, inclusion probabilities proportional to size can be defined as

$$\pi_k = n\frac{x_k}{t_x}, \qquad k \in U,$$

provided that $\pi_k \leq 1$ for every unit.

Example 3 (Poisson sampling). *Under Poisson sampling, each population unit is selected independently with a prescribed inclusion probability $\pi_k \in (0, 1)$. More precisely,*

$$I_k \sim \text{Bernoulli}(\pi_k), \qquad k \in U,$$

and the indicators $I_1, \ldots, I_N$ are mutually independent. The probability of selecting a particular sample s is

$$p(s) = \prod_{k \in U} \pi_k^{I_k}(1 - \pi_k)^{1-I_k}.$$

For $k \neq l$,

$$\pi_{kl} = P(I_k = 1, I_l = 1) = \pi_k \pi_l,$$

and consequently,

$$\text{Cov}_p(I_k, I_l) = 0.$$

2.3 The Horvitz-Thompson estimator and its properties

The Horvitz-Thompson estimator, also known as the expansion estimator or the π-estimator, was proposed by Narain (1951) and Horvitz and Thompson (1952). It is an

estimator of the total $t_y = \sum_{k \in U} y_k$. It is given by

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} w_k y_k = \sum_{k \in U} w_k y_k I_k.$$

Proposition 2.1. *If $\pi_k > 0 \ \forall k \in U$, the Horvitz-Thompson estimator $\hat{t}_{y,\pi}$ is design-unbiased for t_y.*

Proof.

$$\begin{aligned}
\mathbb{E}_p(\hat{t}_{y,\pi}) &= \mathbb{E}_p\left(\sum_{k \in S} \frac{y_k}{\pi_k}\right) \\
&= \mathbb{E}_p\left(\sum_{k \in U} \frac{y_k}{\pi_k} I_k\right) \\
&= \sum_{k \in U} \frac{y_k}{\pi_k} \mathbb{E}_p(I_k) \\
&= \sum_{k \in U} \frac{y_k}{\pi_k} \pi_k \\
&= \sum_{k \in U} y_k \\
&= t_y
\end{aligned}$$

□

The case of the population mean follows directly from that of the population total. The Horvitz-Thompson estimator of μ is

$$\hat{\mu}_\pi = \frac{\hat{t}_{y,\pi}}{N}.$$

Since N is known, $\hat{\mu}_\pi$ is design-unbiased for μ. Indeed, using Proposition 2.1,

$$\mathbb{E}_p(\hat{\mu}_\pi) = \mathbb{E}_p\left(\frac{\hat{t}_{y,\pi}}{N}\right) = \frac{\mathbb{E}_p(\hat{t}_{y,\pi})}{N} = \frac{t_y}{N} = \mu.$$

Proposition 2.2. *Suppose that $\pi_k > 0$ for all $k \in U$. The design variance of the Horvitz–Thompson estimator is*

$$\begin{aligned}\mathbb{V}_p(\hat{t}_{y,\pi}) &= \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \\ &= \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \end{aligned} \tag{2.3.1}$$

Proof.

$$\begin{aligned}\mathbb{V}_p(\hat{t}_{y,\pi}) &= \mathbb{V}_p\left(\sum_{k\in S}\frac{y_k}{\pi_k}\right) \\ &= \mathbb{V}_p\left(\sum_{k\in U}\frac{y_k}{\pi_k}I_k\right) \\ &= \sum_{k\in U}\sum_{l\in U}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\operatorname{Cov}_p(I_k, I_l) \\ &= \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\end{aligned}$$

□

The design variance $\mathbb{V}_p(\hat{t}_{y,\pi})$ measures the variability of the Horvitz–Thompson estimator induced by the sampling design. More precisely, if the sampling process could be repeated many times according to the same sampling design, a different sample would generally be selected at each repetition, leading to a different value of $\hat{t}_{y,\pi}$. The design variance measures how much these possible values of the estimator vary around their expectation. Thus, a small design variance indicates that the estimator is relatively stable from one possible sample to another, whereas a large design variance indicates greater sampling variability. Expression (2.3.1) is valid for both fixed-size and random-size sampling designs.

In practice, $\mathbb{V}_p(\hat{t}_{y,\pi})$ cannot generally be computed directly because it involves the values y_k for all units in the finite population, whereas these values are observed only for the

sampled units. Provided that $\pi_{kl} > 0$ for all $k, l \in U$, a standard estimator of $\mathbb{V}_p(\hat{t}_{y,\pi})$ is

$$\widehat{\mathbb{V}}_p(\hat{t}_{y,\pi}) = \sum_{k\in S}\sum_{l\in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

This estimator is design-unbiased:

$$\mathbb{E}_p \left\{ \widehat{\mathbb{V}}_p(\hat{t}_{y,\pi}) \right\} = \mathbb{V}_p(\hat{t}_{y,\pi}).$$

For SRSWOR, the expression (2.3.1) simplifies to

$$\mathbb{V}_p(\hat{t}_{y,\pi}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n},$$

where

$$S_y^2 = \frac{1}{N-1} \sum_{k\in U} (y_k - \mu)^2$$

is the finite population variance. Since S_y^2 is unknown, it is estimated by the sample variance

$$s_y^2 = \frac{1}{n-1} \sum_{k\in S} (y_k - \bar{y})^2,$$

where

$$\bar{y} = \frac{1}{n} \sum_{k\in S} y_k.$$

This leads to the design-unbiased variance estimator

$$\widehat{\mathbb{V}}_p(\hat{t}_{y,\pi}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}.$$

Since $\hat{\mu}_\pi = \hat{t}_{y,\pi}/N$, the design variance of the Horvitz–Thompson estimator of the population mean is

$$\mathbb{V}_p(\hat{\mu}_\pi) = \frac{1}{N^2} \mathbb{V}_p(\hat{t}_{y,\pi}) = \frac{1}{N^2} \sum_{k\in U}\sum_{l\in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

In the SRSWOR case, we obtain

$$\mathbb{V}_p(\hat{\mu}_\pi) = \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n},$$

which can be estimated unbiasedly by

$$\widehat{\mathbb{V}}_p(\hat{\mu}_\pi) = \left(1 - \frac{n}{N}\right) \frac{s_y^2}{n}.$$

2.4 Nonresponse in surveys

2.4.1 Unit vs. item nonresponse

One of the major problems in survey sampling is missing data. Missing data may result from nonresponse, which is generally divided into two types: unit nonresponse and item nonresponse. Unit nonresponse refers to the situation in which no information on the variables of interest can be obtained for a sampled unit k. Reasons for unit nonresponse range from the inability to contact a sampled unit to outright refusal to participate. Unit nonresponse is commonly handled through a weight adjustment procedure, whereby the sampling weights of responding units are increased to compensate for the nonrespondents. In contrast, item nonresponse occurs when information is missing for only some of the survey variables. This may occur, for example, when questions are sensitive, when respondents do not know the answer, or when reported values are invalid or inconsistent. Item nonresponse is typically treated using imputation. Imputation refers to methods used to replace a missing value with a plausible value. In general, auxiliary variables available in the dataset are used to construct an imputation model based on information from the respondents. Even after applying weight adjustment or imputation, however, unit and item nonresponse may lead to invalid inference. In particular, if respondents and nonrespondents differ systematically with respect to the variable of interest, estimators may be biased. For instance, if the objective is to estimate the average income of Canadians and individuals with lower incomes are less likely to respond, an estimator that does not adequately account for nonresponse may overestimate the true average income. Nonresponse may also affect the variance of survey estimators. In the case of unit nonresponse, weight adjustment generally results in more variable weights and a reduced effective sample size, which may increase the variance compared with the full-response case. A similar

issue arises with item nonresponse. When imputation is based on a model fitted using respondent data, additional uncertainty may arise from the nonresponse mechanism, the estimation of the imputation model, and, for some imputation procedures, the generation of the imputed values themselves. Therefore, the main objective of nonresponse treatment methods is to reduce nonresponse bias while keeping the resulting variance as small as possible.

2.4.2 Nonresponse mechanisms: MCAR, MAR, and NMAR

To introduce the different nonresponse mechanisms, let r_k denote the response indicator associated with unit k, defined as

$$r_k = \begin{cases} 1, & \text{if unit } k \text{ is a respondent,} \\ 0, & \text{otherwise.} \end{cases}$$

We denote by

$$\boldsymbol{r} = (r_1, \ldots, r_k, \ldots, r_N)^\top$$

the vector of response indicators in the population. For the selected sample S, let

$$S_r = \{k \in S : r_k = 1\}$$

denote the set of respondents, and let

$$S_m = \{k \in S : r_k = 0\}$$

denote the set of nonrespondents.

For each unit $k \in U$, let $\boldsymbol{v}_k = (v_{k1}, \ldots, v_{kp})^\top$ denote the vector of p fully observed auxiliary variables associated with unit k. We denote by $\mathbf{V}$ the corresponding population matrix of auxiliary variables, whose k-th row is $\boldsymbol{v}_k^\top$, that is,

$$\mathbf{V} = \begin{pmatrix} \boldsymbol{v}_1^\top \\ \vdots \\ \boldsymbol{v}_N^\top \end{pmatrix}.$$

Thus, $\mathbf{V}$ is an $N \times p$ matrix containing the auxiliary information for all units in the population.

The nonresponse mechanism is characterized by the conditional distribution

$$q(\boldsymbol{r} \mid \boldsymbol{y}, \mathbf{V}).$$

We denote by $\mathbb{E}_q(\cdot)$, $\mathbb{V}_q(\cdot)$, and $\mathrm{Cov}_q(\cdot, \cdot)$ the expectation, variance, and covariance with respect to the nonresponse mechanism. Under this framework, the response indicators are treated as random, whereas the population values $\boldsymbol{y}$ and the auxiliary information $\mathbf{V}$ are treated as fixed.

For each unit $k \in U$, the response probability is defined as

$$p_k = P(r_k = 1 \mid y_k, \boldsymbol{v}_k) = \mathbb{E}_q(r_k \mid y_k, \boldsymbol{v}_k).$$

We assume that

$$0 < p_k \leq 1, \qquad k \in U.$$

This positivity assumption ensures that every population unit has a positive probability of responding. We also assume that the response indicators $r_1, \ldots, r_N$ are conditionally independent given $\boldsymbol{y}$ and $\mathbf{V}$.

We now introduce three commonly used nonresponse mechanisms: *Missing Completely At Random (MCAR)*, *Missing At Random (MAR)*, and *Not Missing At Random (NMAR)*:

i) The data are said to be Missing Completely At Random when the probability of response does not depend on either the survey variable or the auxiliary variables. In its simplest form,

$$P(r_k = 1 \mid y_k, \boldsymbol{v}_k) = P(r_k = 1) = p_0, \qquad k \in U,$$

for some $p_0 \in (0, 1)$. Under MCAR, there is no systematic difference between respondents and nonrespondents with respect to the variables under study. Nonresponse can therefore be viewed as an additional random mechanism that is independent of

both y_k and $\boldsymbol{v}_k$. For example, in an online survey, some participants may fail to complete the questionnaire because of purely technical problems, such as a browser malfunction or a temporary Internet connection failure. If these events occur independently of the characteristics and answers of the participants, the resulting nonresponse mechanism can reasonably be regarded as MCAR. Under MCAR, nonresponse does not introduce systematic selection with respect to the variable of interest. However, the number of observed units is reduced, resulting in a loss of information and generally an increase in variance compared with the full-response case.

ii) The data are said to be Missing At Random when the probability of response may depend on fully observed auxiliary variables but, conditionally on these variables, does not depend on the survey variable subject to nonresponse. More precisely, the MAR assumption is

$$P(r_k = 1 \mid y_k, \boldsymbol{v}_k) = P(r_k = 1 \mid \boldsymbol{v}_k) = p(\boldsymbol{v}_k).$$

Equivalently, under MAR,

$$r_k \perp y_k \mid \boldsymbol{v}_k.$$

That is, conditionally on $\boldsymbol{v}_k$, the response indicator r_k and the survey variable y_k are independent. This also implies that

$$f(y_k \mid \boldsymbol{v}_k, r_k = 1) = f(y_k \mid \boldsymbol{v}_k, r_k = 0).$$

In other words, after conditioning on the auxiliary variables $\boldsymbol{v}_k$, respondents and nonrespondents have the same conditional distribution of the survey variable.

iii) The data are said to be Not Missing At Random when the probability of response depends on the survey variable even after conditioning on the available auxiliary variables. In other words,

$$P(r_k = 1 \mid y_k, \boldsymbol{v}_k)$$

still depends on y_k after $\boldsymbol{v}_k$ has been taken into account. Equivalently,

$$r_k \not\perp y_k \mid \boldsymbol{v}_k.$$

In this case,

$$f(y_k \mid \boldsymbol{v}_k, r_k = 1) \neq f(y_k \mid \boldsymbol{v}_k, r_k = 0).$$

Thus, respondents and nonrespondents may have different conditional distributions of the survey variable even after controlling for the observed auxiliary variables. NMAR is generally more difficult to handle because the observed auxiliary information does not fully explain the nonresponse mechanism. For example, in a health survey, individuals with a particular medical condition may be less likely to answer a question about that condition because they consider the question sensitive. If this difference in response propensity remains after conditioning on the available auxiliary variables, the response mechanism is NMAR. Under NMAR, standard weighting or imputation procedures based only on observed auxiliary variables may lead to biased estimators. Additional assumptions or information about the nonresponse mechanism are generally required to obtain valid inference.

2.5 Imputation methods

As mentioned in the previous section, imputation refers to methods used to replace missing values with plausible values. The main objective of imputation is generally not to predict each missing value as accurately as possible, but rather to obtain valid estimates of finite population parameters from the completed data. Let y be a survey variable subject to nonresponse, and suppose that the parameter of interest is the population mean

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

An imputed estimator of μ is given by

$$\hat{\mu}_I = \frac{1}{N} \left(\sum_{k \in S_r} w_k y_k + \sum_{k \in S_m} w_k y_k^* \right), \tag{2.5.1}$$

where y_k^* denotes the imputed value used to replace the missing value y_k for unit $k \in S_m$. The estimator in (2.5.1) can equivalently be written as

$$\hat{\mu}_I = \frac{1}{N} \sum_{k \in S} w_k \left\{ r_k y_k + (1 - r_k) y_k^* \right\} = \frac{1}{N} \sum_{k \in S} w_k \tilde{y}_k,$$

where

$$\tilde{y}_k = r_k y_k + (1 - r_k) y_k^*$$

denotes the completed value for unit k.

Imputation procedures can be classified according to several characteristics. A first distinction is between deterministic and random imputation. Under deterministic imputation, applying the same imputation procedure to the same observed data always produces the same completed dataset. A simple example is mean imputation, in which every missing value is replaced by the mean of the observed values. Other examples include regression imputation and some forms of nearest-neighbour imputation. In contrast, random imputation introduces an additional source of randomness into the imputation procedure. Consequently, applying the same procedure several times to the same observed data may produce different completed datasets. Depending on the method, this randomness may be introduced by drawing a residual, selecting a donor at random, or generating an imputed value from an estimated conditional distribution.

A second distinction is between donor imputation and predicted-value imputation. Under donor imputation, a missing value is replaced by an observed value from a responding unit, referred to as a donor. In contrast, predicted-value imputation replaces a missing value by a value obtained from a fitted statistical model. Donor methods have the advantage of preserving values that are actually observed in the data and are therefore particularly useful when the variable of interest is categorical or has a restricted support.

Many imputation procedures make use of auxiliary variables $\boldsymbol{v}_k$ that are observed for both respondents and nonrespondents. The relationship between the survey variable y_k and the auxiliary variables can be modeled in different ways. In particular, we distinguish between parametric and nonparametric imputation models.

Under a parametric imputation model, we assume that

$$y_k = f(\boldsymbol{v}_k; \boldsymbol{\beta}) + \varepsilon_k, \tag{2.5.2}$$

where $f(\boldsymbol{v}_k; \boldsymbol{\beta})$ is a known function indexed by an unknown finite-dimensional parameter vector $\boldsymbol{\beta}$, and ε_k is a model error satisfying

$$\mathbb{E}_m(\varepsilon_k \mid \boldsymbol{v}_k) = 0$$

and

$$\mathbb{V}_m(\varepsilon_k \mid \boldsymbol{v}_k) = c_k\sigma^2,$$

where c_k is a known positive quantity. We also assume that

$$\mathrm{Cov}_m(\varepsilon_k, \varepsilon_l \mid \boldsymbol{v}_k, \boldsymbol{v}_l) = 0, \qquad k \neq l.$$

We denote by $\mathbb{E}_m(\cdot)$, $\mathbb{V}_m(\cdot)$, and $\mathrm{Cov}_m(\cdot,\cdot)$ the expectation, variance, and covariance with respect to the imputation model. Under model (2.5.2),

$$\mathbb{E}_m(y_k \mid \boldsymbol{v}_k) = f(\boldsymbol{v}_k;\boldsymbol{\beta})$$

and

$$\mathbb{V}_m(y_k \mid \boldsymbol{v}_k) = c_k\sigma^2.$$

The parametric specification assumes that the conditional mean of y_k belongs to a known family of functions indexed by the finite-dimensional parameter $\boldsymbol{\beta}$. For example, under a linear regression model,

$$f(\boldsymbol{v}_k;\boldsymbol{\beta}) = \boldsymbol{v}_k^\top\boldsymbol{\beta}.$$

The parameter $\boldsymbol{\beta}$ is estimated using the respondents, and the fitted model is then used to construct imputed values for the nonrespondents. The performance of a parametric imputation procedure depends on how well the assumed form of $f(\boldsymbol{v}_k;\boldsymbol{\beta})$ represents the relationship between the survey variable and the auxiliary variables. If the parametric model is misspecified, for example because important nonlinearities or interactions are omitted, the resulting imputed estimator may be biased or inefficient.

A more flexible approach is to consider a nonparametric imputation model. In this case, we write

$$y_k = f(\boldsymbol{v}_k) + \varepsilon_k,$$

where the regression function $f(\cdot)$ is unknown and is not assumed to belong to a family indexed by a finite-dimensional parameter vector. Equivalently,

$$\mathbb{E}_m(y_k \mid \boldsymbol{v}_k) = f(\boldsymbol{v}_k).$$

Thus, unlike the parametric case, there is no finite-dimensional parameter $\boldsymbol{\beta}$ that completely characterizes the regression function. Instead, the unknown object to be estimated

is the function $f(\cdot)$ itself. Nonparametric methods estimate this function directly from the observed data and therefore allow for more complex relationships between y_k and $\boldsymbol{v}_k$. This additional flexibility can be useful when the relationship between the survey variable and the auxiliary variables is nonlinear or involves complex interactions that are difficult to specify in advance. However, the increased flexibility generally comes at the cost of more demanding estimation and may require larger sample sizes. In later sections, we will consider imputation procedures based on both parametric and nonparametric models.

2.5.1 Deterministic imputation

We first consider deterministic imputation methods. Under a parametric imputation model, a missing value y_k is replaced by

$$y_k^* = f(\boldsymbol{v}_k; \hat{\boldsymbol{\beta}}_r), \qquad k \in S_m,$$

where $\hat{\boldsymbol{\beta}}_r$ is an estimator of the unknown parameter $\boldsymbol{\beta}$ computed using the respondents. Depending on the model, $\hat{\boldsymbol{\beta}}_r$ may be obtained, for example, by least squares or maximum likelihood.

2.5.1.1 Linear regression imputation

Suppose that the conditional mean of y_k is linear in the auxiliary variables, so that $f(\boldsymbol{v}_k; \boldsymbol{\beta}) = \boldsymbol{v}_k^\top \boldsymbol{\beta}$. Linear regression imputation replaces the missing value y_k by

$$y_k^* = \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS}, \qquad k \in S_m,$$

where

$$\hat{\boldsymbol{\beta}}_{WLS} = \left(\sum_{k \in S_r} \phi_k c_k^{-1} \boldsymbol{v}_k \boldsymbol{v}_k^\top \right)^{-1} \sum_{k \in S_r} \phi_k c_k^{-1} \boldsymbol{v}_k y_k. \tag{2.5.3}$$

The quantities ϕ_k determine the weights used when fitting the imputation model. Under a noninformative sampling design, it is common to use $\phi_k = 1$. When the sampling design is informative, incorporating the sampling weights through $\phi_k = w_k$ may be important to account for the unequal selection probabilities. Linear regression imputation is appropriate when the conditional mean of y_k is reasonably well represented by a linear function of the

auxiliary variables. Ratio imputation and mean imputation can be viewed as particular cases of linear regression imputation.

2.5.1.2 Ratio imputation

Ratio imputation is obtained by considering a single scalar auxiliary variable v_k and a regression model through the origin,

$$y_k = \beta v_k + \varepsilon_k.$$

Setting $c_k = v_k$ in (2.5.3) gives

$$\hat{\beta}_{WLS} = \frac{\sum_{k \in S_r} \phi_k y_k}{\sum_{k \in S_r} \phi_k v_k}.$$

Equivalently,

$$\hat{\beta}_{WLS} = \frac{\bar{y}_{r,\phi}}{\bar{v}_{r,\phi}},$$

where

$$\bar{y}_{r,\phi} = \frac{\sum_{k \in S_r} \phi_k y_k}{\sum_{k \in S_r} \phi_k}, \qquad \bar{v}_{r,\phi} = \frac{\sum_{k \in S_r} \phi_k v_k}{\sum_{k \in S_r} \phi_k}.$$

The imputed value is therefore

$$y_k^* = \hat{\beta}_{WLS} v_k = \frac{\bar{y}_{r,\phi}}{\bar{v}_{r,\phi}} v_k, \qquad k \in S_m.$$

Ratio imputation is particularly appropriate when the relationship between y_k and v_k is approximately linear and passes through the origin.

2.5.1.3 Mean imputation

Mean imputation is another particular case of linear regression imputation. It is obtained by considering an intercept-only model, that is,

$$\boldsymbol{v}_k = 1 \qquad \text{and} \qquad c_k = 1.$$

The imputed value is then

$$y_k^* = \bar{y}_{r,\phi}, \qquad k \in S_m,$$

where

$$\bar{y}_{r,\phi} = \frac{\sum_{k \in S_r} \phi_k y_k}{\sum_{k \in S_r} \phi_k}.$$

In the unweighted case, $\phi_k = 1$, this reduces to the usual respondent mean. Mean imputation relies on the restrictive assumption that the conditional mean of y_k is constant across units. In particular, it does not make use of auxiliary information that could explain differences in the variable of interest.

2.5.1.4 Nearest-neighbour imputation

Nearest-neighbour imputation is a nonparametric donor imputation method. For each nonrespondent $k \in S_m$, a respondent whose auxiliary information is as close as possible to that of unit k is selected as a donor. Let

$$l(k) = \arg\min_{i \in S_r} D(\boldsymbol{v}_k, \boldsymbol{v}_i)$$

denote the index of the nearest respondent to unit k. The missing value is then imputed by

$$y_k^* = y_{l(k)}, \qquad k \in S_m.$$

A general distance between two vectors of auxiliary variables can be defined as

$$D(\boldsymbol{v}_i, \boldsymbol{v}_j) = \left\{ \sum_{h=1}^{p} a_h |v_{ih} - v_{jh}|^b \right\}^{1/b},$$

where $b \geq 1$ and $a_h \geq 0$ is a weight reflecting the relative importance of the h-th auxiliary variable. When $b = 1$, this gives a weighted L_1 distance, whereas $b = 2$ gives a weighted Euclidean distance.

When auxiliary variables are measured on very different scales, they are often standardized before computing the distance so that variables with larger numerical scales do not dominate the matching procedure. Since the imputed value is an observed value from a respondent, nearest-neighbour imputation preserves the support of the original variable. This property makes it particularly useful for categorical variables and other variables for

which only certain values are admissible.

2.5.2 Random imputation

Random imputation methods introduce an additional random mechanism when constructing the imputed values. Thus, two applications of the same imputation procedure to the same observed data may lead to different completed datasets. For regression-based methods, randomness can be introduced by adding a randomly generated residual to the fitted value. For donor methods, randomness can instead be introduced by randomly selecting a donor from a set of eligible respondents.

2.5.2.1 Random linear regression imputation

Random linear regression imputation extends deterministic regression imputation by adding a random residual to the predicted value. Under the regression model, the imputed value is

$$y_k^* = \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS} + \hat{\sigma}\sqrt{c_k}\, e_k^*, \qquad k \in S_m,$$

where e_k^* is a random variable with mean zero and, typically, variance equal to one. If a normal error distribution is assumed, one possibility is to generate

$$e_k^* \sim \mathcal{N}(0,1).$$

Alternatively, to avoid imposing a parametric distribution on the model errors, the random residuals can be generated from the empirical distribution of the residuals observed among the respondents. Define the standardized residual associated with respondent j by

$$\tilde{e}_j = \frac{y_j - \boldsymbol{v}_j^\top \hat{\boldsymbol{\beta}}_{WLS}}{\hat{\sigma}\sqrt{c_j}}, \qquad j \in S_r.$$

These residuals may be centered as

$$e_j = \tilde{e}_j - \frac{\sum_{l \in S_r} \phi_l \tilde{e}_l}{\sum_{l \in S_r} \phi_l}.$$

For each nonrespondent k, a residual e_k^* is then selected with replacement from the set

$\{e_j : j \in S_r\}$. For example, respondent j may be selected with probability

$$P(e_k^* = e_j) = \frac{\phi_j}{\sum_{l \in S_r} \phi_l}, \qquad j \in S_r.$$

2.5.2.2 Random hot-deck imputation

Random hot-deck imputation is a commonly used random donor imputation method. For each nonrespondent $k \in S_m$, a donor l is randomly selected from the set of respondents S_r, and the missing value is replaced by the donor value: $y_k^* = y_l$. For example, donor l may be selected with probability

$$P(l \text{ is selected for } k) = \frac{\phi_l}{\sum_{j \in S_r} \phi_j}, \qquad l \in S_r.$$

Random hot-deck imputation can also be viewed as a particular case of random regression imputation based on an intercept-only model. Indeed, let

$$\bar{y}_{r,\phi} = \frac{\sum_{j \in S_r} \phi_j y_j}{\sum_{j \in S_r} \phi_j}$$

and define the respondent residuals by

$$e_j = y_j - \bar{y}_{r,\phi}.$$

If one of these residuals is randomly selected, say e_l, then

$$\begin{aligned} y_k^* &= \bar{y}_{r,\phi} + e_l \\ &= \bar{y}_{r,\phi} + (y_l - \bar{y}_{r,\phi}) \\ &= y_l. \end{aligned}$$

Thus, random hot-deck imputation preserves observed values and introduces randomness through the selection of the donor. This feature is particularly useful for categorical variables, since the imputed values automatically belong to the set of observed categories.

2.5.3 Bias of the imputed estimator

We now study the bias of the imputed estimator $\hat{\mu}_I$. Throughout this section, we focus on deterministic imputation methods. The total error of the imputed estimator can be decomposed as

$$\hat{\mu}_I - \mu = \underbrace{(\hat{\mu}_\pi - \mu)}_{\text{sampling error}} + \underbrace{(\hat{\mu}_I - \hat{\mu}_\pi)}_{\text{nonresponse error}} . \tag{2.5.4}$$

The first term in (2.5.4) is the error that would be present even under complete response. It results from observing only a sample of the finite population. The second term is the additional error introduced by nonresponse and the imputation procedure. In the absence of nonresponse, $\hat{\mu}_I = \hat{\mu}_\pi$ and the second term is equal to zero.

To study the bias, we use the mpq framework, in which the expectation is taken with respect to the model m generating the survey variable, the sampling design p, and the nonresponse mechanism q. We define

$$\mathbb{E}_{mpq}(\cdot) = \mathbb{E}_m \mathbb{E}_p \mathbb{E}_q(\cdot).$$

The mpq-bias of $\hat{\mu}_I$ is therefore

$$\text{Bias}_{mpq}(\hat{\mu}_I) = \mathbb{E}_{mpq}(\hat{\mu}_I - \mu).$$

Using (2.5.4),

$$\begin{aligned} \text{Bias}_{mpq}(\hat{\mu}_I) &= \mathbb{E}_{mpq}\left(\hat{\mu}_\pi - \mu\right) + \mathbb{E}_{mpq}\left(\hat{\mu}_I - \hat{\mu}_\pi\right) \\ &= \mathbb{E}_{mpq}\left(\hat{\mu}_I - \hat{\mu}_\pi\right), \end{aligned}$$

since the Horvitz–Thompson estimator is design-unbiased:

$$\mathbb{E}_p\left(\hat{\mu}_\pi - \mu\right) = 0.$$

Consequently, when studying the bias induced by nonresponse, it is sufficient to study the difference $\hat{\mu}_I - \hat{\mu}_\pi$. For any deterministic imputation procedure,

$$\hat{\mu}_I = \frac{1}{N}\left(\sum_{k \in S_r} w_k y_k + \sum_{k \in S_m} w_k y_k^*\right).$$

It follows that

$$\hat{\mu}_I - \hat{\mu}_\pi = \frac{1}{N} \sum_{k \in S_m} w_k \left(y_k^* - y_k \right). \tag{2.5.5}$$

Expression (2.5.5) is useful because it shows directly that the bias of the imputed estimator depends on how well the imputed values reproduce, on average, the missing values. For a parametric imputation procedure,

$$y_k^* = f(\boldsymbol{v}_k; \hat{\boldsymbol{\beta}}_r), \qquad k \in S_m,$$

and hence

$$\hat{\mu}_I - \hat{\mu}_\pi = -\frac{1}{N} \sum_{k \in S_m} w_k \left\{ y_k - f(\boldsymbol{v}_k; \hat{\boldsymbol{\beta}}_r) \right\}. \tag{2.5.6}$$

In general, the fact that the imputation model is correctly specified does not automatically imply exact unbiasedness in finite samples, since the unknown model parameters must be estimated from the respondents. Under suitable regularity conditions, however, a correctly specified imputation model can lead to a negligible bias.

2.5.3.1 Linear regression imputation

Consider the linear regression model

$$y_k = \boldsymbol{v}_k^\top \boldsymbol{\beta} + \varepsilon_k, \tag{2.5.7}$$

with

$$\mathbb{E}_m(\varepsilon_k \mid \boldsymbol{v}_k) = 0,$$

and suppose that the model errors are independent across population units.

Recall from (2.5.3) that the parameter $\boldsymbol{\beta}$ is estimated from the respondents by

$$\hat{\boldsymbol{\beta}}_{WLS} = \left(\sum_{k \in S_r} \phi_k c_k^{-1} \boldsymbol{v}_k \boldsymbol{v}_k^\top \right)^{-1} \sum_{k \in S_r} \phi_k c_k^{-1} \boldsymbol{v}_k y_k.$$

The imputed values are then

$$y_k^* = \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS}, \qquad k \in S_m.$$

Proposition 2.3. *Suppose that the linear imputation model* (2.5.7) *is correctly specified, the response mechanism is MAR given* $\boldsymbol{v}_k$*, the sampling design is noninformative with respect to the model errors, and the matrix involved in the definition of* $\hat{\boldsymbol{\beta}}_{WLS}$ *is nonsingular. Then the linear regression imputed estimator is mpq-unbiased for* μ*.*

Proof. Under the linear regression model,

$$\hat{\boldsymbol{\beta}}_{WLS} = \boldsymbol{\beta} + \left(\sum_{k \in S_r} \phi_k c_k^{-1} \boldsymbol{v}_k \boldsymbol{v}_k^\top \right)^{-1} \sum_{k \in S_r} \phi_k c_k^{-1} \boldsymbol{v}_k \varepsilon_k.$$

Therefore, conditionally on the selected sample, the response indicators, and the auxiliary variables, $\mathbb{E}_m \left(\hat{\boldsymbol{\beta}}_{WLS} \right) = \boldsymbol{\beta}$. For a nonrespondent $k \in S_m$,

$$\begin{aligned} \mathbb{E}_m \left[y_k - \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS} \right] &= \boldsymbol{v}_k^\top \boldsymbol{\beta} - \boldsymbol{v}_k^\top \mathbb{E}_m \left(\hat{\boldsymbol{\beta}}_{WLS} \right) \\ &= 0. \end{aligned}$$

Using (2.5.6), it follows that

$$\mathbb{E}_m \left(\hat{\mu}_I - \hat{\mu}_\pi \right) = 0.$$

Consequently,

$$\mathbb{E}_{mpq} \left(\hat{\mu}_I - \hat{\mu}_\pi \right) = 0,$$

and, since $\hat{\mu}_\pi$ is design-unbiased,

$$\mathbb{E}_{mpq} (\hat{\mu}_I - \mu) = 0.$$

□

Thus, under the stated assumptions, linear regression imputation provides an exactly unbiased estimator of the population mean in the *mpq* framework.

2.5.3.2 Nearest-neighbour imputation

The situation is different for nearest-neighbour imputation. Recall that, for each nonrespondent k, the imputed value is

$$y_k^* = y_{l(k)},$$

where $l(k) \in S_r$ denotes the respondent whose auxiliary information is closest to that of unit k.

Suppose that

$$y_k = f(\boldsymbol{v}_k) + \varepsilon_k, \qquad \mathbb{E}_m(\varepsilon_k \mid \boldsymbol{v}_k) = 0.$$

Conditionally on the sample, the response indicators, and the auxiliary variables,

$$\mathbb{E}_m\left(y_k^* - y_k\right) = f(\boldsymbol{v}_{l(k)}) - f(\boldsymbol{v}_k).$$

Therefore,

$$\mathbb{E}_m\left(\hat{\mu}_I - \hat{\mu}_\pi\right) = \frac{1}{N} \sum_{k \in S_m} w_k \left\{ f(\boldsymbol{v}_{l(k)}) - f(\boldsymbol{v}_k) \right\}.$$

The bias of nearest-neighbour imputation is therefore directly related to the quality of the matches between nonrespondents and their donors. If $\boldsymbol{v}_{l(k)}$ is close to $\boldsymbol{v}_k$ and the function $f(\cdot)$ is sufficiently smooth, the difference $f(\boldsymbol{v}_{l(k)}) - f(\boldsymbol{v}_k)$ will be small. However, the quality of nearest-neighbour matching generally deteriorates as the number of auxiliary variables increases. Consequently, the bias properties of nearest-neighbour imputation depend on several factors, including the dimension of the auxiliary information, the smoothness of $f(\cdot)$, the distribution of the auxiliary variables, and the distance used to define the nearest neighbour. For this reason, no general finite-sample unbiasedness result can be stated for nearest-neighbour imputation. Its asymptotic bias requires a separate analysis under appropriate regularity conditions.

2.6 Variance of the imputed estimator

We now consider the variance of the imputed estimator. In this section, we restrict attention to deterministic parametric imputation methods. More precisely, for each nonrespondent $k \in S_m$, the missing value is replaced by

$$y_k^* = f(\boldsymbol{v}_k; \hat{\boldsymbol{\beta}}_r),$$

where $\hat{\boldsymbol{\beta}}_r$ is an estimator of the finite-dimensional parameter $\boldsymbol{\beta}$ based on the respondents. Variance estimation after imputation is more involved than under complete response because the imputed estimator is affected by several sources of variability. In addition to the

variability resulting from the selection of the sample, one must account for the additional variability associated with nonresponse and with the estimation of the imputation model. Throughout this section, we assume that the parametric imputation model is correctly specified and that the resulting imputed estimator is unbiased under the inferential framework considered. This allows us to focus on the variance components rather than on the mean squared error.

We present two approaches for decomposing the total variance of $\hat{\mu}_I$: the two-phase approach and the reverse approach.

2.6.1 The two-phase approach

Under the two-phase approach, nonresponse is viewed as a second phase of selection. The first phase consists of selecting the sample S from the finite population U, while the second phase determines which sampled units respond. Schematically,

$$U \longrightarrow S \longrightarrow S_r,$$

where $S_m = S \setminus S_r$ denotes the set of nonrespondents.

Following Särndal (1992), the error of the imputed estimator is decomposed as

$$\hat{\mu}_I - \mu = (\hat{\mu}_\pi - \mu) + (\hat{\mu}_I - \hat{\mu}_\pi)\,, \tag{2.6.1}$$

where $\hat{\mu}_\pi$ is the Horvitz–Thompson estimator that would be computed under complete response. The first term in (2.6.1) is the sampling error, whereas the second term is the additional error resulting from nonresponse. The total variance can be decomposed as

$$\mathbb{V}_{\text{tot,tp}}(\hat{\mu}_I) = \mathbb{V}_{\text{sam}}(\hat{\mu}_I) + \mathbb{V}_{\text{nr}}(\hat{\mu}_I) + 2\mathbb{V}_{\text{mix}}(\hat{\mu}_I),$$

where

$$\mathbb{V}_{\text{sam}}(\hat{\mu}_I) = \mathbb{E}_m \left\{\mathbb{V}_p(\hat{\mu}_\pi)\right\},$$
$$\mathbb{V}_{\text{nr}}(\hat{\mu}_I) = \mathbb{E}_p\mathbb{E}_q \left\{\mathbb{V}_m \left(\hat{\mu}_I - \hat{\mu}_\pi\right)\right\},$$

and

$$\mathbb{V}_{\text{mix}}(\hat{\mu}_I) = \mathbb{E}_p\mathbb{E}_q \left\{\text{Cov}_m \left(\hat{\mu}_\pi - \mu, \hat{\mu}_I - \hat{\mu}_\pi\right)\right\}.$$

The first component, $\mathbb{V}_{\text{sam}}(\hat{\mu}_I)$, represents the variability that would be present even under complete response. It is therefore the usual sampling variance associated with the Horvitz–Thompson estimator, averaged with respect to the imputation model. The second component, $\mathbb{V}_{\text{nr}}(\hat{\mu}_I)$, measures the additional variability caused by nonresponse and imputation. It depends, among other things, on the response rate, the relationship between the survey variable and the auxiliary variables, and the precision with which the parameters of the imputation model are estimated. The third component, $\mathbb{V}_{\text{mix}}(\hat{\mu}_I)$, is a covariance term between the sampling error and the nonresponse error. Unlike the first two components, it is not itself a variance and may be either positive or negative. Thus, under the two-phase approach, variance estimation requires accounting for the usual sampling variability, the additional variability due to nonresponse, and the dependence between these two sources of error.

2.6.2 The reverse approach

An alternative way of decomposing the variance is the reverse approach (Fay, 1991; Shao and Steel, 1999). This approach requires strong invariance of the response mechanism (Beaumont and Haziza, 2016). Under this property, the response indicators can be conceptually generated for all units in the finite population before the sample is selected. In particular, the response probability attached to a given population unit does not depend on which particular sample happens to be selected. Let $U_r = \{k \in U : r_k = 1\}$ and $U_m = \{k \in U : r_k = 0\}$ denote, respectively, the sets of potential respondents and nonrespondents in the finite population. Under the reverse approach, the randomization can then be viewed schematically as

$$U \longrightarrow (U_r, U_m) \longrightarrow S.$$

This representation does not change the actual data collection process. It is a theoretical device that allows the order of the sampling and response mechanisms to be reversed when deriving the variance.

The total variance under the reverse approach can be written as

$$\mathbb{V}_{\text{tot,ra}}(\hat{\mu}_I) = \mathbb{V}_1(\hat{\mu}_I) + \mathbb{V}_2(\hat{\mu}_I),$$

where

$$\mathbb{V}_1(\hat{\mu}_I) = \mathbb{E}_q \mathbb{E}_m \left\{ \mathbb{V}_p(\hat{\mu}_I) \right\},$$

and

$$\mathbb{V}_2(\hat{\mu}_I) = \mathbb{E}_q \left[\mathbb{V}_m \left\{ \mathbb{E}_p(\hat{\mu}_I - \mu) \right\} \right].$$

The reverse approach will be particularly useful later in this thesis when considering imputation procedures based on machine learning methods. Also, an important simplification occurs when the sampling fraction is negligible. In this case, the contribution of $\mathbb{V}_2(\hat{\mu}_I)$ to the total variance is asymptotically negligible. Therefore, when the sampling fraction is negligible, the total variance can be approximated by

$$\mathbb{V}_{\text{tot,ra}}(\hat{\mu}_I) \simeq \mathbb{V}_1(\hat{\mu}_I).$$

This result is particularly useful because variance estimation can then focus primarily on the sampling variance component. When the sampling fraction is not negligible, however, the second component cannot in general be ignored; see Shao and Steel (1999).

2.7 The Oracle estimator and its properties

We now introduce the Oracle imputed estimator, which will serve as a benchmark for the imputation procedures considered later in this thesis. The survey variable y is assumed to satisfy

$$y_k = f(\boldsymbol{v}_k) + \varepsilon_k,$$

where

$$\mathbb{E}_m(\varepsilon_k \mid \boldsymbol{v}_k) = 0$$

and

$$\mathbb{V}_m(\varepsilon_k \mid \boldsymbol{v}_k) = \sigma^2.$$

We also assume that the model errors are independent across population units, that is,

$$\text{Cov}_m(\varepsilon_k, \varepsilon_l \mid \boldsymbol{v}_k, \boldsymbol{v}_l) = 0, \qquad k \neq l.$$

The Oracle setting assumes that the conditional mean function $f(\cdot)$ is known. Thus, for each nonrespondent $k \in S_m$, the missing value is replaced by

$$y_k^* = f(\boldsymbol{v}_k).$$

The resulting Oracle imputed estimator of the population mean is

$$\hat{\mu}_{\text{oracle}} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k f(\boldsymbol{v}_k)\right\}. \tag{2.7.1}$$

Since the true conditional mean function $f(\cdot)$ is generally unknown, the Oracle estimator cannot be implemented in practice. Nevertheless, it provides a useful benchmark for understanding the properties of feasible imputation procedures in which $f(\cdot)$ must be estimated.

Proposition 2.4. *Under the assumptions stated above, the Oracle imputed estimator $\hat{\mu}_{\text{oracle}}$ is mpq-unbiased for the population mean μ.*

Proof. Recall that the Horvitz–Thompson estimator under complete response is

$$\hat{\mu}_\pi = \frac{1}{N}\sum_{k\in S} w_k y_k.$$

From (2.7.1),

$$\hat{\mu}_{\text{oracle}} - \hat{\mu}_\pi = -\frac{1}{N}\sum_{k\in S} w_k(1-r_k)\left\{y_k - f(\boldsymbol{v}_k)\right\}. \tag{2.7.2}$$

Under the MAR assumption and the noninformative sampling assumption, conditioning on the sample, the response indicators, and the auxiliary variables does not change the conditional mean of the model error. Therefore,

$$\mathbb{E}_m\left[y_k - f(\boldsymbol{v}_k)\right] = 0.$$

It follows from (2.7.2) that

$$\mathbb{E}_m\left(\hat{\mu}_{\text{oracle}} - \hat{\mu}_\pi\right) = 0.$$

Hence,

$$\mathbb{E}_{mpq}\left(\hat{\mu}_{\text{oracle}} - \hat{\mu}_\pi\right) = 0.$$

Since the Horvitz–Thompson estimator is design-unbiased,

$$\mathbb{E}_{mpq}\left(\hat{\mu}_{\pi} - \mu\right) = 0.$$

As a result,

$$\mathbb{E}_{mpq}\left(\hat{\mu}_{\text{oracle}} - \mu\right) = 0.$$

□

2.7.1 Variance of the Oracle estimator

We now derive the variance of the Oracle estimator using the two approaches introduced in the previous section. In this section, we treat σ^2 as known.

2.7.1.1 Two-phase approach

Recall that the total variance under the two-phase approach can be written as

$$\mathbb{V}_{\text{tot,tp}}\left(\hat{\mu}_{\text{oracle}}\right) = \mathbb{V}_{\text{sam}}\left(\hat{\mu}_{\text{oracle}}\right) + \mathbb{V}_{\text{nr}}\left(\hat{\mu}_{\text{oracle}}\right) + 2\mathbb{V}_{\text{mix}}\left(\hat{\mu}_{\text{oracle}}\right).$$

The sampling component is

$$\mathbb{V}_{\text{sam}}\left(\hat{\mu}_{\text{oracle}}\right) = \mathbb{E}_{m}\left\{\mathbb{V}_{p}(\hat{\mu}_{\pi})\right\}.$$

Using the variance expression of the Horvitz–Thompson estimator and the model assumptions, we obtain

$$\mathbb{V}_{\text{sam}}\left(\hat{\mu}_{\text{oracle}}\right) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{f(\boldsymbol{v}_k)}{\pi_k}\frac{f(\boldsymbol{v}_l)}{\pi_l} + \frac{\sigma^2}{N^2}\sum_{k\in U}(w_k - 1). \tag{2.7.3}$$

Provided that $\pi_{kl} > 0$, an unbiased estimator of (2.7.3) is

$$\widehat{\mathbb{V}}_{\text{sam}}\left(\hat{\mu}_{\text{oracle}}\right) = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{f(\boldsymbol{v}_k)}{\pi_k}\frac{f(\boldsymbol{v}_l)}{\pi_l} + \frac{\sigma^2}{N^2}\sum_{k\in S}w_k(w_k - 1).$$

Next, the nonresponse component is

$$\begin{aligned}\mathbb{V}_{\mathrm{nr}}\left(\hat{\mu}_{\mathrm{oracle}}\right) &= \mathbb{E}_p\mathbb{E}_q\left\{\mathbb{V}_m\left(\hat{\mu}_{\mathrm{oracle}}-\hat{\mu}_\pi\right)\right\} \\ &= \frac{\sigma^2}{N^2}\mathbb{E}_p\mathbb{E}_q\left\{\sum_{k\in S} w_k^2(1-r_k)\right\} \\ &= \frac{\sigma^2}{N^2}\sum_{k\in U} w_k\left\{1-p(\boldsymbol{v}_k)\right\}. \end{aligned} \tag{2.7.4}$$

An unbiased estimator of (2.7.4) is

$$\widehat{\mathbb{V}}_{\mathrm{nr}}\left(\hat{\mu}_{\mathrm{oracle}}\right) = \frac{\sigma^2}{N^2}\sum_{k\in S} w_k^2(1-r_k).$$

Finally, the mixed component is

$$\begin{aligned}\mathbb{V}_{\mathrm{mix}}\left(\hat{\mu}_{\mathrm{oracle}}\right) &= \mathbb{E}_p\mathbb{E}_q\left\{\mathrm{Cov}_m\left(\hat{\mu}_\pi-\mu,\hat{\mu}_{\mathrm{oracle}}-\hat{\mu}_\pi\right)\right\} \\ &= -\frac{\sigma^2}{N^2}\sum_{k\in U}\left\{1-p(\boldsymbol{v}_k)\right\}(w_k-1). \end{aligned} \tag{2.7.5}$$

Since $0 < p(\boldsymbol{v}_k) \leq 1$ and $w_k \geq 1$, the mixed component is nonpositive. An unbiased estimator of (2.7.5) is

$$\widehat{\mathbb{V}}_{\mathrm{mix}}\left(\hat{\mu}_{\mathrm{oracle}}\right) = -\frac{\sigma^2}{N^2}\sum_{k\in S} w_k(w_k-1)(1-r_k).$$

Combining the three components, we obtain

$$\widehat{\mathbb{V}}_{\mathrm{tot,tp}}\left(\hat{\mu}_{\mathrm{oracle}}\right) = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{f(\boldsymbol{v}_k)}{\pi_k}\frac{f(\boldsymbol{v}_l)}{\pi_l} + \frac{\sigma^2}{N^2}\left\{\sum_{k\in S_m} w_k + \sum_{k\in S_r} w_k(w_k-1)\right\}.$$

2.7.1.2 Reverse approach

We now derive the variance of the Oracle estimator under the reverse approach. Recall that the completed value for unit k is

$$\tilde{y}_k = r_k y_k + (1 - r_k) f(\boldsymbol{v}_k).$$

The Oracle estimator can therefore be written as

$$\hat{\mu}_{\text{oracle}} = \frac{1}{N} \sum_{k \in S} w_k \tilde{y}_k.$$

The total variance under the reverse approach can then be decomposed as

$$\mathbb{V}_{\text{tot,ra}} \left(\hat{\mu}_{\text{oracle}} \right) = \mathbb{V}_1 \left(\hat{\mu}_{\text{oracle}} \right) + \mathbb{V}_2 \left(\hat{\mu}_{\text{oracle}} \right),$$

where

$$\mathbb{V}_1 \left(\hat{\mu}_{\text{oracle}} \right) = \mathbb{E}_q \mathbb{E}_m \left\{ \mathbb{V}_p \left(\hat{\mu}_{\text{oracle}} \right) \right\},$$

and

$$\mathbb{V}_2 \left(\hat{\mu}_{\text{oracle}} \right) = \mathbb{E}_q \left[\mathbb{V}_m \left\{ \mathbb{E}_p \left(\hat{\mu}_{\text{oracle}} - \mu \right) \right\} \right].$$

The first component represents the sampling variability of the estimator, averaged with respect to the imputation model and the nonresponse mechanism. Conditionally on the completed population values,

$$\mathbb{V}_p \left(\hat{\mu}_{\text{oracle}} \right) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{\tilde{y}_k}{\pi_k} \frac{\tilde{y}_l}{\pi_l}.$$

Taking expectations with respect to the model and the nonresponse mechanism gives

$$\mathbb{V}_1 \left(\hat{\mu}_{\text{oracle}} \right) = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{f(\boldsymbol{v}_k)}{\pi_k} \frac{f(\boldsymbol{v}_l)}{\pi_l} + \frac{\sigma^2}{N^2} \sum_{k \in U} p(\boldsymbol{v}_k)(w_k - 1). \tag{2.7.6}$$

An unbiased estimator of $\mathbb{V}_1$ is obtained by applying the usual Horvitz–Thompson variance

estimator to the completed values:

$$\widehat{\mathbb{V}}_1\left(\hat{\mu}_{\text{oracle}}\right) = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{\tilde{y}_k}{\pi_k}\frac{\tilde{y}_l}{\pi_l},$$

provided that $\pi_{kl} > 0$. We now consider the second component. We have

$$\begin{aligned}\mathbb{E}_p\left(\hat{\mu}_{\text{oracle}} - \mu\right) &= \frac{1}{N}\sum_{k\in U}\left(\tilde{y}_k - y_k\right)\\ &= -\frac{1}{N}\sum_{k\in U}(1-r_k)\left\{y_k - f(\boldsymbol{v}_k)\right\}.\end{aligned}$$

Therefore,

$$\mathbb{V}_2\left(\hat{\mu}_{\text{oracle}}\right) = \frac{\sigma^2}{N^2}\sum_{k\in U}\left\{1 - p(\boldsymbol{v}_k)\right\}. \tag{2.7.7}$$

An unbiased estimator of $\mathbb{V}_2$ is

$$\widehat{\mathbb{V}}_2\left(\hat{\mu}_{\text{oracle}}\right) = \frac{\sigma^2}{N^2}\sum_{k\in S} w_k(1-r_k).$$

Combining (2.7.6) and (2.7.7), the total variance of the Oracle estimator under the reverse approach is

$$\mathbb{V}_{\text{tot,ra}}\left(\hat{\mu}_{\text{oracle}}\right) = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{f(\boldsymbol{v}_k)}{\pi_k}\frac{f(\boldsymbol{v}_l)}{\pi_l} + \frac{\sigma^2}{N^2}\sum_{k\in U}\left[\left\{1 - p(\boldsymbol{v}_k)\right\} + p(\boldsymbol{v}_k)(w_k - 1)\right].$$

The corresponding variance estimator is simply

$$\begin{aligned}\widehat{\mathbb{V}}_{\text{tot,ra}}(\hat{\mu}_{\text{oracle}}) &= \widehat{\mathbb{V}}_1(\hat{\mu}_{\text{oracle}}) + \widehat{\mathbb{V}}_2(\hat{\mu}_{\text{oracle}})\\ &= \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{\tilde{y}_k}{\pi_k}\frac{\tilde{y}_l}{\pi_l} + \frac{\sigma^2}{N^2}\sum_{k\in S} w_k(1-r_k).\end{aligned}$$

An important feature of the reverse approach is that the second variance component becomes negligible when the sampling fraction is small. Indeed, since $0 < 1 - p(\boldsymbol{v}_k) \leq 1$, we

have

$$\mathbb{V}_2\left(\hat{\mu}_{\text{oracle}}\right) \leq \frac{\sigma^2}{N}.$$

Hence, if

$$\frac{n}{N} \longrightarrow 0,$$

then

$$\mathbb{V}_2\left(\hat{\mu}_{\text{oracle}}\right) = o\left(\frac{1}{n}\right).$$

Under standard sampling-design conditions, $\mathbb{V}_1(\hat{\mu}_{\text{oracle}})$ is typically of order n^{-1}. Therefore,

$$\mathbb{V}_{\text{tot,ra}}\left(\hat{\mu}_{\text{oracle}}\right) = \mathbb{V}_1\left(\hat{\mu}_{\text{oracle}}\right)\{1 + o(1)\}.$$

2.7.2 Root-n consistency of the Oracle estimator

Before studying the asymptotic properties of the Oracle estimator, it is useful to clarify the asymptotic framework. Since the population is finite, asymptotic results cannot be obtained by letting the sample size increase while keeping the population fixed. Instead, we consider a sequence of finite populations $U_N = \{1, \ldots, N\}$, with corresponding sample sizes $n = n_N$, such that

$$N \longrightarrow \infty \qquad \text{and} \qquad n \longrightarrow \infty.$$

We also assume that the sampling fraction satisfies

$$\frac{n}{N} \longrightarrow \pi^*, \qquad \pi^* \in [0, 1).$$

The case $\pi^* > 0$ corresponds to a non-negligible sampling fraction, whereas $\pi^* = 0$ corresponds to a negligible sampling fraction. All asymptotic statements in the remainder of this thesis are understood along such a sequence of finite populations and sampling designs.

Proposition 2.5. *Suppose that the assumptions of Proposition 2.4 hold and that standard regularity conditions on the sampling design and the finite population ensure that*

$$\mathbb{V}_1\left(\hat{\mu}_{\text{oracle}}\right) = O\left(\frac{1}{n}\right).$$

Then,

$$\hat{\mu}_{\text{oracle}} - \mu = O_p\left(n^{-1/2}\right).$$

Hence, the Oracle estimator is root-n consistent for μ.

Proof. Under the reverse approach,

$$\mathbb{V}_{\text{tot,ra}}\left(\hat{\mu}_{\text{oracle}}\right) = \mathbb{V}_1\left(\hat{\mu}_{\text{oracle}}\right) + \mathbb{V}_2\left(\hat{\mu}_{\text{oracle}}\right),$$

where

$$\mathbb{V}_2\left(\hat{\mu}_{\text{oracle}}\right) = \frac{\sigma^2}{N^2}\sum_{k\in U}\left\{1 - p(\boldsymbol{v}_k)\right\}.$$

Since $0 < p(\boldsymbol{v}_k) \leq 1$,

$$\mathbb{V}_2\left(\hat{\mu}_{\text{oracle}}\right) \leq \frac{\sigma^2}{N}.$$

Because $n \leq N$, it follows that

$$\mathbb{V}_2\left(\hat{\mu}_{\text{oracle}}\right) = O\left(\frac{1}{n}\right).$$

Therefore,

$$\mathbb{V}_{\text{tot,ra}}\left(\hat{\mu}_{\text{oracle}}\right) = O\left(\frac{1}{n}\right).$$

Moreover, from Proposition 2.4,

$$\mathbb{E}_{mpq}\left(\hat{\mu}_{\text{oracle}} - \mu\right) = 0.$$

Hence,

$$\mathbb{E}_{mpq}\left[\left(\hat{\mu}_{\text{oracle}} - \mu\right)^2\right] = \mathbb{V}_{\text{tot,ra}}\left(\hat{\mu}_{\text{oracle}}\right) = O\left(\frac{1}{n}\right).$$

By Chebyshev's inequality,

$$\hat{\mu}_{\text{oracle}} - \mu = O_p\left(n^{-1/2}\right).$$

□

If, in addition, the sampling fraction is negligible, that is,

$$\frac{n}{N} \longrightarrow 0,$$

then

$$\mathbb{V}_2\left(\hat{\mu}_{\text{oracle}}\right) = o\left(\frac{1}{n}\right).$$

In this case, the first-order asymptotic variance of the Oracle estimator is entirely determined by $\mathbb{V}_1(\hat{\mu}_{\text{oracle}})$.

2.8 The linear regression imputed estimator and its properties

Recall that, under linear regression imputation, the missing value for unit $k \in S_m$ is replaced by

$$y_k^* = \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS},$$

where $\hat{\boldsymbol{\beta}}_{WLS}$ is defined in (2.5.3). The corresponding imputed estimator of the population mean is

$$\hat{\mu}_{LRI} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS}\right\}.$$

As shown previously, under a correctly specified linear imputation model, the MAR assumption, and a noninformative sampling design, $\hat{\mu}_{LRI}$ is mpq-unbiased for the population mean. We therefore focus here on variance estimation.

2.8.1 Variance of the linear regression imputed estimator

2.8.1.1 Two-phase approach

We first consider the sampling component of the variance. Recall that, under complete response, the usual design-unbiased variance estimator of the Horvitz–Thompson estimator

would be

$$\widehat{\mathbb{V}}_p\left(\hat{\mu}_\pi\right) = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \tag{2.8.1}$$

Thus, (2.8.1) is the variance estimator that would have been available if there had been no nonresponse.

In the presence of nonresponse, the values y_k are not observed for $k \in S_m$. Define the completed value

$$\tilde{y}_k = r_k y_k + (1 - r_k)\boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS}.$$

A natural, but naive, approach consists of replacing y_k by $\tilde{y}_k$ in the complete-response variance estimator (2.8.1). This gives

$$\widehat{\mathbb{V}}_{\text{naive}}\left(\hat{\mu}_{LRI}\right) = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{\tilde{y}_k}{\pi_k}\frac{\tilde{y}_l}{\pi_l}. \tag{2.8.2}$$

The difficulty with (2.8.2) is that the imputed values are treated as if they were observed values. In particular, for a nonrespondent, the deterministic imputed value $\boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS}$ contains the fitted part of y_k but does not reproduce its residual variability. Consequently, the completed values are generally less variable than the values that would have been observed under full response. The naive variance estimator therefore tends to underestimate the variance that would have been estimated from the complete sample.

This loss of variability can be quantified by comparing the complete-response variance estimator with the naive variance estimator. Following the conditional approach of Beaumont and Bocci (2009), define

$$\begin{aligned}\mathbb{V}_{\text{diff}}\left(\hat{\mu}_{LRI}\right) &= \mathbb{E}_m\left[\widehat{\mathbb{V}}_p\left(\hat{\mu}_\pi\right) - \widehat{\mathbb{V}}_{\text{naive}}\left(\hat{\mu}_{LRI}\right) \mid \boldsymbol{I}, \boldsymbol{r}, \mathbf{V}, \boldsymbol{y}_r\right] \\ &= \mathbb{E}_m\left[\widehat{\mathbb{V}}_p\left(\hat{\mu}_\pi\right) \mid \boldsymbol{I}, \boldsymbol{r}, \mathbf{V}, \boldsymbol{y}_r\right] - \widehat{\mathbb{V}}_{\text{naive}}\left(\hat{\mu}_{LRI}\right),\end{aligned}$$

where $\boldsymbol{y}_r$ denotes the vector of observed values for the respondents. The second equality follows because, conditionally on the respondent data, the completed values and therefore $\widehat{\mathbb{V}}_{\text{naive}}(\hat{\mu}_{LRI})$ are fixed.

Under the linear imputation model, this correction is given by

$$\mathbb{V}_{\text{diff}}\left(\hat{\mu}_{LRI}\right) = \frac{\sigma^2}{N^2}\sum_{k\in S} w_k^2(1-\pi_k)(1-r_k)c_k. \tag{2.8.3}$$

Notice that every term in (2.8.3) is nonnegative. This makes explicit the downward bias of the naive variance estimator under the model: the correction restores the residual variability that is lost when the missing values are replaced by deterministic regression predictions.

Replacing σ^2 by an estimator $\hat{\sigma}^2$ gives

$$\widehat{\mathbb{V}}_{\text{diff}}\left(\hat{\mu}_{LRI}\right) = \frac{\hat{\sigma}^2}{N^2}\sum_{k\in S} w_k^2(1-\pi_k)(1-r_k)c_k.$$

A possible estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{k\in S_r} w_k\left(y_k - \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS}\right)^2}{\sum_{k\in S_r} w_k c_k}.$$

The sampling component of the variance can therefore be estimated by

$$\widehat{\mathbb{V}}_{\text{sam}}\left(\hat{\mu}_{LRI}\right) = \widehat{\mathbb{V}}_{\text{naive}}\left(\hat{\mu}_{LRI}\right) + \widehat{\mathbb{V}}_{\text{diff}}\left(\hat{\mu}_{LRI}\right).$$

We now consider the additional variability due to nonresponse. To simplify the notation, define

$$\widehat{\mathbf{A}}_r = \sum_{l\in S_r} w_l c_l^{-1}\boldsymbol{v}_l\boldsymbol{v}_l^\top$$

and

$$\widehat{\boldsymbol{b}}_m = \sum_{l\in S_m} w_l\boldsymbol{v}_l.$$

For each unit $k \in S$, let

$$g_k = 1 + \widehat{\boldsymbol{b}}_m^\top \widehat{\mathbf{A}}_r^{-1} c_k^{-1}\boldsymbol{v}_k.$$

The nonresponse component is then

$$\mathbb{V}_{\text{nr}}\left(\hat{\mu}_{LRI}\right) = \frac{\sigma^2}{N^2}\mathbb{E}_p\mathbb{E}_q\left\{\sum_{k\in S} w_k^2(r_k g_k - 1)^2 c_k\right\},$$

and can be estimated by

$$\widehat{\mathbb{V}}_{\text{nr}}\left(\hat{\mu}_{LRI}\right) = \frac{\hat{\sigma}^2}{N^2}\sum_{k\in S} w_k^2(r_k g_k - 1)^2 c_k.$$

Finally, the mixed component is

$$\mathbb{V}_{\text{mix}}\left(\hat{\mu}_{LRI}\right) = \frac{\sigma^2}{N^2}\mathbb{E}_p\mathbb{E}_q\left\{\sum_{k\in S} w_k^2(1-\pi_k)(r_k g_k - 1) c_k\right\},$$

with estimator

$$\widehat{\mathbb{V}}_{\text{mix}}\left(\hat{\mu}_{LRI}\right) = \frac{\hat{\sigma}^2}{N^2}\sum_{k\in S} w_k^2(1-\pi_k)(r_k g_k - 1) c_k.$$

Therefore, using the variance decomposition introduced previously, the total variance estimator under the two-phase approach is

$$\begin{aligned}\widehat{\mathbb{V}}_{\text{tot,tp}}\left(\hat{\mu}_{LRI}\right) = \widehat{\mathbb{V}}_{\text{sam}}\left(\hat{\mu}_{LRI}\right) + \widehat{\mathbb{V}}_{\text{nr}}\left(\hat{\mu}_{LRI}\right) \\ + 2\widehat{\mathbb{V}}_{\text{mix}}\left(\hat{\mu}_{LRI}\right).\end{aligned}$$

2.8.1.2 Reverse approach

We now consider the reverse approach. As discussed previously, this approach is particularly convenient because it separates the total variance into two components,

$$\mathbb{V}_{\text{tot,ra}}\left(\hat{\mu}_{LRI}\right) = \mathbb{V}_1\left(\hat{\mu}_{LRI}\right) + \mathbb{V}_2\left(\hat{\mu}_{LRI}\right).$$

Because $\hat{\boldsymbol{\beta}}_{WLS}$ is estimated from the respondents, $\hat{\mu}_{LRI}$ is first linearized with respect to

$\hat{\boldsymbol{\beta}}_{WLS}$. This leads to the linearized variable

$$\hat{\delta}_k = \tilde{y}_k + \widehat{\boldsymbol{b}}_m^\top \widehat{\mathbf{A}}_r^{-1} r_k c_k^{-1} \boldsymbol{v}_k \left(y_k - \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS} \right).$$

The first variance component can then be estimated using the usual Horvitz–Thompson variance estimator applied to the linearized variable:

$$\widehat{\mathbb{V}}_1 \left(\hat{\mu}_{LRI} \right) = \frac{1}{N^2} \sum_{k \in S} \sum_{l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{\delta}_k}{\pi_k} \frac{\hat{\delta}_l}{\pi_l}.$$

For the second component, a first-order linearization gives

$$\mathbb{E}_p \left(\hat{\mu}_{LRI} - \mu \right) \approx \frac{1}{N} \sum_{k \in U} \xi_k \varepsilon_k,$$

where

$$\xi_k = \left\{ \sum_{l \in U} (1 - r_l) \boldsymbol{v}_l \right\}^\top \left\{ \sum_{l \in U} r_l c_l^{-1} \boldsymbol{v}_l \boldsymbol{v}_l^\top \right\}^{-1} r_k c_k^{-1} \boldsymbol{v}_k - (1 - r_k).$$

It follows that

$$\mathbb{V}_2 \left(\hat{\mu}_{LRI} \right) \approx \frac{\sigma^2}{N^2} \mathbb{E}_q \left\{ \sum_{k \in U} \xi_k^2 c_k \right\}.$$

The sample analogue of ξ_k is

$$\hat{\xi}_k = \widehat{\boldsymbol{b}}_m^\top \widehat{\mathbf{A}}_r^{-1} r_k c_k^{-1} \boldsymbol{v}_k - (1 - r_k).$$

Therefore,

$$\widehat{\mathbb{V}}_2 \left(\hat{\mu}_{LRI} \right) = \frac{\hat{\sigma}^2}{N^2} \sum_{k \in S} w_k \hat{\xi}_k^2 c_k.$$

The resulting variance estimator under the reverse approach is

$$\widehat{\mathbb{V}}_{\text{tot,ra}} \left(\hat{\mu}_{LRI} \right) = \widehat{\mathbb{V}}_1 \left(\hat{\mu}_{LRI} \right) + \widehat{\mathbb{V}}_2 \left(\hat{\mu}_{LRI} \right).$$

2.8.2 Root-n consistency of the linear regression imputed estimator

We now study the rate of convergence of the linear regression imputed estimator. Suppose that the linear imputation model is correctly specified, so that

$$y_k = \boldsymbol{v}_k^\top \boldsymbol{\beta} + \varepsilon_k,$$

with

$$\mathbb{E}_m(\varepsilon_k \mid \boldsymbol{v}_k) = 0.$$

Under this model, the corresponding Oracle estimator is

$$\hat{\mu}_{\text{oracle}} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k \boldsymbol{v}_k^\top \boldsymbol{\beta}\right\},$$

whereas the linear regression imputed estimator is

$$\hat{\mu}_{LRI} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS}\right\}.$$

Therefore,

$$\begin{aligned}\hat{\mu}_{LRI} - \hat{\mu}_{\text{oracle}} &= \frac{1}{N}\sum_{k\in S_m} w_k \boldsymbol{v}_k^\top \left(\hat{\boldsymbol{\beta}}_{WLS} - \boldsymbol{\beta}\right) \\ &= \left\{\frac{1}{N}\sum_{k\in S_m} w_k \boldsymbol{v}_k\right\}^\top \left(\hat{\boldsymbol{\beta}}_{WLS} - \boldsymbol{\beta}\right).\end{aligned}$$

It follows that

$$\hat{\mu}_{LRI} - \mu = \underbrace{(\hat{\mu}_{\text{oracle}} - \mu)}_{\text{I}} + \underbrace{\left\{\frac{1}{N}\sum_{k\in S_m} w_k \boldsymbol{v}_k\right\}^\top \left(\hat{\boldsymbol{\beta}}_{WLS} - \boldsymbol{\beta}\right)}_{\text{II}}. \tag{2.8.4}$$

From the previous section,

$$\hat{\mu}_{\text{oracle}} - \mu = O_p\left(n^{-1/2}\right).$$

Hence, term I is of order $O_p(n^{-1/2})$. For term II, suppose that standard regularity conditions ensure that

$$\hat{\boldsymbol{\beta}}_{WLS} - \boldsymbol{\beta} = O_p\left(n^{-1/2}\right)$$

and

$$\frac{1}{N}\sum_{k\in S_m} w_k \boldsymbol{v}_k = O_p(1).$$

Then,

$$\hat{\mu}_{LRI} - \hat{\mu}_{\text{oracle}} = O_p\left(n^{-1/2}\right).$$

Combining the two terms in (2.8.4), we obtain

$$\hat{\mu}_{LRI} - \mu = O_p\left(n^{-1/2}\right).$$

Thus, under a correctly specified linear imputation model and standard regularity conditions, the linear regression imputed estimator is root-n consistent for the population mean μ.

2.9 Simulation study

We conducted a Monte Carlo simulation study to evaluate the variance estimators derived for the linear regression imputed estimator $\hat{\mu}_{LRI}$ and the Oracle imputed estimator $\hat{\mu}_{\text{oracle}}$. In particular, we considered variance estimation under both the two-phase and the reverse approaches.

2.9.1 Simulation design

We generated a finite population U of size $N = 10,000$. For each population unit $k \in U$, let $\boldsymbol{v}_k = (v_{k1}, \ldots, v_{kp})^\top$ denote a vector of $p = 10$ auxiliary variables. The auxiliary variables were generated independently according to

$$v_{kj} \sim \chi^2_2, \qquad j = 1, \ldots, p, \qquad k = 1, \ldots, N.$$

The auxiliary variables were generated once and kept fixed throughout the Monte Carlo simulation.

The survey variable was generated according to the linear model

$$y_k = v_{k1} + 2.5v_{k4} + 1.3v_{k6} + \varepsilon_k. \tag{2.9.1}$$

We considered three levels of residual variability:

$$\varepsilon_k \sim \mathcal{N}(0, \sigma^2),$$

with $\sigma^2 \in \{100, 49, 9\}$. These three settings correspond to low, moderate, and high predictive strength, respectively. Under the data-generating mechanism considered here, the corresponding theoretical values of R^2 are approximately 0.26, 0.42, and 0.80.

Nonresponse was generated according to a logistic response model. More precisely,

$$p(\boldsymbol{v}_k) = P(r_k = 1 \mid \boldsymbol{v}_k) = \frac{\exp(0.5 + \boldsymbol{\alpha}^\top \boldsymbol{v}_k)}{1 + \exp(0.5 + \boldsymbol{\alpha}^\top \boldsymbol{v}_k)},$$

where

$$\boldsymbol{\alpha} = (-0.234, -0.152, -0.355, 0.1096, 0.339, 0.315, 0.019, 0.2612, -0.213, -0.430)^\top.$$

The vector $\boldsymbol{\alpha}$ was chosen so that the overall response rate was approximately 50%.

For each Monte Carlo replication $b = 1, \ldots, B$, with $B = 10,000$, the following steps were performed:

1. For each of the three values of σ^2, generate $y_1^{(b)}, \ldots, y_N^{(b)}$ according to (2.9.1) and compute the corresponding finite population mean

$$\mu^{(b)} = \frac{1}{N} \sum_{k \in U} y_k^{(b)}.$$

2. Select a sample $S^{(b)}$ of size $n = 500$ by simple random sampling without replacement. The sampling fraction is therefore $n/N = 0.05$.

3. For each sampled unit, independently generate the response indicator according to

$$r_k^{(b)} \sim \text{Bernoulli}\left\{p(\boldsymbol{v}_k)\right\}.$$

4. Compute the linear regression and Oracle imputed estimators. For linear regression imputation, the missing values are replaced by

$$y_k^* = \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS}, \qquad k \in S_m,$$

whereas the Oracle imputed values are

$$y_k^* = f(\boldsymbol{v}_k) = v_{k1} + 2.5v_{k4} + 1.3v_{k6}.$$

5. Compute the variance estimators derived previously. Under the two-phase approach, the sampling, nonresponse, and mixed components were computed, as well as the resulting total variance estimator. Under the reverse approach, the components $\widehat{\mathbb{V}}_1$ and $\widehat{\mathbb{V}}_2$ were computed together with the corresponding total variance estimator.

2.9.2 Monte Carlo measures

Because a new realization of the survey variable is generated at each Monte Carlo replication, the true finite population mean $\mu^{(b)}$ also changes from one replication to another. Consequently, the Monte Carlo variance must be computed from the estimation errors rather than from the values of $\hat{\mu}^{(b)}$ alone.

For a given estimator $\hat{\mu}$, define

$$d^{(b)} = \hat{\mu}^{(b)} - \mu^{(b)}$$

and

$$\bar{d} = \frac{1}{B}\sum_{b=1}^{B} d^{(b)}.$$

The Monte Carlo variance of the estimation error is estimated by

$$\mathbb{V}_{MC}(\hat{\mu}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(d^{(b)} - \bar{d}\right)^2.$$

For a variance estimator $\widehat{\mathbb{V}}(\hat{\mu})$, its Monte Carlo average is

$$\mathbb{E}_{MC}\left\{\widehat{\mathbb{V}}(\hat{\mu})\right\} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mathbb{V}}^{(b)}(\hat{\mu}).$$

The Monte Carlo percent relative bias is then defined as

$$RB_{MC}\left\{\widehat{\mathbb{V}}(\hat{\mu})\right\} = 100\frac{\mathbb{E}_{MC}\left\{\widehat{\mathbb{V}}(\hat{\mu})\right\} - \mathbb{V}_{MC}(\hat{\mu})}{\mathbb{V}_{MC}(\hat{\mu})}.$$

We also evaluated the coverage rate of nominal 95% confidence intervals. At replication b, the confidence interval is given by

$$\hat{\mu}^{(b)} \pm 1.96\sqrt{\widehat{\mathbb{V}}^{(b)}(\hat{\mu})}.$$

The Monte Carlo coverage rate is therefore

$$CR_{MC} = \frac{100}{B}\sum_{b=1}^{B} 1\left\{\mu^{(b)} \in \left[\hat{\mu}^{(b)} - 1.96\sqrt{\widehat{\mathbb{V}}^{(b)}(\hat{\mu})},\ \hat{\mu}^{(b)} + 1.96\sqrt{\widehat{\mathbb{V}}^{(b)}(\hat{\mu})}\right]\right\}.$$

The Monte Carlo relative bias and the corresponding coverage rate are reported in Table 2.1 for each variance estimator.

2.9.3 Results

Table 2.1 shows that the variance estimators for the linear regression imputed estimator perform well overall under both the two-phase and the reverse approaches. The performance is particularly good in the moderate- and high-R^2 settings, where the Monte Carlo relative biases are small and the coverage rates are close to the nominal level of 95%.

Some underestimation of the variance is observed in the low-R^2 setting. In this setting, a larger fraction of the variability of y is due to the model errors, so that the uncertainty associated with the residual component plays a more important role. The observed

underestimation should therefore be viewed as a finite-sample feature of the variance approximation and estimation rather than as the contribution of an additional variance component.

As expected, the naive variance estimator systematically underestimates the true variance. Its coverage rates are approximately 63%, 74%, and 87% in the low-, moderate-, and high-R^2 settings, respectively. The improvement as R^2 increases is consistent with the fact that deterministic regression predictions reproduce a larger proportion of the variability of y when the regression model is highly predictive. Nevertheless, even in the high-R^2 setting, the naive variance estimator fails to achieve the nominal coverage rate. This confirms the importance of accounting for the additional uncertainty introduced by imputation.

The variance estimator for the Oracle estimator performs very well in all three settings, with Monte Carlo relative bias close to zero and coverage rates close to 95%. This is expected since the true conditional mean function $f(\boldsymbol{v}_k)$ is known in the Oracle setting and no additional variability is introduced by estimating the regression coefficients.

Table 2.1: Monte Carlo percent relative bias and coverage rates of the variance estimators.

	$R^2 \approx 0.26$		$R^2 \approx 0.42$		$R^2 \approx 0.80$	
Variance estimator	RB (%)	CR (%)	RB (%)	CR (%)	RB (%)	CR (%)
Linear regression imputation						
$\widehat{\mathbb{V}}_{\text{tot,tp}}(\hat{\mu}_{LRI})$	−2.9	94.9	0.2	95.3	−0.4	95.0
$\widehat{\mathbb{V}}_{\text{tot,ra}}(\hat{\mu}_{LRI})$	−7.3	94.0	−3.6	94.7	−2.7	94.7
$\widehat{\mathbb{V}}_{\text{naive}}(\hat{\mu}_{LRI})$	−78.6	62.9	−66.8	74.0	−40.1	86.7
Oracle imputation						
$\widehat{\mathbb{V}}_{\text{tot,tp}}(\hat{\mu}_{\text{oracle}})$	−1.8	94.7	0.4	95.1	−0.2	94.7
$\widehat{\mathbb{V}}_{\text{tot,ra}}(\hat{\mu}_{\text{oracle}})$	−4.7	94.3	−1.3	94.9	−0.8	94.7

We also examined the relative contribution of the different components to the total variance. Under the two-phase approach, the contributions of the sampling and nonresponse

components are defined as

$$C_{\text{sam}} = 100\frac{\widehat{\mathbb{V}}_{\text{sam}}}{\widehat{\mathbb{V}}_{\text{tot,tp}}}$$

and

$$C_{\text{nr}} = 100\frac{\widehat{\mathbb{V}}_{\text{nr}}}{\widehat{\mathbb{V}}_{\text{tot,tp}}}.$$

Since the mixed component enters the total variance with a factor of two, its contribution is defined as

$$C_{\text{mix}} = 100\frac{2\widehat{\mathbb{V}}_{\text{mix}}}{\widehat{\mathbb{V}}_{\text{tot,tp}}}.$$

With this definition,

$$C_{\text{sam}} + C_{\text{nr}} + C_{\text{mix}} = 100.$$

Notice that C_{mix} may be negative because the mixed component is a covariance term rather than a variance.

Under the reverse approach, the corresponding contributions are

$$C_1 = 100\frac{\widehat{\mathbb{V}}_1}{\widehat{\mathbb{V}}_{\text{tot,ra}}}$$

and

$$C_2 = 100\frac{\widehat{\mathbb{V}}_2}{\widehat{\mathbb{V}}_{\text{tot,ra}}},$$

so that

$$C_1 + C_2 = 100.$$

The contributions were computed at each Monte Carlo replication and then averaged over the B replications. The results are reported in Table 2.2.

Under the two-phase approach, Table 2.2 shows that the relative contribution of the nonresponse component decreases as the predictive strength of the imputation model increases. Conversely, the relative contribution of the sampling component increases. This is consistent with the fact that, as the conditional mean explains a larger proportion

Table 2.2: Relative contributions of the variance components to the total variance estimator.

Variance component	$R^2 \approx 0.26$	$R^2 \approx 0.42$	$R^2 \approx 0.80$
Two-phase approach			
$\widehat{\mathbb{V}}_{\text{sam}}(\hat{\mu}_{LRI})$	35.4%	44.7%	66.9%
$\widehat{\mathbb{V}}_{\text{nr}}(\hat{\mu}_{LRI})$	64.6%	55.3%	33.1%
Reverse approach			
$\widehat{\mathbb{V}}_1(\hat{\mu}_{LRI})$	96.6%	97.1%	98.3%
$\widehat{\mathbb{V}}_2(\hat{\mu}_{LRI})$	3.4%	2.9%	1.7%

of the variability of y, the residual variability associated with the missing values becomes smaller. The mixed component is essentially zero in the settings considered here, so that the total variance is explained almost entirely by the sampling and nonresponse components..

Under the reverse approach, $\widehat{\mathbb{V}}_1$ accounts for most of the total variance, whereas the contribution of $\widehat{\mathbb{V}}_2$ is small. This result is consistent with the discussion in the previous section. Indeed, the sampling fraction in the simulation is only $n/N = 0.05$, and the contribution of $\mathbb{V}_2$ is expected to become negligible relative to $\mathbb{V}_1$ as the sampling fraction tends to zero.

3

Inference of doubly robust estimators based on machine learning methods

3.1 The imputed estimator under machine learning methods

3.1.1 Overview of machine learning methods

Machine learning refers to a broad class of methods that can be used to estimate complex relationships between variables. Unlike simple parametric models, such as linear or logistic regression, machine learning methods do not always require the exact form of the relationship to be specified in advance. This makes them particularly attractive when the relationship between the variable of interest and the auxiliary variables is complicated or not well understood.

In the previous chapter, we considered imputation methods based mainly on parametric models. For example, under linear regression imputation, the conditional mean of the survey variable is assumed to follow a linear model. When this model is correctly specified, the properties of the resulting imputed estimator can be studied using standard statistical tools. In particular, under suitable regularity conditions, the estimator is root-n consistent. We also saw how its variance can be studied using the two-phase framework or the reverse approach.

The main limitation of parametric imputation is that the form of the model must be chosen in advance. In practice, the relationship between the survey variable y_k and the auxiliary variables $\boldsymbol{v}_k$ may be more complicated than a simple linear or logistic model. It may contain nonlinear effects, interactions between variables, or other structures that are difficult to identify beforehand. If the assumed model does not describe the true relationship adequately, the imputation model is misspecified and the resulting estimator may be biased.

Machine learning provides a possible way to reduce this problem. Instead of assuming a specific parametric form, a machine learning method can be used to estimate the conditional mean

$$f(\boldsymbol{v}_k) = \mathbb{E}_m(y_k \mid \boldsymbol{v}_k)$$

directly from the observed data. Methods such as regression trees, random forest, gradient boosting, and neural networks are able to represent nonlinear relationships and interactions without requiring all of them to be specified explicitly. These methods can therefore be useful for constructing imputed values when the relationship between y_k and $\boldsymbol{v}_k$ is complex.

However, the use of machine learning creates new difficulties. In a parametric model, the unknown regression function is described by a finite-dimensional parameter, such as the vector $\boldsymbol{\beta}$ in linear regression. The behaviour of the estimator can then often be studied using standard tools such as Taylor linearization. With machine learning, the fitted function may be much more complicated and there may be no simple finite-dimensional parameter that summarizes the model. As a consequence, studying the bias and variance of an imputed estimator based on machine learning becomes more difficult.

Another important point is that good prediction does not necessarily imply good

estimation of a finite population parameter. A machine learning method may predict the missing values accurately at the individual level, but the resulting estimator of a population mean or total may still have an important bias or variance. In survey sampling, the objective is therefore not only to obtain good predictions for the nonrespondents. The main objective is to construct an estimator of the population parameter with good statistical properties.

The use of machine learning also raises questions about consistency. In the parametric setting considered in Chapter 2, the estimated regression parameter typically converges at the usual root-n rate under standard conditions. Machine learning estimators may converge more slowly, especially when many auxiliary variables are used or when the relationship between the variables is highly complex. Therefore, an imputed estimator obtained by simply replacing the missing values with machine learning predictions is not automatically root-n consistent. Additional conditions on the quality of the estimated regression function are required.

Variance estimation is another important challenge. In Chapter 2, variance estimators could be derived explicitly for parametric imputation methods because the structure of the fitted model was known. The two-phase and reverse approaches provided useful decompositions of the total variance. With machine learning, the additional variability caused by estimating the imputation function is generally more difficult to express analytically. This makes the construction of valid variance estimators and confidence intervals more challenging.

Another difficulty concerns the asymptotic distribution of the imputed estimator. Even if an estimator is consistent and a valid estimator of its variance is available, this is not sufficient to justify the construction of confidence intervals based on a normal approximation. For this purpose, a Central Limit Theorem is also required. In the parametric setting, asymptotic normality can often be established using standard arguments because the imputation model depends on a finite-dimensional parameter. With machine learning, however, the estimated imputation function may have a much more complex behaviour and may converge at a slower rate. Establishing a Central Limit Theorem for an estimator involving machine learning predictions is therefore generally more difficult. Without such

a result, confidence intervals of the usual form

$$\hat{\mu} \pm 1.96\sqrt{\widehat{\mathbb{V}}(\hat{\mu})}$$

are not automatically justified, even if the point estimator is consistent and its variance can be estimated.

Finally, most machine learning methods require the choice of tuning parameters, also called hyperparameters. These parameters control the complexity of the fitted model. For example, they may determine the depth of a regression tree, the number of trees in a random forest, or the learning rate used in gradient boosting. If the model is not sufficiently flexible, it may fail to capture important features of the data. On the other hand, if it is too flexible, it may overfit the observed data. The choice of hyperparameters can therefore affect the bias and variance of the final survey estimator.

The purpose of this chapter is to study how machine learning methods can be used for imputation while preserving valid inference for finite population parameters. In particular, we are interested in methods that retain the flexibility of machine learning while controlling the bias and variance of the resulting estimators.

Many machine learning methods can be used to estimate the relationship between the survey variable y_k and the auxiliary variables $\boldsymbol{v}_k$. In this thesis, we consider several commonly used methods. These methods differ in the way they construct predictions and in the tuning parameters that control their complexity. We briefly describe them below.

Example 4. *Regression and classification trees divide the predictor space into regions using a sequence of binary splits. At each step, the algorithm chooses an auxiliary variable and a cut-off value in order to divide the observations into two groups. The procedure is repeated until a stopping rule is reached. For a continuous response variable, a regression tree predicts the value of a new observation using the average response among the observations belonging to the same terminal node, or leaf. For a categorical response variable, a classification tree typically predicts the most frequent category within the corresponding leaf. The complexity of a tree is controlled by tuning parameters such as the maximum tree depth, the minimum number of observations in a terminal node, and pruning parameters. A very deep tree can adapt closely to the observed data but may overfit, whereas a tree that is too shallow may fail to capture important features of the relationship.*

Example 5. *Random forests (RF), introduced by Breiman (2001), combine a large number of regression or classification trees. Instead of relying on a single tree, many trees are constructed using randomization. In particular, each tree is usually fitted using a bootstrap sample of the observations, and only a random subset of the auxiliary variables is considered when selecting a split. For a continuous response, the final prediction is obtained by averaging the predictions from all the trees. For a categorical response, the predictions of the individual trees are combined to determine the predicted class. Averaging many trees generally produces more stable predictions than using a single tree. Important tuning parameters include the number of trees, the number of auxiliary variables considered at each split, often denoted by* `mtry`*, and parameters controlling the size of the individual trees.*

Example 6. *Gradient boosting also combines several trees, but the trees are constructed sequentially. The first tree provides an initial approximation to the relationship between the response and the auxiliary variables. Additional trees are then added to improve the current prediction. In this way, the fitted model is built progressively by combining a sequence of relatively simple trees. The main tuning parameters include the number of trees, the learning rate, and the depth of the individual trees. The learning rate controls the contribution of each new tree to the final prediction. A smaller learning rate generally requires a larger number of trees, whereas the tree depth controls the complexity of the relationships and interactions that can be captured.*

Example 7. *Cubist combines tree-based methods with linear regression. The predictor space is divided into regions using a set of rules, similarly to a regression tree. However, instead of assigning a constant prediction to each region, a linear regression model is used to obtain the prediction. This allows Cubist to capture nonlinear relationships by using different linear models in different regions of the predictor space. The method can also combine several successive models, called committees, and can use a nearest-neighbour adjustment to improve the final predictions. Two important tuning parameters are the number of committees and the number of nearest neighbours used in the adjustment.*

Example 8. *The K-nearest-neighbours (KNN) method was introduced in Chapter 2 in the context of nearest-neighbour imputation. For a unit with auxiliary information $\boldsymbol{v}_k$, the method identifies the K respondents whose auxiliary variables are closest to those of unit k. For a continuous response variable, the prediction is typically obtained by averaging the observed values of the K nearest neighbours. For a categorical response variable, the predicted category can be determined from the categories observed among these neighbours. The main tuning parameter is K. A small value of K produces a very local and flexible*

fit but may lead to highly variable predictions. A large value of K produces smoother predictions but may combine observations that are less similar. The choice of the distance function and the scale of the auxiliary variables are also important.

All the methods described above involve tuning parameters that control the complexity of the fitted model. If the model is too simple, it may fail to capture important relationships in the data. If it is too complex, it may overfit the observed sample and produce unstable predictions. A common approach is to select the tuning parameters using cross-validation, where candidate models are compared according to their predictive performance on observations that were not used for fitting.

3.1.2 Rate of convergence of the imputed estimator

We now study more closely the effect of estimating the conditional mean function on the imputed estimator. Throughout this section, we consider the model

$$y_k = f(\boldsymbol{v}_k) + \varepsilon_k,$$

where

$$f(\boldsymbol{v}_k) = \mathbb{E}_m(y_k \mid \boldsymbol{v}_k)$$

and

$$\mathbb{E}_m(\varepsilon_k \mid \boldsymbol{v}_k) = 0.$$

The function f is unknown and must be estimated from the respondents. We denote its estimator by $\hat{f}$.

The function f will be referred to as a nuisance function. The term "nuisance" does not mean that f is unimportant. Rather, it indicates that f is not the parameter of primary interest. Our objective is to estimate the finite population mean μ, while f is an intermediate quantity that must be estimated in order to construct an estimator of μ. This terminology will become particularly useful later when other nuisance functions, such as the response probability, are introduced.

To keep the notation consistent with Chapter 2, we write the imputed estimator based on

$\hat{f}$ as

$$\hat{\mu}_I(\hat{f}) = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k \hat{f}(\boldsymbol{v}_k)\right\}.$$

For example, under linear regression imputation,

$$\hat{f}(\boldsymbol{v}_k) = \boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{WLS},$$

whereas in this chapter $\hat{f}$ may be obtained using a machine learning method.

If the true conditional mean function f were known, the corresponding Oracle estimator would be

$$\hat{\mu}_I(f) = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k f(\boldsymbol{v}_k)\right\}.$$

This is the Oracle estimator introduced in Chapter 2. The difference between the feasible estimator $\hat{\mu}_I(\hat{f})$ and the Oracle estimator is therefore

$$\hat{\mu}_I(\hat{f}) - \hat{\mu}_I(f) = \frac{1}{N}\sum_{k\in S_m} w_k\left\{\hat{f}(\boldsymbol{v}_k) - f(\boldsymbol{v}_k)\right\} \equiv R_{1,n}(\hat{f}, f).$$

Consequently,

$$\hat{\mu}_I(\hat{f}) - \mu = \{\hat{\mu}_I(f) - \mu\} + R_{1,n}(\hat{f}, f). \tag{3.1.1}$$

The decomposition in (3.1.1) separates two sources of error. The first term is the error that would remain even if the true function f were known. As shown in Chapter 2, under suitable regularity conditions,

$$\hat{\mu}_I(f) - \mu = O_p(n^{-1/2}).$$

The second term, $R_{1,n}(\hat{f}, f)$, is the additional error caused by replacing f with the estimated function $\hat{f}$. Thus, the rate of convergence of the feasible estimator depends on how large this additional term is. In a parametric model, as in Section 2.8.2, $\hat{f}$ converges to f at the rate $n^{-1/2}$, so that the resulting imputed estimator is root-n consistent.

The situation is more complicated for nonparametric methods. The rate at which $\hat{f}$ converges to f depends on the complexity of the function and on the dimension of the auxiliary information. A classical result of Stone (1982) provides a useful illustration. Let p denote the dimension of the vector of auxiliary variables used to estimate f, and suppose that f belongs to a smoothness class of order s. Under suitable regularity conditions, the optimal nonparametric rate in L_2 norm is of the order

$$\|\hat{f} - f\|_2 = O_p\left(n^{-s/(2s+p)}\right),$$

or, equivalently, the corresponding squared error is of order $O_p\left(n^{-2s/(2s+p)}\right)$.

This result highlights two important features. First, estimation becomes easier when f is smoother, since the rate improves as s increases. Second, for a fixed level of smoothness, the rate deteriorates as the dimension p increases. This phenomenon is commonly referred to as the curse of dimensionality. Intuitively, when the number of auxiliary variables increases, the available observations become increasingly sparse in the predictor space, making it more difficult to estimate f accurately.

Nearest-neighbour imputation provides a simple illustration of this dimensionality problem. Let $l(k)$ denote the respondent selected as the nearest neighbour of a nonrespondent k. We can regard nearest-neighbour imputation as estimating $f(\boldsymbol{v}_k)$ by

$$\hat{f}_{NN}(\boldsymbol{v}_k) = y_{l(k)}.$$

The corresponding imputed estimator is therefore

$$\hat{\mu}_I(\hat{f}_{NN}) = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k y_{l(k)}\right\}.$$

To understand its behaviour, we compare it with the Oracle estimator $\hat{\mu}_I(f)$. We have

$$\begin{aligned}\hat{\mu}_I(\hat{f}_{NN}) - \hat{\mu}_I(f) &= \frac{1}{N}\sum_{k\in S_m} w_k\left\{y_{l(k)} - f(\boldsymbol{v}_k)\right\} \\ &= \frac{1}{N}\sum_{k\in S_m} w_k\left\{f(\boldsymbol{v}_{l(k)}) - f(\boldsymbol{v}_k)\right\} + \frac{1}{N}\sum_{k\in S_m} w_k\varepsilon_{l(k)}. \qquad (3.1.2)\end{aligned}$$

The decomposition in (3.1.2) is useful because the two terms have different interpretations. The second term is generated by the model residuals of the selected donors. Since these residuals have mean zero under the model, this term represents additional random variability. The first term is different. Even after averaging over the model errors, it remains because the auxiliary variables of the donor, $\boldsymbol{v}_{l(k)}$, are generally not exactly equal to those of the nonrespondent, $\boldsymbol{v}_k$. This systematic discrepancy is what gives rise to the matching bias.

More precisely, conditionally on the selected sample, the response pattern, and the auxiliary variables, we define the matching bias as

$$B_{NN} = \mathbb{E}_m \left[\hat{\mu}_I(\hat{f}_{NN}) - \hat{\mu}_I(f) \mid \boldsymbol{I}, \boldsymbol{r}, \mathbf{V} \right].$$

Thus, B_{NN} is the conditional bias of the nearest-neighbour imputed estimator relative to the Oracle estimator. It measures the systematic error introduced by using imperfect matches.

Because the donor $l(k)$ is determined from the auxiliary information and

$$\mathbb{E}_m \left(\varepsilon_{l(k)} \mid \boldsymbol{I}, \boldsymbol{r}, \mathbf{V} \right) = 0,$$

taking the conditional model expectation in (3.1.2) gives

$$B_{NN} = \frac{1}{N} \sum_{k \in S_m} w_k \left\{ f(\boldsymbol{v}_{l(k)}) - f(\boldsymbol{v}_k) \right\}. \tag{3.1.3}$$

Equation (3.1.3) shows precisely where the matching bias comes from. If the donor and the nonrespondent had exactly the same auxiliary information, then $\boldsymbol{v}_{l(k)} = \boldsymbol{v}_k$ and their contribution to the matching bias would be zero. In practice, nearest-neighbour matching only makes these two vectors close, not identical.

We can now study the order of B_{NN}. Suppose that nearest-neighbour matching is performed using a p-dimensional vector of continuous auxiliary variables. Also suppose that the response rate is bounded away from zero, so that the number of available respondents

is of the same order as the original sample size n. Under standard regularity conditions, the distance between a nonrespondent and its nearest respondent is typically of order

$$\left\|\boldsymbol{v}_{l(k)} - \boldsymbol{v}_k\right\| = O_p(n^{-1/p}).$$

This result has a simple interpretation. With one matching variable, it is relatively easy to find a respondent very close to a given nonrespondent. As the dimension increases, observations become more dispersed in the predictor space and the closest available donor is generally farther away.

Suppose in addition that f is Lipschitz continuous, so that for some constant C,

$$\left|f(\boldsymbol{v}_{l(k)}) - f(\boldsymbol{v}_k)\right| \leq C \left\|\boldsymbol{v}_{l(k)} - \boldsymbol{v}_k\right\|.$$

It follows that

$$f(\boldsymbol{v}_{l(k)}) - f(\boldsymbol{v}_k) = O_p(n^{-1/p}).$$

If the sampling weights are of order N/n and the number of nonrespondents is of order n, equation (3.1.3) then gives

$$B_{NN} = O_p(n^{-1/p}). \tag{3.1.4}$$

We can now explain why this rate matters for inference. From Chapter 2, the Oracle estimator satisfies

$$\hat{\mu}_I(f) - \mu = O_p(n^{-1/2}).$$

Thus, the natural scale of the Oracle estimation error is $n^{-1/2}$. If we want nearest-neighbour imputation to have the same first-order behaviour as the Oracle estimator, its matching bias must be smaller than this Oracle error. In other words, we need

$$B_{NN} = o_p(n^{-1/2}).$$

Using (3.1.4), this requirement can be examined by multiplying B_{NN} by $\sqrt{n}$:

$$\sqrt{n}\, B_{NN} = O_p\left(n^{1/2-1/p}\right).$$

The quantity $n^{1/2-1/p}$ is therefore not the matching bias itself. It is the order of the

matching bias after it has been rescaled by $\sqrt{n}$, which is the scale used to study the limiting behaviour of a root-n estimator.

The consequences depend directly on the dimension p :

i) If $p = 1$, then

$$B_{NN} = O_p(n^{-1}),$$

so that

$$\sqrt{n}\, B_{NN} = O_p(n^{-1/2}) = o_p(1).$$

The matching bias is therefore negligible relative to the Oracle error. In this case, nearest-neighbour imputation can retain the same first-order root-n behaviour as the Oracle estimator, provided that the donor-residual term and the remaining random terms are also well controlled.

ii) If $p = 2$, then

$$B_{NN} = O_p(n^{-1/2}),$$

and hence

$$\sqrt{n}\, B_{NN} = O_p(1).$$

The matching bias may then be of the same order as the Oracle estimation error. The estimator may still converge at the root-n rate, but the matching bias may no longer be negligible and may affect its first-order limiting distribution.

iii) If $p > 2$, then $n^{-1/p}$ converges to zero more slowly than $n^{-1/2}$. Consequently,

$$\sqrt{n}\, B_{NN} = O_p\left(n^{1/2-1/p}\right)$$

may diverge. The matching bias is then not negligible in general and may dominate the Oracle error, so that root-n inference may require additional structure or bias correction. This dimensionality problem is closely related to the results of Abadie and Imbens (2006) for matching estimators with a fixed number of matches.

The same issue appears in the survey-sampling literature on nearest-neighbour imputation. Much of the classical asymptotic theory has therefore been developed using

a scalar matching variable. For example, Chen and Shao (2000) study nearest-neighbour imputation using a single matching variable. More recently, Yang and Kim (2019) establish asymptotic properties for nearest-neighbour imputation of several finite population parameters under conditions that reduce the relevant matching problem to a low-dimensional setting.

For more complex machine learning methods, the situation is less transparent. There is generally no single convergence rate that applies to every implementation of a given algorithm. Theoretical rates have been obtained for particular versions of some methods under specific assumptions. For example, Bühlmann and Yu (2003) derive convergence results for a particular L_2- boosting procedure, while Biau (2012) and Scornet, Biau, and Vert (2015) establish theoretical results for particular random-forest settings. These results do not imply a universal convergence rate for an arbitrary implementation of gradient boosting or random forest. For methods such as Cubist, a general rate theory directly applicable to the imputation problem considered here is even less developed.

For our purposes, however, the rate of convergence of $\hat{f}$ is not the final quantity of interest. The parameter we wish to estimate is μ, not f. What matters is therefore the effect of the nuisance estimation error on $\hat{\mu}_I(\hat{f})$, represented by $R_{1,n}(\hat{f}, f)$. To obtain the same first-order behaviour as the Oracle estimator, we would like

$$R_{1,n}(\hat{f}, f) = o_p(n^{-1/2}). \tag{3.1.5}$$

If (3.1.5) holds, then

$$\hat{\mu}_I(\hat{f}) = \hat{\mu}_I(f) + o_p(n^{-1/2}),$$

so that estimating the nuisance function has no first-order effect on the estimator of μ. For the simple plug-in imputed estimator, however, the estimation error $\hat{f} - f$ enters directly into $R_{1,n}(\hat{f}, f)$. When the nuisance function is estimated slowly, this first-order dependence may prevent the feasible estimator from behaving like the Oracle estimator. The objective of the following sections will therefore be to construct estimators whose first-order behaviour is less sensitive to estimation errors in the nuisance function. This idea will lead naturally to orthogonal estimators and, subsequently, to double machine learning.

3.2 The double/debiased machine learning framework

The double/debiased machine learning (DML) framework was introduced by Chernozhukov et al. (2018) to allow valid statistical inference when high-dimensional or machine learning methods are used to estimate nuisance functions. The framework relies on two main ingredients: Neyman orthogonality and cross-fitting. Neyman orthogonality reduces the first-order effect of errors in the estimated nuisance functions, while cross-fitting reduces the dependence created when the same observations are used both to estimate the nuisance functions and to construct the estimator of the parameter of interest. The developments in this section follow closely the survey-sampling adaptation discussed by Dagdoug and Haziza (2026b), with a particular focus on item nonresponse and estimation of a finite population mean.

We consider two nuisance functions. The first is the outcome regression function

$$f(\boldsymbol{v}_k) = \mathbb{E}_m(y_k \mid \boldsymbol{v}_k),$$

which describes the relationship between the survey variable and the auxiliary variables. The second is the response propensity

$$p(\boldsymbol{v}_k) = P(r_k = 1 \mid \boldsymbol{v}_k).$$

Under the MAR assumption introduced previously,

$$P(r_k = 1 \mid y_k, \boldsymbol{v}_k) = P(r_k = 1 \mid \boldsymbol{v}_k) = p(\boldsymbol{v}_k).$$

We also assume the usual positivity condition, so that $p(\boldsymbol{v}_k)$ is bounded away from zero.

Neither $f(\cdot)$ nor $p(\cdot)$ is the parameter of primary interest. The parameter of interest remains the finite population mean μ. The functions $f(\cdot)$ and $p(\cdot)$ are therefore nuisance functions: they are estimated because they are useful ingredients in the construction of an estimator of μ.

A natural estimator combining these two nuisance functions is the Augmented Inverse

Probability Weighting (AIPW) estimator. AIPW estimators have an important double robustness property: they remain unbiased, or consistent under the appropriate asymptotic framework, if either the outcome regression model or the response propensity model is correctly specified. This property has been studied extensively in the missing-data and survey-sampling literature; see, for example, Robins, Rotnitzky, and Zhao (1994), Kang and Schafer (2007), Cao, Tsiatis, and Davidian (2009), and Kim and Haziza (2014).

Double robustness and Neyman orthogonality are related but distinct properties. Double robustness protects the estimator against misspecification of one of the two nuisance functions. Neyman orthogonality, on the other hand, controls the effect of estimation errors in these nuisance functions. The latter property is particularly important when $f(\cdot)$ and $p(\cdot)$ are estimated using flexible machine learning methods that may converge at rates slower than $n^{-1/2}$.

3.2.1 The Oracle AIPW estimator

We first consider the Oracle AIPW estimator, for which both nuisance functions $f(\cdot)$ and $p(\cdot)$ are assumed to be known. Although this estimator is not feasible in practice, it provides a useful benchmark for constructing and studying feasible estimators.

For each sampled unit k, define the pseudo-value

$$\eta_k = f(\boldsymbol{v}_k) + \frac{r_k}{p(\boldsymbol{v}_k)} \left\{ y_k - f(\boldsymbol{v}_k) \right\}.$$

The Oracle AIPW estimator of μ is

$$\begin{aligned} \hat{\mu}_{aipw}(f, p) &= \frac{1}{N} \left\{ \sum_{k \in S} w_k f(\boldsymbol{v}_k) + \sum_{k \in S_r} w_k \frac{y_k - f(\boldsymbol{v}_k)}{p(\boldsymbol{v}_k)} \right\} \\ &= \frac{1}{N} \sum_{k \in S} w_k \eta_k. \end{aligned}$$

Thus, the Oracle AIPW estimator can be viewed as a Horvitz–Thompson estimator applied to the pseudo-values η_k.

Proposition 3.1. *Under the MAR assumption and the positivity condition, the Oracle AIPW estimator $\hat{\mu}_{aipw}(f,p)$ is mpq-unbiased for the finite population mean μ.*

Proof. Conditionally on y_k and $\boldsymbol{v}_k$,

$$\begin{aligned}\mathbb{E}_q\left(\eta_k \mid y_k, \boldsymbol{v}_k\right) &= f(\boldsymbol{v}_k) + \frac{\mathbb{E}_q(r_k \mid y_k, \boldsymbol{v}_k)}{p(\boldsymbol{v}_k)}\left\{y_k - f(\boldsymbol{v}_k)\right\} \\ &= f(\boldsymbol{v}_k) + \left\{y_k - f(\boldsymbol{v}_k)\right\} \\ &= y_k.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{mpq}\left\{\hat{\mu}_{aipw}(f,p)\right\} &= \frac{1}{N}\mathbb{E}_m\mathbb{E}_p\left(\sum_{k\in S} w_k y_k\right) \\ &= \frac{1}{N}\mathbb{E}_m\left(\sum_{k\in U} y_k\right) \\ &= \mathbb{E}_m(\mu).\end{aligned}$$

Hence,

$$\mathbb{E}_{mpq}\left\{\hat{\mu}_{aipw}(f,p) - \mu\right\} = 0.$$

□

Under suitable regularity conditions, the Oracle AIPW estimator is also root-n consistent and asymptotically normal; see Dagdoug and Haziza (2026b). In particular,

$$\hat{\mu}_{aipw}(f,p) - \mu = O_p(n^{-1/2}).$$

These properties make the Oracle AIPW estimator the benchmark that we would like a feasible machine-learning-based estimator to reproduce.

3.2.2 Double robustness

The double robustness property can be seen directly by replacing the true functions $f(\cdot)$ and $p(\cdot)$ by two fixed working functions $\tilde{f}(\cdot)$ and $\tilde{p}(\cdot)$. These working functions may be misspecified. The corresponding AIPW estimator is

$$\hat{\mu}_{aipw}(\tilde{f}, \tilde{p}) = \frac{1}{N} \sum_{k \in S} w_k \left[\tilde{f}(\boldsymbol{v}_k) + \frac{r_k}{\tilde{p}(\boldsymbol{v}_k)} \left\{ y_k - \tilde{f}(\boldsymbol{v}_k) \right\} \right].$$

Taking expectations with respect to the sampling design, the nonresponse mechanism, and the outcome model gives

$$\text{Bias}_{mpq} \left\{ \hat{\mu}_{aipw}(\tilde{f}, \tilde{p}) \right\} = \frac{1}{N} \sum_{k \in U} \left\{ 1 - \frac{p(\boldsymbol{v}_k)}{\tilde{p}(\boldsymbol{v}_k)} \right\} \left\{ \tilde{f}(\boldsymbol{v}_k) - f(\boldsymbol{v}_k) \right\}.$$

From the previous expression, we can see that the bias is equal to zero if either of the two working models is correctly specified. More precisely, the bias vanishes if $\tilde{p}(\boldsymbol{v}_k) = p(\boldsymbol{v}_k)$ for all k, even if $\tilde{f}(\cdot)$ is misspecified. Similarly, the bias is equal to zero if $\tilde{f}(\boldsymbol{v}_k) = f(\boldsymbol{v}_k)$ for all k, even if $\tilde{p}(\cdot)$ is misspecified. This property is referred to as double robustness: only one of the two nuisance models needs to be correctly specified for the estimator to remain unbiased.

3.2.3 Variance of the Oracle AIPW estimator

We now derive the variance of the Oracle AIPW estimator using the reverse approach introduced in Chapter 2. Let $\sigma_k^2 = \mathbb{V}_m(y_k \mid \boldsymbol{v}_k)$ denote the conditional model variance. In the homoscedastic case, $\sigma_k^2 = \sigma^2$ for all k.

Proposition 3.2. *Under the assumptions stated above, the total variance of the Oracle AIPW estimator is*

$$\mathbb{V}_{\text{tot}} \left\{ \hat{\mu}_{aipw}(f, p) \right\} = \frac{1}{N^2} \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{f(\boldsymbol{v}_k)}{\pi_k} \frac{f(\boldsymbol{v}_l)}{\pi_l} + \frac{1}{N^2} \sum_{k \in U} \frac{w_k - p(\boldsymbol{v}_k)}{p(\boldsymbol{v}_k)} \sigma_k^2. \tag{3.2.1}$$

Proof. Under the reverse approach,

$$\mathbb{V}_{\text{tot,ra}}\left\{\hat{\mu}_{aipw}(f,p)\right\} = \mathbb{V}_1\left\{\hat{\mu}_{aipw}(f,p)\right\} + \mathbb{V}_2\left\{\hat{\mu}_{aipw}(f,p)\right\}.$$

The first component is

$$\mathbb{V}_1\left\{\hat{\mu}_{aipw}(f,p)\right\} = \mathbb{E}_q\mathbb{E}_m\mathbb{V}_p\left\{\hat{\mu}_{aipw}(f,p)\right\}.$$

Since $\hat{\mu}_{aipw}(f,p)$ is a Horvitz–Thompson-type estimator applied to the pseudo-values η_k,

$$\mathbb{V}_1\left\{\hat{\mu}_{aipw}(f,p)\right\} = \frac{1}{N^2}\mathbb{E}_q\mathbb{E}_m\left[\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{\eta_k}{\pi_k}\frac{\eta_l}{\pi_l}\right].$$

For $k \neq l$, the model and response assumptions imply

$$\mathbb{E}_{mq}(\eta_k\eta_l) = f(\boldsymbol{v}_k)f(\boldsymbol{v}_l),$$

whereas

$$\mathbb{E}_{mq}(\eta_k^2) = f(\boldsymbol{v}_k)^2 + \frac{\sigma_k^2}{p(\boldsymbol{v}_k)}.$$

Using

$$\Delta_{kk} = \pi_k(1-\pi_k),$$

we obtain

$$\mathbb{V}_1\left\{\hat{\mu}_{aipw}(f,p)\right\} = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{f(\boldsymbol{v}_k)}{\pi_k}\frac{f(\boldsymbol{v}_l)}{\pi_l} + \frac{1}{N^2}\sum_{k\in U}\frac{w_k-1}{p(\boldsymbol{v}_k)}\sigma_k^2. \tag{3.2.2}$$

For the second component,

$$\begin{aligned}\mathbb{E}_p\left\{\hat{\mu}_{aipw}(f,p)-\mu\right\} &= \frac{1}{N}\sum_{k\in U}\left[f(\boldsymbol{v}_k) + \frac{r_k}{p(\boldsymbol{v}_k)}\left\{y_k - f(\boldsymbol{v}_k)\right\} - y_k\right] \\ &= \frac{1}{N}\sum_{k\in U}\left\{\frac{r_k}{p(\boldsymbol{v}_k)} - 1\right\}\left\{y_k - f(\boldsymbol{v}_k)\right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{V}_2\left\{\hat{\mu}_{aipw}(f,p)\right\} &= \mathbb{E}_q\mathbb{V}_m\mathbb{E}_p\left\{\hat{\mu}_{aipw}(f,p)-\mu\right\}\\ &= \frac{1}{N^2}\sum_{k\in U}\frac{1-p(\boldsymbol{v}_k)}{p(\boldsymbol{v}_k)}\sigma_k^2. \end{aligned} \tag{3.2.3}$$

Adding (3.2.2) and (3.2.3) gives (3.2.1).

□

Provided that $\pi_{kl} > 0$, an unbiased estimator of (3.2.1), when the conditional variances σ_k^2 are known, is

$$\widehat{\mathbb{V}}_{\text{tot,ra}}\left\{\hat{\mu}_{aipw}(f,p)\right\} = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{\eta_k}{\pi_k}\frac{\eta_l}{\pi_l} + \frac{1}{N^2}\sum_{k\in S_r} w_k\frac{1-p(\boldsymbol{v}_k)}{p(\boldsymbol{v}_k)^2}\sigma_k^2.$$

The second component of the reverse variance is of order N^{-1} under standard boundedness conditions. Consequently, when the sampling fraction is negligible, $n/N \longrightarrow 0$, its contribution is negligible relative to the first component, which is typically of order n^{-1}. This is the same simplification discussed in Chapter 2; see Shao and Steel (1999). In the AIPW and machine-learning setting, this point is discussed further by Dagdoug and Haziza (2026b).

3.2.4 The feasible AIPW estimator

In practice, neither $f(\cdot)$ nor $p(\cdot)$ is known. Let $\hat{f}(\cdot)$ and $\hat{p}(\cdot)$ denote estimators of these two nuisance functions. The two functions do not need to be estimated using the same method. For example, the outcome regression may be estimated using a random forest, while the response propensity may be estimated using a machine learning method appropriate for a binary response.

Define

$$\hat{\eta}_k = \hat{f}(\boldsymbol{v}_k) + \frac{r_k}{\hat{p}(\boldsymbol{v}_k)}\left\{y_k - \hat{f}(\boldsymbol{v}_k)\right\}.$$

The feasible AIPW estimator is

$$\hat{\mu}_{aipw}(\hat{f}, \hat{p}) = \frac{1}{N} \sum_{k \in S} w_k \hat{\eta}_k.$$

To understand the effect of estimating the nuisance functions, we compare the feasible estimator with its Oracle counterpart. We can write

$$\hat{\mu}_{aipw}(\hat{f}, \hat{p}) = \hat{\mu}_{aipw}(f, p) + R_{2,n}(\hat{f}, \hat{p}),$$

where

$$R_{2,n}(\hat{f}, \hat{p}) = T_{1n} + T_{2n} + T_{3n},$$

with

$$\begin{aligned}
T_{1n} &= \frac{1}{N} \sum_{k \in S} w_k \left\{ \hat{f}(\boldsymbol{v}_k) - f(\boldsymbol{v}_k) \right\} \left\{ 1 - \frac{r_k}{p(\boldsymbol{v}_k)} \right\}, \\
T_{2n} &= \frac{1}{N} \sum_{k \in S} w_k r_k \left\{ y_k - f(\boldsymbol{v}_k) \right\} \left\{ \frac{p(\boldsymbol{v}_k) - \hat{p}(\boldsymbol{v}_k)}{p(\boldsymbol{v}_k)\hat{p}(\boldsymbol{v}_k)} \right\}, \\
T_{3n} &= -\frac{1}{N} \sum_{k \in S} w_k r_k \left\{ \hat{f}(\boldsymbol{v}_k) - f(\boldsymbol{v}_k) \right\} \left\{ \frac{p(\boldsymbol{v}_k) - \hat{p}(\boldsymbol{v}_k)}{p(\boldsymbol{v}_k)\hat{p}(\boldsymbol{v}_k)} \right\}.
\end{aligned}$$

This decomposition is useful because it separates first- and second-order effects. The term T_{1n} is linear in the estimation error $\hat{f} - f$, while T_{2n} is linear in the estimation error $\hat{p} - p$. In contrast, T_{3n} involves the product of the two nuisance estimation errors and is therefore a second-order term.

For the feasible estimator to be asymptotically equivalent to the Oracle estimator, we require

$$R_{2,n}(\hat{f}, \hat{p}) = o_p(n^{-1/2}). \tag{3.2.4}$$

Under (3.2.4),

$$\sqrt{n} \left\{ \hat{\mu}_{aipw}(\hat{f}, \hat{p}) - \hat{\mu}_{aipw}(f, p) \right\} = o_p(1),$$

and the feasible estimator inherits the root-n consistency and asymptotic normality of the Oracle AIPW estimator.

When flexible machine learning methods are fitted and evaluated on the same observations, however, Neyman orthogonality alone is generally not sufficient to guarantee (3.2.4). The estimation errors $\hat{f}(\boldsymbol{v}_k) - f(\boldsymbol{v}_k)$ and $\hat{p}(\boldsymbol{v}_k) - p(\boldsymbol{v}_k)$ remain statistically related to the same response indicators and outcomes that appear in the AIPW estimating equation. This data reuse can prevent the first-order terms T_{1n} and T_{2n} from being sufficiently small. It can also lead to overfitting and artificially small fitted residuals, which is problematic for plug-in variance estimation. This issue is emphasized in Dagdoug and Haziza (2026b).

For this reason, simply replacing f and p by flexible machine learning estimators in the Oracle AIPW formula does not, by itself, provide the full DML result. Cross-fitting is needed to reduce the dependence between nuisance estimation and evaluation.

3.2.5 Cross-fitting and the product-rate condition

Let the realized sample S be partitioned into J disjoint folds $S_1, \ldots, S_J$. For a sampled unit k, let $j(k)$ denote the fold containing unit k, and define $S_{-j(k)} = S \setminus S_{j(k)}$. For each fold, the nuisance functions are estimated using only observations outside that fold. We denote the resulting estimators by $\hat{f}^{(-j(k))}(\cdot)$ and $\hat{p}^{(-j(k))}(\cdot)$. The outcome regression is fitted using the respondents in the training folds, whereas the response propensity is fitted using the response indicators available for all sampled units in the training folds.

The cross-fitted pseudo-value for unit k is

$$\hat{\eta}_k^{cf} = \hat{f}^{(-j(k))}(\boldsymbol{v}_k) + \frac{r_k}{\hat{p}^{(-j(k))}(\boldsymbol{v}_k)} \left\{ y_k - \hat{f}^{(-j(k))}(\boldsymbol{v}_k) \right\}.$$

The cross-fitted AIPW estimator is then

$$\hat{\mu}_{aipw}^{cf} = \frac{1}{N} \sum_{k \in S} w_k \hat{\eta}_k^{cf}.$$

The purpose of cross-fitting is that the prediction used for unit k is obtained from a model that was not trained using unit k. This substantially reduces the dependence responsible for the first-order terms in the remainder. Combined with Neyman orthogonality, the leading contribution of nuisance estimation is then governed by the interaction between

the two nuisance estimation errors.

The relevant convergence condition is therefore a product-rate condition. Using an appropriate L_2 norm, assume that

$$\left\|\hat{f} - f\right\|_2 \|\hat{p} - p\|_2 = o_p(n^{-1/2}). \tag{3.2.5}$$

Under suitable regularity conditions and cross-fitting, (3.2.5) implies

$$\hat{\mu}^{cf}_{aipw} - \hat{\mu}_{aipw}(f, p) = o_p(n^{-1/2}).$$

Consequently,

$$\sqrt{n}\left\{\hat{\mu}^{cf}_{aipw} - \hat{\mu}_{aipw}(f, p)\right\} = o_p(1),$$

and the cross-fitted estimator is asymptotically equivalent to the Oracle AIPW estimator. This is the main result underlying the use of DML for item nonresponse in the survey-sampling framework considered here; see Dagdoug and Haziza (2026b).

An important feature of (3.2.5) is that neither nuisance function needs to be estimated at the parametric rate $n^{-1/2}$. For example, if

$$\|\hat{f} - f\|_2 = O_p(n^{-a}) \qquad \text{and} \qquad \|\hat{p} - p\|_2 = O_p(n^{-b}),$$

then the product-rate condition is satisfied whenever

$$a + b > \frac{1}{2}.$$

Thus, two nuisance estimators that each converge relatively slowly may still lead to valid root-n inference when their errors decrease sufficiently fast in combination.

As discussed previously, the rates of nonparametric and machine learning estimators are often difficult to determine in practice. They may depend on the smoothness of the unknown nuisance functions, the dimension of the auxiliary variables, and the particular architecture and tuning parameters used; see, for example, Stone (1982). Hence, the product-rate condition is theoretically transparent but may be difficult to verify empirically for a given machine learning procedure.

We now consider variance estimation. In what follows, we assume that the sampling fraction is negligible, that is, $n/N \longrightarrow 0$. Under this assumption, the first-order variance of the cross-fitted AIPW estimator can be estimated by applying the usual Horvitz–Thompson variance estimator to the cross-fitted pseudo-values $\hat{\eta}_k^{cf}$. This gives

$$\widehat{\mathbb{V}}_{\text{tot}}\left\{\hat{\mu}_{aipw}^{cf}\right\} = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{\eta}_k^{cf}}{\pi_k}\frac{\hat{\eta}_l^{cf}}{\pi_l}.$$

Under suitable regularity conditions, this estimator is consistent for the first-order variance of the cross-fitted AIPW estimator; see Dagdoug and Haziza (2026b).

We can now establish the asymptotic distribution of the cross-fitted AIPW estimator. Recall that, under the product-rate condition,

$$\left\|\hat{f} - f\right\|_2 \|\hat{p} - p\|_2 = o_p(n^{-1/2}),$$

cross-fitting and Neyman orthogonality imply that

$$\hat{\mu}_{aipw}^{cf} - \hat{\mu}_{aipw}(f,p) = o_p(n^{-1/2}). \tag{3.2.6}$$

Therefore,

$$\sqrt{n}\left\{\hat{\mu}_{aipw}^{cf} - \hat{\mu}_{aipw}(f,p)\right\} = o_p(1).$$

Thus, the estimation of the nuisance functions has no first-order effect on the cross-fitted AIPW estimator.

Suppose now that the Oracle AIPW estimator satisfies a Central Limit Theorem, so that

$$\frac{\hat{\mu}_{aipw}(f,p) - \mu}{\sqrt{\mathbb{V}_{\text{tot}}\left\{\hat{\mu}_{aipw}(f,p)\right\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

We can write

$$\sqrt{n}\left\{\hat{\mu}_{aipw}^{cf} - \mu\right\} = \sqrt{n}\left\{\hat{\mu}_{aipw}(f,p) - \mu\right\} + \sqrt{n}\left\{\hat{\mu}_{aipw}^{cf} - \hat{\mu}_{aipw}(f,p)\right\}.$$

The second term is $o_p(1)$ by (3.2.6). Consequently, by Slutsky's theorem, the cross-fitted AIPW estimator has the same first-order asymptotic distribution as the Oracle AIPW estimator. In particular,

$$\frac{\hat{\mu}^{cf}_{aipw} - \mu}{\sqrt{\widehat{\mathbb{V}}_{\text{tot}}\left\{\hat{\mu}^{cf}_{aipw}\right\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Therefore, an asymptotically valid $100(1-\alpha)\%$ confidence interval for the population mean is

$$\hat{\mu}^{cf}_{aipw} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathbb{V}}_{\text{tot}}\left\{\hat{\mu}^{cf}_{aipw}\right\}}.$$

3.3 Simulation study

We conducted a simulation study with two main objectives. The first objective was to illustrate the potentially slow convergence of the usual imputed estimator $\hat{\mu}_I(\hat{f})$ when the imputation function is estimated using machine learning methods. The second objective was to assess the performance of the cross-fitted AIPW estimator introduced in the previous section.

3.3.1 Simulation setup

For each unit k in the finite population, let $\boldsymbol{v}_k = (v_{k1}, \ldots, v_{k10})^\top$ denote a vector of $p = 10$ auxiliary variables. We considered three population sizes, $N \in \{2{,}500, 5{,}000, 10{,}000\}$. For each population size, the auxiliary variables were generated independently according to

$$v_{kj} \sim \mathcal{U}(0, 100), \qquad j = 1, \ldots, 10, \qquad k = 1, \ldots, N.$$

The auxiliary variables were generated once for each population size and were then kept fixed throughout the Monte Carlo experiment. For each value of N, we performed $B = 10{,}000$ Monte Carlo replications.

At each replication, two survey variables were generated. The first variable followed a linear model,

$$y_{1k} = 0.5v_{k3} + 1.1v_{k5} + 2.7v_{k8} + \varepsilon_{1k},$$

where

$$\varepsilon_{1k} \sim \mathcal{N}(0, 50^2).$$

The second survey variable followed a nonlinear model,

$$y_{2k} = 10\,1(v_{k1} > 50) + \log(v_{k5}) + \sqrt{v_{k8}v_{k9}} + \varepsilon_{2k},$$

where

$$\varepsilon_{2k} \sim \mathcal{N}(0, 15^2).$$

The error variances were chosen so that the coefficient of determination was approximately $R^2 = 0.7$ for both survey variables. The two variables provide two different situations for imputation. For y_1, the conditional mean is linear in the auxiliary variables, so linear regression imputation is correctly specified. In contrast, the conditional mean of y_2 contains nonlinear terms, and a standard linear regression imputation model is misspecified.

For each Monte Carlo replication, the finite population means

$$\mu_1 = \frac{1}{N}\sum_{k\in U} y_{1k} \qquad \text{and} \qquad \mu_2 = \frac{1}{N}\sum_{k\in U} y_{2k}$$

were computed before sampling.

A sample was then selected by simple random sampling without replacement. The sampling fraction was fixed at 10%, so that $n/N = 0.10$. As a result, the three population sizes corresponded to sample sizes $n \in \{250, 500, 1{,}000\}$. This allowed us to examine how the behaviour of the estimators changed as the sample size increased while keeping the sampling fraction constant.

After selecting the sample, item nonresponse was generated independently for the sampled units. The response probability for unit k was

$$p(\boldsymbol{v}_k) = \frac{\exp\left(-0.0132v_{k3} + 0.01202v_{k5} + 0.00106v_{k8}\right)}{1 + \exp\left(-0.0132v_{k3} + 0.01202v_{k5} + 0.00106v_{k8}\right)}.$$

The coefficients were chosen so that the expected response rate was approximately 50%. Since the response probability depends only on the fully observed auxiliary variables,

the response mechanism is MAR. When the response probabilities were estimated, the estimated values were truncated from below, $\hat{p}_k \leftarrow \max(\hat{p}_k, 0.01)$, so as to avoid extreme inverse-probability weights.

Several methods were used to estimate the conditional mean function $f(\boldsymbol{v}_k)$. Linear regression imputation was included as a benchmark. We also considered random forest, K-nearest neighbours, Cubist, and gradient boosting.

For random forest, we used 500 trees, a minimum terminal-node size of 5, and `mtry` = 8. For K-nearest neighbours, we used $K = 12$. For Cubist, we used 10 committees, with 3 neighbours when $n = 250$ or 500, and no nearest-neighbour adjustment when $n = 1{,}000$. Gradient boosting was fitted using 500 trees, an interaction depth of 1, and a shrinkage parameter equal to 0.05.

For the cross-fitted AIPW estimator, the sample was randomly partitioned at each replication into $J = 5$ folds of (almost) equal size, without stratification on the response indicator.

3.3.2 Slow convergence of the imputed estimator

The first part of the simulation study was designed to illustrate the convergence behaviour of the usual imputed estimator $\hat{\mu}_I(\hat{f})$. For each imputation method, the missing values were replaced by the corresponding predictions

$$y_k^* = \hat{f}(\boldsymbol{v}_k), \qquad k \in S_m,$$

and the population mean was estimated by

$$\hat{\mu}_I(\hat{f}) = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k \hat{f}(\boldsymbol{v}_k)\right\}.$$

For each survey variable and each imputation method, we examined the Monte Carlo distribution of the estimation error $\hat{\mu}_I(\hat{f}) - \mu$ for the three sample sizes. If the estimator is consistent, this distribution should become increasingly concentrated around zero as n increases. To investigate whether the estimator converges at the root-n rate, we also

examined the scaled error $\sqrt{n}\left\{\hat{\mu}_I(\hat{f}) - \mu\right\}$. This scaling provides a simple way to assess root-n convergence. If $\hat{\mu}_I(\hat{f})$ is root-n consistent and has a stable first-order limiting distribution, the distribution of $\sqrt{n}\{\hat{\mu}_I(\hat{f}) - \mu\}$ should remain approximately stable as n increases. In contrast, if the estimation error decreases more slowly than $n^{-1/2}$, multiplying it by $\sqrt{n}$ will magnify the remaining error and the distribution of the scaled estimator will not stabilize.

The distributions of the unscaled errors and scaled errors are displayed in Figures 3.1 and 3.2, respectively. The comparison across the three sample sizes allows us to assess both consistency and the rate at which the different imputation estimators approach the true finite population mean.

First, as shown in Figure 3.1, when the true model is linear, linear regression imputation (LRI) yields errors centered around zero, which indicates a small bias in these experiments. Moreover, the variance decreases as the sample size increases, which is consistent with the consistency of the estimator. In contrast, using LRI for the nonlinear model results in a noticeable bias and an overestimation of the true mean even though the variance decreases as the sample size increases.

Now, we consider the estimators based on ML methods. The imputed estimators based on RF, KNN, and Cubist imputation present a non-negligible bias for both the linear and the nonlinear variables, especially KNN, for which the estimates remain far from the true value for all the sample sizes considered. Imputation based on Cubist performs well for the nonlinear variable for the larger sample size $n = 1,000$. Overall, for all methods, the bias decreases with the sample size, but the decrease appears slow. Boosting imputation produces better results in terms of bias for the linear variable, since the errors are close to zero for each sample size. This is not the case for the nonlinear variable, for which the bias is substantial and decreases slowly, as observed for the RF, KNN and Cubist cases.

Figure 3.2 provides more details about the behaviour of the different methods as the sample size grows. When the true model is linear, LRI produces errors centered around zero. The width of the boxplots remains approximately constant, which is consistent with a variance of the imputed estimator decreasing at a rate of $\frac{1}{n}$, and hence with $\sqrt{n}$-consistency. In contrast, when LRI is applied to a nonlinear variable, the variance remains

approximately stable, but the errors increase with the sample size. This suggests that the imputed estimator is biased and that the bias does not shrink as the sample size grows, which is consistent with the misspecification of the first moment in our imputation model. Considering the ML methods, the errors tend to increase with the sample size in the KNN, RF, Cubist (linear variable), and Boosting (nonlinear variable) cases. This is consistent with a slow convergence of the imputed estimators, which do not appear to reach the target rate of $\frac{1}{\sqrt{n}}$ in these experiments.

3.3.3 Behaviour of the cross-fitted AIPW estimator

We compared two estimators of the mean: the cross-fitted AIPW estimator $\hat{\mu}^{cf}_{aipw}$ and the feasible AIPW estimator (without cross-fitting), $\hat{\mu}_{aipw}(\hat{f}, \hat{p})$.

For the outcome model $f(\cdot)$, we used random forest, Cubist, and gradient boosting. For the response propensity $p(\cdot)$, we used random forest, gradient boosting, and classification trees.

For each estimator of the mean, the Monte Carlo percent relative bias (RB) is reported alongside the relative efficiency (RE), that is, the ratio of its Monte Carlo variance to that of the Horvitz–Thompson estimator under full response. For each variance estimator, the Monte Carlo percent relative bias of the variance estimator is reported alongside the empirical coverage rate of the 95% confidence interval.

Results for each variable are shown in Table 3.1 and Table 3.2. For both mean estimators, the percent relative bias is close to zero across all sample sizes and method combinations, which indicates a small bias in these experiments. Regarding the variance estimators, we note a substantial underestimation of the variance of the feasible AIPW estimator $\hat{\mu}_{aipw}(\hat{f}, \hat{p})$, reaching as low as -30% for some combinations. This is consistent with overfitting: since the machine learning models were trained on the same data used for prediction, the residuals are overly optimistic, and the coverage rates fall below 95%. In contrast, the variance estimator for the cross-fitted AIPW estimator $\hat{\mu}^{cf}_{aipw}$ exhibits percent relative biases close to zero across all combinations, which is reflected in coverage rates close to the nominal 95% level.

Table 3.1: Monte Carlo results for the linear variable y_1.

		Feasible AIPW				Cross-fitted AIPW			
Outcome / response	n	RB (%)	RE	VRB (%)	CR (%)	RB (%)	RE	VRB (%)	CR (%)
Random Forest / Random Forest									
	250	0.2	129	−26.3	91.3	−0.4	157	3.7	95.4
	500	0.3	127	−22.4	91.6	−0.2	142	3.0	95.5
	1,000	0.3	130	−24.2	91.0	−0.2	138	1.2	94.8
Random Forest / Classification tree									
	250	0.0	129	−21.0	92.1	−0.7	146	0.6	94.5
	500	0.0	126	−14.1	93.3	−0.4	137	−0.6	94.8
	1,000	0.1	128	−16.3	92.5	−0.3	139	−0.5	94.4
Cubist / Random Forest									
	250	0.1	130	−18.2	92.5	−0.1	145	3.1	95.3
	500	0.0	125	−16.3	93.3	0.0	135	2.0	95.4
	1,000	0.0	124	−17.8	92.6	0.0	132	0.7	94.7
Cubist / Classification tree									
	250	0.0	130	−6.3	94.3	0.0	133	−0.3	95.1
	500	0.0	125	−4.7	94.6	0.0	128	−1.7	95.0
	1,000	0.0	128	−6.0	94.1	0.0	131	−0.9	94.7
Cubist / Gradient Boosting									
	250	0.1	130	−12.1	93.6	−0.1	145	2.6	95.4
	500	0.0	124	−6.8	94.1	0.0	128	0.8	95.2
	1,000	0.0	124	−7.5	94.1	0.0	125	−3.3	94.7
Gradient Boosting / Classification tree									
	250	−0.1	131	−11.3	93.5	−0.4	138	0.5	94.9
	500	−0.2	126	−7.4	94.1	−0.3	132	−0.7	95.0
	1,000	−0.1	129	−7.2	94.0	−0.3	135	0.6	94.6

Note: The first and second methods reported in each row group correspond, respectively, to the outcome regression and response propensity estimators. RB denotes the Monte Carlo percent relative bias of the point estimator; RE denotes its relative efficiency with respect to the Horvitz–Thompson estimator; VRB denotes the Monte Carlo percent relative bias of the variance estimator; and CR denotes the empirical coverage rate of the nominal 95% confidence interval.

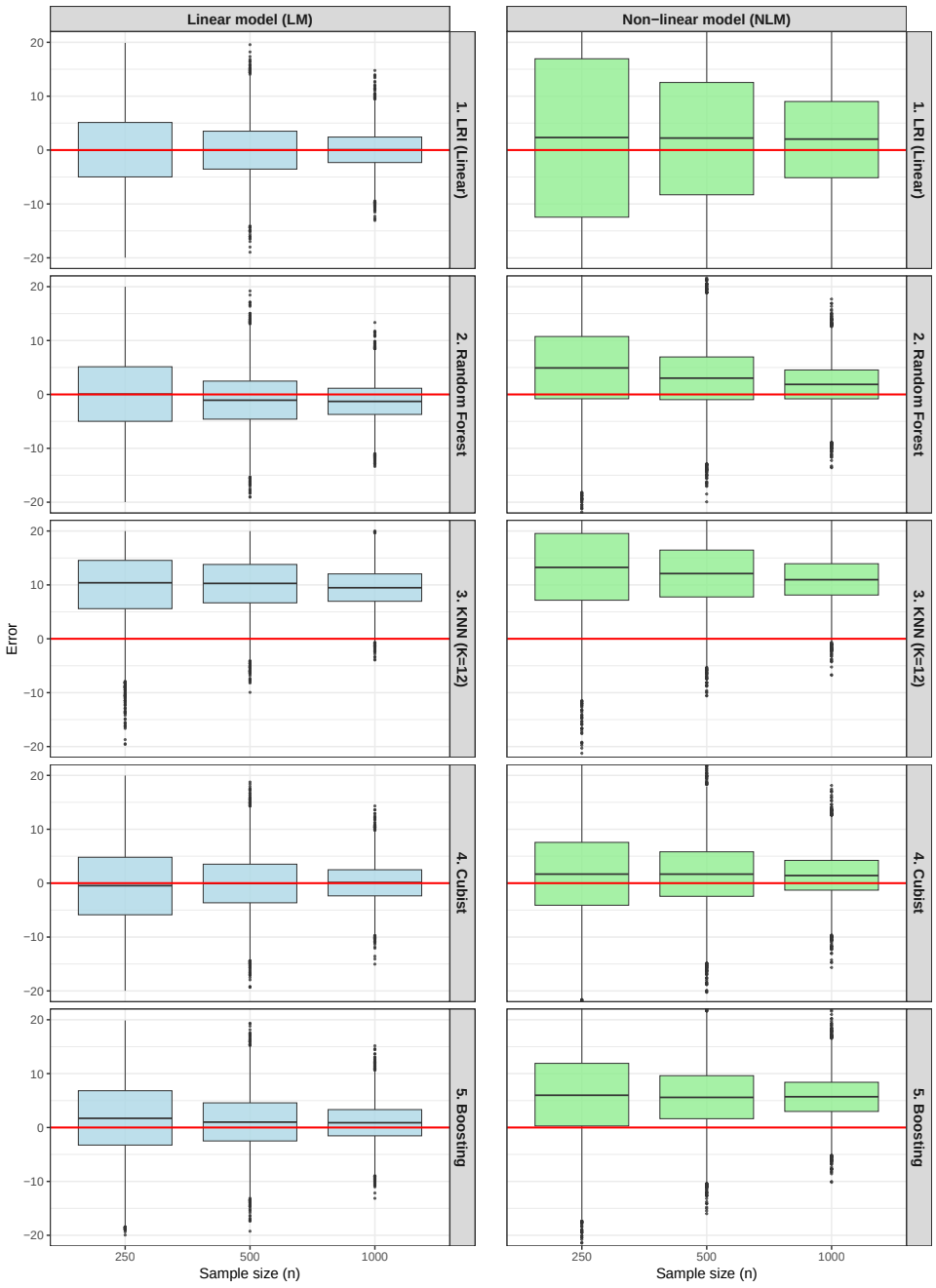

Figure 3.1: Distribution of the errors $\hat{\mu}_I - \mu$

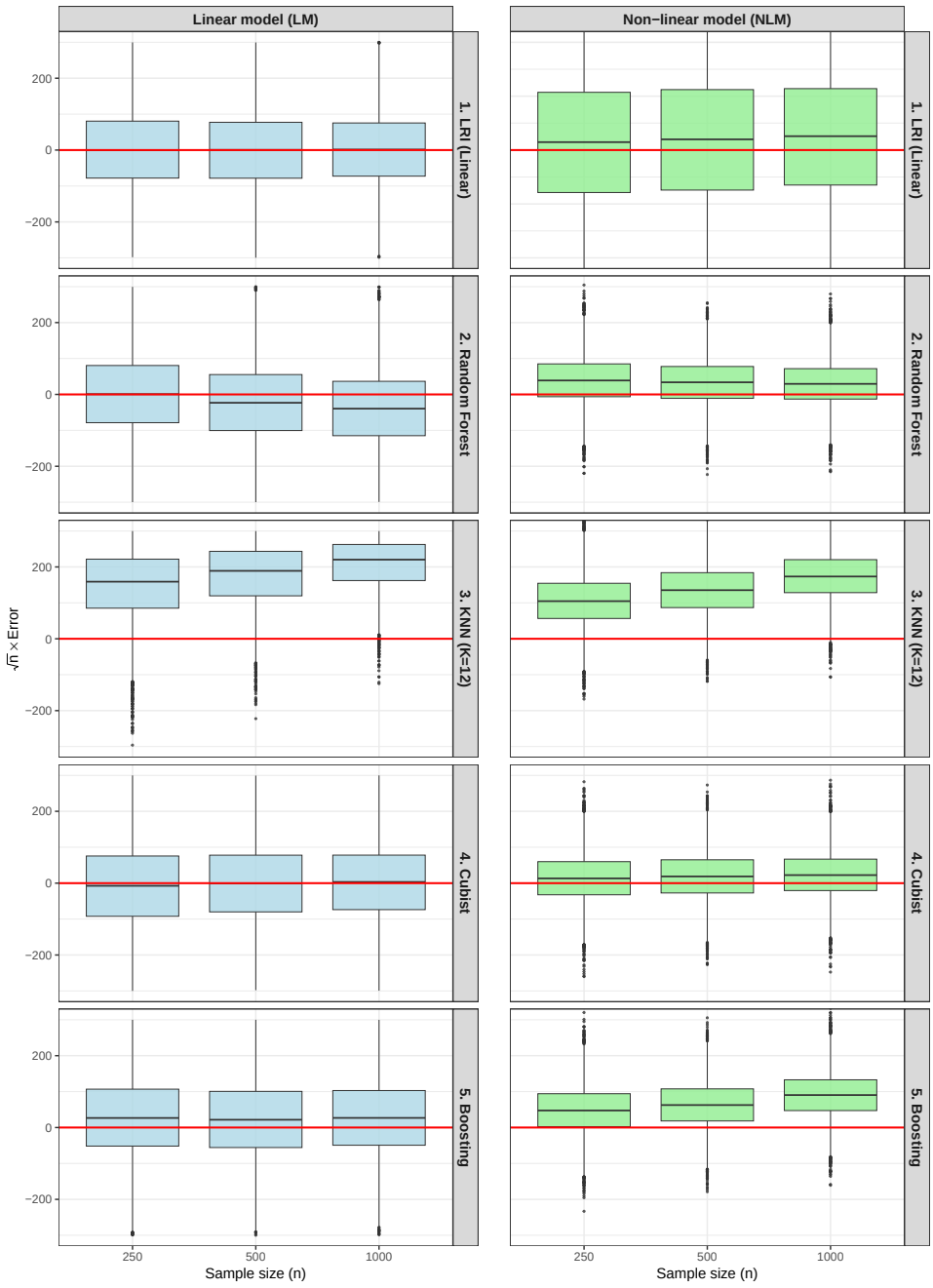

Figure 3.2: Distribution of the errors $\sqrt{n}(\hat{\mu}_I - \mu)$

Table 3.2: Monte Carlo results for the nonlinear variable y_2.

		Feasible AIPW				Cross-fitted AIPW			
Outcome / response	n	RB (%)	RE	VRB (%)	CR (%)	RB (%)	RE	VRB (%)	CR (%)
Random Forest / Random Forest									
	250	−0.3	131	−30.0	90.2	−0.1	160	3.8	94.8
	500	−0.3	125	−22.6	91.5	−0.1	144	5.5	95.5
	1,000	−0.2	126	−25.0	90.9	0.0	143	−0.7	94.7
Random Forest / Classification tree									
	250	−0.3	132	−24.1	91.4	−0.3	143	−0.1	94.6
	500	−0.4	128	−13.9	92.9	−0.3	135	1.2	95.1
	1,000	−0.2	129	−16.9	92.4	−0.1	139	−0.9	95.0
Cubist / Classification tree									
	250	0.0	135	−8.4	94.1	−0.2	142	−1.4	94.9
	500	−0.1	128	−4.9	94.2	−0.2	134	0.1	95.1
	1,000	0.0	133	−8.0	94.0	−0.1	139	−1.9	94.6
Gradient Boosting / Random Forest									
	250	−0.1	132	−26.2	91.0	−0.1	159	2.7	94.6
	500	−0.2	125	−20.2	91.9	−0.1	141	4.5	95.5
	1,000	−0.1	126	−21.9	91.5	0.0	141	−0.7	94.8
Gradient Boosting / Classification tree									
	250	−0.1	133	−17.5	92.7	−0.2	143	−1.6	94.6
	500	−0.2	127	−9.6	93.5	−0.2	133	0.2	95.1
	1,000	−0.1	131	−10.1	93.4	−0.1	137	−1.2	94.7
Gradient Boosting / Gradient Boosting									
	250	−0.1	132	−21.7	92.0	0.0	158	3.8	95.3
	500	−0.2	126	−12.2	93.1	−0.1	134	2.1	95.2
	1,000	−0.1	127	−11.8	93.4	0.0	131	−4.6	94.5

Note: The first and second methods reported in each row group correspond, respectively, to the outcome regression and response propensity estimators. RB denotes the Monte Carlo percent relative bias of the point estimator; RE denotes its relative efficiency with respect to the Horvitz–Thompson estimator; VRB denotes the Monte Carlo percent relative bias of the variance estimator; and CR denotes the empirical coverage rate of the nominal 95% confidence interval.

4

Imputation of categorical variables

In Chapter 3, we showed how double/debiased machine learning can be used to estimate a finite population mean in the presence of item nonresponse. We introduced the cross-fitted AIPW estimator and showed that, under suitable conditions, it is root-n consistent and asymptotically normal. We also presented an estimator of its variance.

In this chapter, we consider the problem of imputing a categorical variable. In practice, statistical agencies do not provide only an estimate of the population mean. They also produce a completed data file that can be used by other researchers and data users. In this file, the observed and imputed values of a survey variable appear in the same column.

Let y_k^* denote the imputed value for a nonrespondent. The completed value for sampled unit k is

$$\tilde{y}_k = r_k y_k + (1 - r_k) y_k^*.$$

Thus, $\tilde{y}_k = y_k$ when unit k is a respondent, and $\tilde{y}_k = y_k^*$ when unit k is a nonrespondent. A data user should be able to estimate the population mean directly from this completed

variable. The corresponding estimator is

$$\hat{\mu}_I = \frac{1}{N}\sum_{k\in S} w_k \tilde{y}_k = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k y_k^*\right\}. \tag{4.0.1}$$

The user should not need to estimate the outcome model again, estimate the response probabilities, or reconstruct the cross-fitting procedure. The estimate should be obtained simply by using the completed values and the survey weights.

The cross-fitted AIPW estimator introduced in Chapter 3 does not have this convenient form. It depends on the estimated outcome function, the estimated response probabilities, and an adjustment term. It gives a valid estimate of the population mean, but a user cannot obtain it simply by taking the weighted mean of one completed variable. Our first objective is therefore to construct a completed data file whose usual weighted mean is equal, or very close, to the cross-fitted AIPW estimate.

A second problem appears when the survey variable is categorical. Consider, for example, a binary variable taking the values zero and one. The imputed values stored in the completed file should also be zeros and ones. A prediction such as $\hat{f}(\boldsymbol{v}_k) = 0.72$ may be interpreted as an estimated probability, but it is not a possible value of the original binary variable. We cannot record that an individual has an employment status equal to 0.72, for example.

Stochastic imputation provides one possible solution. If $\hat{f}(\boldsymbol{v}_k)$ is the estimated probability that $y_k = 1$, we can generate

$$y_k^* \sim \text{Bernoulli}\left\{\hat{f}(\boldsymbol{v}_k)\right\}.$$

The imputed value is then either zero or one. However, ordinary Bernoulli imputation introduces additional variability. Moreover, the weighted mean of the completed values will not generally be equal to the cross-fitted AIPW estimator.

The objective of this chapter is to solve these two problems at the same time. We want to generate imputed values that belong to the original categories, and we want the usual estimator computed from the completed file to remain close to the cross-fitted AIPW

estimator.

We will use rejective stochastic imputation for this purpose. We first generate possible imputed values from Bernoulli distributions. We then keep only the imputed vectors for which the weighted mean of the completed data is sufficiently close to the cross-fitted AIPW estimate. In this way, the completed file contains only zeros and ones, while the estimator obtained from that file keeps the good statistical properties of the cross-fitted AIPW estimator.

We begin with the case of a binary variable. We first review deterministic and stochastic imputation for this type of variable. We then explain why ordinary stochastic imputation is not sufficient for our purpose and introduce the rejective imputation procedure. Finally, we extend the method to categorical variables with more than two classes.

4.1 Imputation for a binary variable

Let $y_k \in \{0,1\}, k \in U$. The two values indicate whether unit k has a particular characteristic. For example, in a survey on employment, we may define

$$y_k = \begin{cases} 1, & \text{if individual } k \text{ is employed,} \\ 0, & \text{if individual } k \text{ is not employed.} \end{cases}$$

The finite population mean is

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

In this example, μ is the proportion of employed individuals in the finite population. More generally, for any binary variable, μ is the population proportion of units for which $y_k = 1$.

Since the original variable is binary, we require $y_k^* \in \{0,1\}, k \in S_m$. The completed value is

$$\tilde{y}_k = r_k y_k + (1 - r_k) y_k^*.$$

Therefore, every value in the completed column also belongs to $\{0, 1\}$. The imputed estimator of the population proportion is given by (4.0.1).

For a binary variable,

$$f(\boldsymbol{v}_k) = \mathbb{E}_m\left(y_k \mid \boldsymbol{v}_k\right) = P\left(y_k = 1 \mid \boldsymbol{v}_k\right)$$

is the probability that unit k has the characteristic of interest, given its auxiliary information. Logistic regression and machine learning classification methods can be used to estimate this probability.

We first consider Oracle imputation, where the true conditional probability $f(\boldsymbol{v}_k)$ is assumed to be known.

4.1.1 Oracle imputation for a binary variable

We start by describing deterministic Oracle imputation. In the Oracle setting, the true conditional mean function is assumed to be known. This assumption is not realistic in practice because the relationship between the survey variable and the auxiliary variables is generally unknown. However, the Oracle setting provides a useful benchmark for studying the properties of feasible imputation methods.

Under deterministic Oracle imputation, the missing value for a nonrespondent $k \in S_m$ is replaced by

$$y_k^* = f(\boldsymbol{v}_k), \quad k \in S_m.$$

The corresponding Oracle imputed estimator of the population mean is

$$\hat{\mu}_{\text{oracle}} = \frac{1}{N}\left\{\sum_{k \in S_r} w_k y_k + \sum_{k \in S_m} w_k f(\boldsymbol{v}_k)\right\}.$$

As shown in Chapter 2, under the assumptions considered there, $\hat{\mu}_{\text{oracle}}$ is mpq-unbiased for μ and is root-n consistent.

The limitation of deterministic Oracle imputation is that the imputed values $f(\boldsymbol{v}_k)$ generally belong to $(0,1)$, whereas the original variable takes values only in $\{0,1\}$.

To overcome this issue, we turn to stochastic Oracle imputation. For every nonrespondent $k \in S_m$, we generate

$$y_k^* \sim \text{Bernoulli}\left\{f(\boldsymbol{v}_k)\right\}.$$

The imputed values are generated independently, conditionally on the realized sample, the response pattern, the auxiliary variables, and the observed survey values. The resulting estimator is

$$\hat{\mu}_{\text{oracle}}^{\text{stoch}} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k y_k^*\right\}.$$

Unlike deterministic Oracle imputation, stochastic Oracle imputation produces only binary imputed values. It therefore preserves the support of the original variable.

We denote by $\mathbb{E}_I(\cdot)$ and $\mathbb{V}_I(\cdot)$ the expectation and variance with respect to the imputation mechanism. When these operators are used, the realized sample, the response indicators, the auxiliary variables, the observed survey values, and the function $f(\cdot)$ are treated as fixed. Only the imputed values y_k^* are treated as random. We write $\mathbb{E}_{mpqI}(\cdot) = \mathbb{E}_m\mathbb{E}_p\mathbb{E}_q\mathbb{E}_I(\cdot)$ for the expectation with respect to the model m, the sampling design p, the nonresponse mechanism q and the imputation mechanism I, and refer to it as the mpqI framework.

Since

$$y_k^* \sim \text{Bernoulli}\left\{f(\boldsymbol{v}_k)\right\},$$

we have

$$\mathbb{E}_I(y_k^*) = f(\boldsymbol{v}_k).$$

It follows that

$$\begin{aligned}
\mathbb{E}_I\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right) &= \frac{1}{N}\mathbb{E}_I\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k y_k^*\right\} \\
&= \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k \mathbb{E}_I(y_k^*)\right\} \\
&= \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k f(\boldsymbol{v}_k)\right\} \\
&= \hat{\mu}_{\text{oracle}}.
\end{aligned}$$

Thus, conditionally on the observed data, the stochastic Oracle estimator is centered on the deterministic Oracle estimator.

Proposition 4.1. *Under the mpqI framework, the stochastic Oracle estimator* $\hat{\mu}_{\text{oracle}}^{\text{stoch}}$ *is unbiased for the finite population mean* μ.

Proof. Using the previous result,

$$\begin{aligned}
\mathbb{E}_{mpqI}\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}} - \mu\right) &= \mathbb{E}_m\mathbb{E}_p\mathbb{E}_q\mathbb{E}_I\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}} - \mu\right) \\
&= \mathbb{E}_m\mathbb{E}_p\mathbb{E}_q\left(\hat{\mu}_{\text{oracle}} - \mu\right) \\
&= 0.
\end{aligned}$$

The last equality follows from the *mpq*-unbiasedness of the deterministic Oracle estimator.

□

Stochastic imputation introduces an additional source of variability because different Bernoulli draws may produce different completed data files. We now show that this additional variability does not change the root-n rate of the Oracle estimator.

Proposition 4.2. *Suppose that the deterministic Oracle estimator is root-n consistent and that the sampling weights satisfy* $N^{-2}\sum_{k\in S_m} w_k^2 = O_p(n^{-1})$. *Then,*

$$\hat{\mu}_{\text{oracle}}^{\text{stoch}} - \mu = O_p(n^{-1/2}).$$

Thus, the stochastic Oracle estimator is also root-n consistent for μ.

Proof. The stochastic Oracle estimator can be written as

$$\hat{\mu}_{\text{oracle}}^{\text{stoch}} = \hat{\mu}_{\text{oracle}} + \frac{1}{N} \sum_{k \in S_m} w_k \left\{ y_k^* - f(\boldsymbol{v}_k) \right\}.$$

Consequently,

$$\hat{\mu}_{\text{oracle}}^{\text{stoch}} - \mu = (\hat{\mu}_{\text{oracle}} - \mu) + D_I,$$

where

$$D_I = \frac{1}{N} \sum_{k \in S_m} w_k \left\{ y_k^* - f(\boldsymbol{v}_k) \right\}.$$

From Chapter 2,

$$\hat{\mu}_{\text{oracle}} - \mu = O_p(n^{-1/2}).$$

Now, conditionally on the observed data,

$$\mathbb{E}_I(D_I) = 0.$$

Since the imputed values are generated independently,

$$\mathbb{V}_I(D_I) = \frac{1}{N^2} \sum_{k \in S_m} w_k^2 \, f(\boldsymbol{v}_k)\{1 - f(\boldsymbol{v}_k)\} \leq \frac{1}{4N^2} \sum_{k \in S_m} w_k^2 = O_p(n^{-1}),$$

because $f(\boldsymbol{v}_k)\{1 - f(\boldsymbol{v}_k)\} \leq 1/4$. Hence, by Chebyshev's inequality applied conditionally on the observed data,

$$D_I = O_p(n^{-1/2}).$$

Combining this result with the root-n consistency of the deterministic Oracle estimator gives

$$\hat{\mu}_{\text{oracle}}^{\text{stoch}} - \mu = O_p(n^{-1/2}).$$

□

To make this additional variability explicit, we use the law of total variance with respect to the imputation mechanism. Recall that $\mathbb{E}_I\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right) = \hat{\mu}_{\text{oracle}}$. Therefore, the total variance

of the stochastic Oracle estimator can be written as

$$\begin{aligned}\mathbb{V}_{\text{tot}}\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right) &= \mathbb{V}_{\text{tot,ra}}\left[\mathbb{E}_I\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right)\right] + \mathbb{E}_m\mathbb{E}_p\mathbb{E}_q\left[\mathbb{V}_I\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right)\right] \\ &= \mathbb{V}_{\text{tot,ra}}\left(\hat{\mu}_{\text{oracle}}\right) + \mathbb{V}_{\text{imp}}\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right), \end{aligned} \tag{4.1.1}$$

where

$$\mathbb{V}_{\text{imp}}\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right) = \mathbb{E}_m\mathbb{E}_p\mathbb{E}_q\left[\mathbb{V}_I\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right)\right].$$

The first term in (4.1.1) is the variance of the deterministic Oracle estimator derived in Chapter 2. It contains the variability due to the model, the sampling design, and the nonresponse mechanism. The second term, $\mathbb{V}_{\text{imp}}$, is the additional variance caused only by the random generation of the imputed values. It follows that

$$\begin{aligned}\mathbb{V}_I\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right) &= \mathbb{V}_I\left[\frac{1}{N}\sum_{k\in S_m} w_k y_k^*\right] \\ &= \frac{1}{N^2}\sum_{k\in S_m} w_k^2 \mathbb{V}_I(y_k^*) \\ &= \frac{1}{N^2}\sum_{k\in S_m} w_k^2 f(\boldsymbol{v}_k)\left\{1 - f(\boldsymbol{v}_k)\right\}. \end{aligned}$$

As a result,

$$\mathbb{V}_{\text{imp}}\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right) = \frac{1}{N^2}\mathbb{E}_m\mathbb{E}_p\mathbb{E}_q\left[\sum_{k\in S_m} w_k^2 f(\boldsymbol{v}_k)\left\{1 - f(\boldsymbol{v}_k)\right\}\right].$$

This expression measures how much variability is added when the fractional values $f(\boldsymbol{v}_k)$ are replaced by random zeros and ones. If deterministic Oracle imputation were used, this component would be equal to zero.

Since $f(\cdot)$ is known in the Oracle setting, an unbiased estimator of the additional imputation variance is

$$\widehat{\mathbb{V}}_{\text{imp}}\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right) = \frac{1}{N^2}\sum_{k\in S_m} w_k^2 f(\boldsymbol{v}_k)\left\{1 - f(\boldsymbol{v}_k)\right\}.$$

A variance estimator for the stochastic Oracle estimator is therefore

$$\widehat{\mathbb{V}}_{\text{tot}}\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right) = \widehat{\mathbb{V}}_{\text{det}}\left(\hat{\mu}_{\text{oracle}}\right) + \widehat{\mathbb{V}}_{\text{imp}}\left(\hat{\mu}_{\text{oracle}}^{\text{stoch}}\right),$$

where $\widehat{\mathbb{V}}_{\text{det}}(\hat{\mu}_{\text{oracle}})$ denotes an estimator of the total variance of the deterministic Oracle estimator. As shown in Chapter 2, this first component can be obtained either under the two-phase framework or under the reverse approach. More precisely, one may use

$$\widehat{\mathbb{V}}_{\text{det}}\left(\hat{\mu}_{\text{oracle}}\right) = \widehat{\mathbb{V}}_{\text{tot,tp}}\left(\hat{\mu}_{\text{oracle}}\right)$$

under the two-phase framework, or

$$\widehat{\mathbb{V}}_{\text{det}}\left(\hat{\mu}_{\text{oracle}}\right) = \widehat{\mathbb{V}}_{\text{tot,ra}}\left(\hat{\mu}_{\text{oracle}}\right)$$

under the reverse approach. These two approaches estimate the same total variance but use different variance decompositions.

The Oracle setting assumes that $f(\cdot)$ is known. In practice, this function must be estimated. We now consider feasible deterministic and stochastic imputation based on logistic regression.

4.1.2 Feasible imputation based on logistic regression

A standard parametric model for a binary variable is logistic regression. This is not the only possible model for a binary response, but it is commonly used because it guarantees that the fitted probabilities belong to the interval $(0, 1)$.

Under the logistic regression model,

$$f(\boldsymbol{v}_k; \boldsymbol{\beta}) = P\left(y_k = 1 \mid \boldsymbol{v}_k\right) = \frac{\exp\left(\boldsymbol{v}_k^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\boldsymbol{v}_k^\top \boldsymbol{\beta}\right)}, \tag{4.1.2}$$

where $\boldsymbol{\beta}$ is a finite-dimensional vector of regression parameters. The parameter $\boldsymbol{\beta}$ is esti-

mated by maximum likelihood using the respondents. The respondent log-likelihood is

$$\ell_r(\boldsymbol{\beta}) = \sum_{k \in S_r} \left[y_k \log \left\{ f(\boldsymbol{v}_k; \boldsymbol{\beta}) \right\} + (1 - y_k) \log \left\{ 1 - f(\boldsymbol{v}_k; \boldsymbol{\beta}) \right\} \right].$$

The maximum likelihood estimator is

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\beta}} \ell_r(\boldsymbol{\beta}).$$

There is generally no closed-form expression for $\hat{\boldsymbol{\beta}}_{\text{MLE}}$, so it is computed using numerical methods. If the sampling design is informative, a survey-weighted likelihood may be used instead.

For unit k, define the estimated conditional probability

$$\hat{f}_k = f\left(\boldsymbol{v}_k; \hat{\boldsymbol{\beta}}_{\text{MLE}}\right) = \frac{\exp\left(\boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{\text{MLE}}\right)}{1 + \exp\left(\boldsymbol{v}_k^\top \hat{\boldsymbol{\beta}}_{\text{MLE}}\right)}.$$

Under deterministic logistic regression imputation, the missing value for $k \in S_m$ is replaced by $y_k^* = \hat{f}_k$. The corresponding logistic regression imputed estimator is

$$\hat{\mu}_{\text{LGRI}} = \frac{1}{N} \left\{ \sum_{k \in S_r} w_k y_k + \sum_{k \in S_m} w_k \hat{f}_k \right\}.$$

This estimator is generally not exactly unbiased in finite samples because the unknown parameter $\boldsymbol{\beta}$ is estimated. Under a correctly specified logistic model and standard regularity conditions, however, it is root-n consistent.

Proposition 4.3. *Suppose that the logistic regression model* (4.1.2) *is correctly specified and that standard regularity conditions ensure that*

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} - \boldsymbol{\beta} = O_p(n^{-1/2}).$$

Suppose also that the weighted average of the auxiliary variables among the nonrespondents

remains bounded in probability. Then

$$\hat{\mu}_{\text{LGRI}} - \mu = O_p(n^{-1/2}).$$

Proof. Under the correctly specified logistic model, the deterministic Oracle estimator is

$$\hat{\mu}_{\text{oracle}} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k f(\boldsymbol{v}_k;\boldsymbol{\beta})\right\}.$$

We can therefore write

$$\hat{\mu}_{\text{LGRI}} - \hat{\mu}_{\text{oracle}} = \frac{1}{N}\sum_{k\in S_m} w_k \left\{f\left(\boldsymbol{v}_k;\hat{\boldsymbol{\beta}}_{\text{MLE}}\right) - f(\boldsymbol{v}_k;\boldsymbol{\beta})\right\}.$$

The logistic function is smooth with respect to $\boldsymbol{\beta}$. Its derivative is

$$\dot{f}(\boldsymbol{v}_k;\boldsymbol{\beta}) = f(\boldsymbol{v}_k;\boldsymbol{\beta})\left\{1 - f(\boldsymbol{v}_k;\boldsymbol{\beta})\right\}\boldsymbol{v}_k.$$

A first-order Taylor expansion gives

$$\hat{\mu}_{\text{LGRI}} - \hat{\mu}_{\text{oracle}} = \boldsymbol{A}_n^{\top}\left(\hat{\boldsymbol{\beta}}_{\text{MLE}} - \boldsymbol{\beta}\right) + o_p(n^{-1/2}),$$

where

$$\boldsymbol{A}_n = \frac{1}{N}\sum_{k\in S_m} w_k \dot{f}(\boldsymbol{v}_k;\boldsymbol{\beta}).$$

Under the mild regularity conditions, $\boldsymbol{A}_n = O_p(1)$. It follows that

$$\hat{\mu}_{\text{LGRI}} - \hat{\mu}_{\text{oracle}} = O_p(n^{-1/2}).$$

From Chapter 2,

$$\hat{\mu}_{\text{oracle}} - \mu = O_p(n^{-1/2}).$$

Using the decomposition

$$\hat{\mu}_{\text{LGRI}} - \mu = (\hat{\mu}_{\text{LGRI}} - \hat{\mu}_{\text{oracle}}) + (\hat{\mu}_{\text{oracle}} - \mu),$$

we conclude that

$$\hat{\mu}_{\text{LGRI}} - \mu = O_p(n^{-1/2}).$$

□

The deterministic procedure produces estimated probabilities rather than binary imputed values. To preserve the binary nature of the completed variable, we can instead use stochastic logistic regression imputation. For each nonrespondent, we independently generate

$$y_k^* \sim \text{Bernoulli}(\hat{f}_k), \qquad k \in S_m.$$

The resulting estimator is

$$\hat{\mu}_{\text{LGRI}}^{\text{stoch}} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k y_k^*\right\}.$$

Conditionally on the observed data and on the fitted logistic model, $\mathbb{E}_I\left(\hat{\mu}_{\text{LGRI}}^{\text{stoch}}\right) = \hat{\mu}_{\text{LGRI}}$. Thus, stochastic logistic regression imputation is centered, under the imputation mechanism, on deterministic logistic regression imputation.

Proposition 4.4. *Suppose that the assumptions of Proposition 4.3 hold and that the sampling weights satisfy* $N^{-2}\sum_{k\in S_m} w_k^2 = O_p(n^{-1})$. *Then*

$$\hat{\mu}_{\text{LGRI}}^{\text{stoch}} - \mu = O_p(n^{-1/2}).$$

Thus, stochastic logistic regression imputation is also root-n consistent.

Proof. We write

$$\hat{\mu}_{\text{LGRI}}^{\text{stoch}} - \mu = (\hat{\mu}_{\text{LGRI}} - \mu) + D_I, \tag{4.1.3}$$

where

$$D_I = \frac{1}{N}\sum_{k\in S_m} w_k\left(y_k^* - \hat{f}_k\right).$$

The first term in (4.1.3) is $O_p(n^{-1/2})$ by Proposition 4.3. Conditionally on the observed data,

$$\mathbb{E}_I(D_I) = 0.$$

Because the Bernoulli draws are conditionally independent,

$$\mathbb{V}_I(D_I) = \frac{1}{N^2} \sum_{k \in S_m} w_k^2 \hat{f}_k \left(1 - \hat{f}_k\right).$$

Since $\hat{f}_k(1 - \hat{f}_k) \leq 1/4$,

$$\mathbb{V}_I(D_I) = \frac{1}{N^2} \sum_{k \in S_m} w_k^2 \, \hat{f}_k(1 - \hat{f}_k) \leq \frac{1}{4N^2} \sum_{k \in S_m} w_k^2 = O_p(n^{-1}).$$

It follows that

$$D_I = O_p(n^{-1/2}).$$

Combining the two terms in (4.1.3) gives

$$\hat{\mu}_{\text{LGRI}}^{\text{stoch}} - \mu = O_p(n^{-1/2}).$$

□

We now briefly discuss variance estimation. The variance of the deterministic logistic regression imputed estimator can be obtained using either the two-phase framework or the reverse approach presented in Chapter 2. Unlike the Oracle case, the variability caused by estimating $\boldsymbol{\beta}$ must be taken into account. Because the logistic function is smooth, this additional variability can be handled by a first-order Taylor linearization. More precisely, the estimator $\hat{\mu}_{\text{LGRI}}$ and the maximum likelihood estimating equation can be expanded around the true parameter $\boldsymbol{\beta}$. This gives a linearized variable to which the usual survey variance formulas can be applied. The resulting variance estimator may be constructed under either the two-phase framework or the reverse approach. The detailed formulas are not presented here; the procedure is the logistic-regression analogue of the linearization used for linear regression imputation in Chapter 2.

Let $\widehat{\mathbb{V}}_{\text{det}}\left(\hat{\mu}_{\text{LGRI}}\right)$ denote the resulting variance estimator for deterministic logistic regression imputation. It may be obtained from either of the two approaches described in Chapter 2.

For stochastic logistic regression imputation, the law of total variance gives

$$\mathbb{V}_{\text{tot}}\left(\hat{\mu}^{\text{stoch}}_{\text{LGRI}}\right) = \mathbb{V}_{\text{tot}}\left(\hat{\mu}_{\text{LGRI}}\right) + \mathbb{V}_{\text{imp}}\left(\hat{\mu}^{\text{stoch}}_{\text{LGRI}}\right), \tag{4.1.4}$$

where

$$\mathbb{V}_{\text{imp}}\left(\hat{\mu}^{\text{stoch}}_{\text{LGRI}}\right) = \mathbb{E}_m\mathbb{E}_p\mathbb{E}_q\left[\mathbb{V}_I\left(\hat{\mu}^{\text{stoch}}_{\text{LGRI}}\right)\right].$$

The first component in (4.1.4) is the variance that would be present under deterministic logistic regression imputation. The second component is the additional variance caused by generating random zeros and ones.

An estimator of the imputation variance is

$$\widehat{\mathbb{V}}_{\text{imp}}\left(\hat{\mu}^{\text{stoch}}_{\text{LGRI}}\right) = \frac{1}{N^2}\sum_{k\in S_m} w_k^2\hat{f}_k\left(1-\hat{f}_k\right).$$

A variance estimator for stochastic logistic regression imputation is therefore

$$\widehat{\mathbb{V}}_{\text{tot}}\left(\hat{\mu}^{\text{stoch}}_{\text{LGRI}}\right) = \widehat{\mathbb{V}}_{\text{det}}\left(\hat{\mu}_{\text{LGRI}}\right) + \widehat{\mathbb{V}}_{\text{imp}}\left(\hat{\mu}^{\text{stoch}}_{\text{LGRI}}\right).$$

Thus, we begin with the variance estimator for deterministic logistic regression imputation, obtained using the two-phase or reverse approach, and then add the variance introduced by the Bernoulli draws.

The properties of both procedures depend on the correct specification of the logistic outcome model. If the logistic model is misspecified, the deterministic and stochastic imputed estimators may have a non-negligible asymptotic bias. Flexible machine learning methods can reduce the need to specify a particular parametric form, but a simple plug-in imputed estimator based on machine learning predictions does not automatically have the desired root-n rate or a simple variance estimator, as discussed in Chapter 3.

Chapter 3 introduced the cross-fitted AIPW estimator to address these inferential difficulties. In the next section, we use this estimator as a statistical benchmark and study how stochastic binary imputation can be combined with it. The objective is to produce imputed values in $\{0, 1\}$ while retaining the statistical properties of the cross-fitted AIPW estimator.

4.2 Direct stochastic versions of the cross-fitted AIPW estimator

In Chapter 3, we showed that combining the AIPW estimator with cross-fitting allows us to use machine learning methods while keeping good statistical properties. Under suitable conditions, the cross-fitted AIPW estimator is root-n consistent and asymptotically normal, and its variance can be estimated. We can therefore use it to estimate the finite population mean and to construct confidence intervals. However, when the variable of interest is binary, the cross-fitted AIPW estimator does not directly solve the problem of producing a completed data file. There are two practical requirements that we would like to satisfy. First, the imputed values must belong to the support of the survey variable; for a binary variable, they must be equal to 0 or 1. Second, the usual imputed estimator, that is, the weighted mean of the observed and imputed values, must be equal, or sufficiently close, to the cross-fitted AIPW estimator.

The cross-fitted AIPW estimator does not immediately satisfy these two requirements. To see this, let $j(k)$ denote the fold containing unit k and define $\hat{f}_k = \hat{f}^{(-j(k))}(\boldsymbol{v}_k)$ and $\hat{p}_k = \hat{p}^{(-j(k))}(\boldsymbol{v}_k)$. The cross-fitted AIPW estimator can be written as

$$\hat{\mu}^{\mathrm{cf}}_{\mathrm{aipw}} = \frac{1}{N}\sum_{k\in S} w_k \left[\hat{f}_k + \frac{r_k}{\hat{p}_k}\left(y_k - \hat{f}_k\right)\right]. \tag{4.2.1}$$

As shown in Chapter 3, under the product-rate condition and the other regularity conditions introduced there,

$$\hat{\mu}^{\mathrm{cf}}_{\mathrm{aipw}} - \hat{\mu}_{\mathrm{aipw}}(f,p) = o_p(n^{-1/2}),$$

where $\hat{\mu}_{\mathrm{aipw}}(f,p)$ denotes the Oracle AIPW estimator. Thus, the cross-fitted AIPW estimator has the same first-order properties as the Oracle estimator.

The difficulty is that the quantity

$$\hat{f}_k + \frac{r_k}{\hat{p}_k}\left(y_k - \hat{f}_k\right)$$

is not a binary value. It may take any real value and may even fall outside the interval $[0,1]$. Therefore, these unit-level quantities cannot be used directly as the completed

values of a binary survey variable. Moreover, (4.2.1) is not written as the usual weighted mean of the observed and imputed values appearing in (4.0.1).

A natural first idea is to replace the estimated probabilities $\hat{f}_k$ by random zeros and ones. This would preserve the binary form of the data. In the remainder of this section, we study two direct ways of introducing Bernoulli draws into the cross-fitted AIPW estimator. These two constructions help us understand the additional variability caused by stochastic imputation. However, as we will see, neither construction produces the convenient estimator in (4.0.1). This limitation will motivate the rejective stochastic imputation procedure introduced in the next section.

To study these two constructions, generate, conditionally independently,

$$Z_k \sim \text{Bernoulli}(\hat{f}_k), \qquad k \in S.$$

The variables Z_k are auxiliary Bernoulli draws used only to define the two stochastic AIPW estimators below. They are generated for every sampled unit because the outcome regression term appears for every unit in the AIPW formula.

A first stochastic version is obtained by keeping the first term of the AIPW estimator unchanged and replacing $\hat{f}_k$ by Z_k only in the residual term. This gives

$$\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,1}} = \frac{1}{N} \sum_{k \in S} w_k \left[\hat{f}_k + \frac{r_k}{\hat{p}_k} (y_k - Z_k) \right].$$

Using (4.2.1), this estimator can also be written as

$$\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,1}} = \hat{\mu}_{\text{aipw}}^{\text{cf}} + \frac{1}{N} \sum_{k \in S} w_k \frac{r_k}{\hat{p}_k} \left(\hat{f}_k - Z_k \right). \tag{4.2.2}$$

Because $r_k = 0$ for a nonrespondent, only the Bernoulli draws associated with respondents affect this first stochastic estimator.

A second stochastic version is obtained by replacing $\hat{f}_k$ by Z_k everywhere in the AIPW

formula:

$$\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,2}} = \frac{1}{N}\sum_{k\in S} w_k \left[Z_k + \frac{r_k}{\hat{p}_k}\left(y_k - Z_k\right)\right].$$

This estimator can be written as

$$\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,2}} = \hat{\mu}_{\text{aipw}}^{\text{cf}} + \frac{1}{N}\sum_{k\in S} w_k \left(1 - \frac{r_k}{\hat{p}_k}\right)\left(Z_k - \hat{f}_k\right). \tag{4.2.3}$$

Unlike the first version, the second version contains Bernoulli variation from both respondents and nonrespondents.

Both stochastic estimators are centered on the cross-fitted AIPW estimator. Indeed,

$$\mathbb{E}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,1}}\right) = \hat{\mu}_{\text{aipw}}^{\text{cf}},$$
$$\mathbb{E}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,2}}\right) = \hat{\mu}_{\text{aipw}}^{\text{cf}}.$$

Thus, averaging over repeated Bernoulli draws would recover the original cross-fitted AIPW estimator.

Using the law of total variance, for $h \in \{1, 2\}$,

$$\begin{aligned}\mathbb{V}_{mpqI}\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch},h}\right) &= \mathbb{V}_{mpq}\left[\mathbb{E}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch},h}\right)\right] + \mathbb{E}_{mpq}\left[\mathbb{V}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch},h}\right)\right] \\ &= \mathbb{V}_{mpq}\left(\hat{\mu}_{\text{aipw}}^{\text{cf}}\right) + \mathbb{E}_{mpq}\left[\mathbb{V}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch},h}\right)\right].\end{aligned}$$

The first component is the variance of the cross-fitted AIPW estimator studied in Chapter 3. The second component is the additional variance caused by the Bernoulli draws.

From (4.2.2), the conditional imputation variance of the first estimator is

$$\mathbb{V}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,1}}\right) = \frac{1}{N^2}\sum_{k\in S} w_k^2 \left(\frac{r_k}{\hat{p}_k}\right)^2 \hat{f}_k\left(1 - \hat{f}_k\right). \tag{4.2.4}$$

Similarly, from (4.2.3),

$$\mathbb{V}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,2}}\right) = \frac{1}{N^2}\sum_{k\in S} w_k^2\left(1-\frac{r_k}{\hat{p}_k}\right)^2 \hat{f}_k\left(1-\hat{f}_k\right). \tag{4.2.5}$$

Under mild regularity conditions, both conditional imputation variances are typically of order n^{-1}. Therefore, the Bernoulli randomization does not change the root-n rate of the estimator, but it adds variability of the same order as the usual sampling error.

We can also compare the two imputation variances directly. Subtracting (4.2.4) from (4.2.5) gives

$$\mathbb{V}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,2}}\right) - \mathbb{V}_I\left(\hat{\mu}_{\text{aipw}}^{\text{cf,stoch,1}}\right) = \frac{1}{N^2}\sum_{k\in S} w_k^2\left(1-\frac{2r_k}{\hat{p}_k}\right)\hat{f}_k\left(1-\hat{f}_k\right). \tag{4.2.6}$$

For a nonrespondent, $r_k = 0$, and the contribution to (4.2.6) is positive. This is because the first stochastic estimator does not use Z_k for nonrespondents, whereas the second estimator does. For a respondent, $r_k = 1$, and

$$1-\frac{2}{\hat{p}_k} < 0.$$

The respondent contribution is therefore nonpositive. Therefore, there is no general ordering between the two variances. The comparison depends on the sampling weights, the number of respondents and nonrespondents, the fitted outcome probabilities, and the fitted response probabilities. Very small values of $\hat{p}_k$ may also make either stochastic estimator unstable because of the inverse response-probability terms. Although these two constructions are centered on the cross-fitted AIPW estimator, neither one fully solves the practical problem considered in this chapter. The first stochastic version still contains the fractional values $\hat{f}_k$ in its first term. It therefore does not produce a completed binary variable.

Our goal is to choose binary imputed values for the nonrespondents so that $\hat{\mu}_I$ in (4.0.1) is equal, when possible, or sufficiently close to $\hat{\mu}_{\text{aipw}}^{\text{cf}}$. Exact equality may not always be possible because the set of weighted means that can be obtained from binary imputed

values is discrete. However, it is sufficient for first-order inference that

$$\hat{\mu}_I - \hat{\mu}_{\text{aipw}}^{\text{cf}} = o_p(n^{-1/2}). \tag{4.2.7}$$

The difference in (4.2.7) is then smaller than the usual root-n estimation error, so the completed-data estimator has the same first-order statistical properties as the cross-fitted AIPW estimator.

This leads to the main idea of the next section. We will generate binary imputed values for the nonrespondents and accept them only when the weighted mean of the completed data is sufficiently close to the cross-fitted AIPW estimate. This procedure is referred to as rejective stochastic imputation.

4.3 Rejective stochastic imputation

The purpose of rejective stochastic imputation is to produce binary imputed values while keeping the estimator computed from the completed data close to the cross-fitted AIPW estimator. More precisely, we want the completed data file to contain only values in $\{0, 1\}$, and we want a data user to be able to estimate the population mean by applying the usual weighted estimator directly to the completed variable.

The idea is related to balanced random imputation, where imputed values are selected so that suitable balancing conditions are satisfied. These conditions are used to reduce or eliminate the additional variability caused by random imputation; see Chauvet, Deville, and Haziza (2011) and Chauvet and Do Paco (2018). The procedure considered here is simpler: we generate Bernoulli values repeatedly and retain only the vectors for which the completed-data estimator is sufficiently close to the cross-fitted AIPW estimate.

4.3.1 The AIPW target for the nonrespondents

Let

$$T_r = \sum_{k \in S_r} w_k y_k$$

denote the weighted total of the observed values. For every nonrespondent $k \in S_m$, let $y_k^* \in \{0,1\}$ denote the imputed value. The estimator computed from the completed data is

$$\hat{\mu}_I = \frac{1}{N}\left\{T_r + \sum_{k \in S_m} w_k y_k^*\right\}.$$

To reproduce the cross-fitted AIPW estimator exactly, the imputed values would have to satisfy

$$\sum_{k \in S_m} w_k y_k^* = \tau_{\text{aipw}},$$

where

$$\tau_{\text{aipw}} = N\hat{\mu}_{\text{aipw}}^{\text{cf}} - T_r.$$

Thus, τ_{aipw} is the weighted total that must be assigned to the nonrespondents for the completed-data estimator to be exactly equal to the cross-fitted AIPW estimator.

The target can also be written as

$$\tau_{\text{aipw}} = \sum_{k \in S_m} w_k \hat{f}_k + \sum_{k \in S_r} w_k \left(\frac{1}{\hat{p}_k} - 1\right)\left(y_k - \hat{f}_k\right).$$

4.3.2 Can the AIPW target always be reached exactly?

The imputed values can only be zero or one. Therefore, their weighted total cannot take any arbitrary value. It can only take values obtained by adding the weights of some subset of the nonrespondents.

Let

$$\mathcal{A}_m = \left\{\sum_{k \in S_m} w_k z_k : z_k \in \{0,1\},\ k \in S_m\right\}$$

denote the set of all weighted totals that can be obtained from binary imputations. Exact equality with the AIPW estimator is possible if and only if $\tau_{\text{aipw}} \in \mathcal{A}_m$. In general, this condition does not have to be satisfied. This issue can be illustrated easily when all nonrespondents have the same weight w. Let $n_m = |S_m|$ denote the number of nonrespondents.

If exactly Q nonrespondents receive an imputed value equal to one, then

$$\sum_{k \in S_m} w y_k^* = wQ,$$

where $Q \in \{0, 1, \ldots, n_m\}$. Thus, the only possible totals are

$$0, w, 2w, \ldots, n_m w.$$

For example, suppose that there are four nonrespondents and that they all have weight 10. The possible totals are

$$0,\ 10,\ 20,\ 30,\ 40.$$

A target equal to 20 can be matched exactly by imputing one to two nonrespondents. A target equal to 23 cannot be matched exactly because no binary completion has a weighted total equal to 23.

With equal weights, exact reproduction is possible only if τ_{aipw}/w is an integer between zero and n_m. However, whenever $0 \le \tau_{\text{aipw}} \le n_m w$, there is an attainable total within $w/2$ of the target. Indeed, let

$$Q_n = \left\lfloor \frac{\tau_{\text{aipw}}}{w} + \frac{1}{2} \right\rfloor.$$

Then

$$|wQ_n - \tau_{\text{aipw}}| \le \frac{w}{2}.$$

Under simple random sampling without replacement, $w = N/n$. Therefore, the corresponding difference on the scale of the population mean is at most

$$\frac{w}{2N} = \frac{1}{2n}.$$

Since

$$\frac{1}{n} = o(n^{-1/2}),$$

this difference is negligible compared with the usual root-n estimation error. This equal-weight example explains why exact matching may fail, but also why a small tolerance is sufficient under SRSWOR. With unequal weights, the same general issue remains.

4.3.3 The rejective imputation procedure

Before defining the acceptance rule, we first introduce the smallest distance that can be achieved between a completed-data estimator and the cross-fitted AIPW estimator. Define

$$d_n = \min_{\boldsymbol{z}\in\{0,1\}^{|S_m|}} \left| \sum_{k\in S_m} w_k z_k - \tau_{\text{aipw}} \right|.$$

The quantity d_n is measured on the scale of the total. It is the smallest possible difference between the AIPW target for the nonrespondents and a weighted total that can actually be obtained by assigning zeros and ones to the nonrespondents. Thus, $d_n = 0$ when the AIPW target can be matched exactly, whereas $d_n > 0$ when exact matching is impossible. The corresponding smallest distance on the scale of the population mean is d_n/N.

Let $e_n \geq 0$ denote a tolerance on the scale of the population mean. The acceptance set is

$$\mathcal{B}_n(e_n) = \left\{ \boldsymbol{z} \in \{0,1\}^{|S_m|} : \left| \frac{1}{N} \left\{ T_r + \sum_{k\in S_m} w_k z_k \right\} - \hat{\mu}^{\text{cf}}_{\text{aipw}} \right| \leq e_n \right\}.$$

Using the definition

$$\tau_{\text{aipw}} = N\hat{\mu}^{\text{cf}}_{\text{aipw}} - T_r,$$

the acceptance condition can equivalently be written as

$$\left| \sum_{k\in S_m} w_k z_k - \tau_{\text{aipw}} \right| \leq N e_n.$$

Therefore, the acceptance set is nonempty if and only if

$$\frac{d_n}{N} \leq e_n.$$

Indeed, if $e_n < d_n/N$, no binary imputation vector can be sufficiently close to the AIPW target. If $e_n \geq d_n/N$, at least one binary vector satisfies the acceptance rule.

The rejective stochastic imputation procedure is carried out as follows:

1. Compute the cross-fitted predictions $\hat{f}_k$ and $\hat{p}_k$ and the cross-fitted AIPW estimator $\hat{\mu}^{\text{cf}}_{\text{aipw}}$.

2. For every nonrespondent, independently generate

$$Z_k \sim \text{Bernoulli}(\hat{f}_k), \qquad k \in S_m.$$

3. Compute the estimator based on the proposed imputed values:

$$\hat{\mu}_I(\boldsymbol{Z}) = \frac{1}{N}\left\{T_r + \sum_{k\in S_m} w_k Z_k\right\}.$$

4. Accept the vector $\boldsymbol{Z} = (Z_k : k \in S_m)$ if

$$\left|\hat{\mu}_I(\boldsymbol{Z}) - \hat{\mu}^{\text{cf}}_{\text{aipw}}\right| \leq e_n.$$

   Otherwise, generate a new vector and repeat the previous steps.

5. For the accepted vector, define

$$y^*_k = Z_k, \qquad k \in S_m.$$

The resulting rejective stochastic imputed estimator is

$$\hat{\mu}^{\text{rej}}_{\text{imp}} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_k + \sum_{k\in S_m} w_k y^*_k\right\}.$$

Proposition 4.5. *For every accepted imputation vector,*

$$\left|\hat{\mu}^{\text{rej}}_{\text{imp}} - \hat{\mu}^{\text{cf}}_{\text{aipw}}\right| \leq e_n. \tag{4.3.1}$$

Moreover, conditionally on the observed sample and the fitted nuisance functions,

$$\mathbb{V}_I\left(\hat{\mu}^{\text{rej}}_{\text{imp}}\right) \leq e_n^2. \tag{4.3.2}$$

If

$$e_n = o(n^{-1/2}),$$

then

$$\hat{\mu}^{\text{rej}}_{\text{imp}} - \hat{\mu}^{\text{cf}}_{\text{aipw}} = o_p(n^{-1/2}).$$

Consequently, the rejective imputed estimator has the same first-order asymptotic properties as the cross-fitted AIPW estimator.

Proof. The inequality in (4.3.1) follows directly from the acceptance rule. Indeed, a proposed imputation vector is accepted only if

$$\left|\hat{\mu}^{\text{rej}}_{\text{imp}} - \hat{\mu}^{\text{cf}}_{\text{aipw}}\right| \leq e_n.$$

We now establish the variance bound. Conditionally on the observed sample and the fitted nuisance functions, the cross-fitted AIPW estimator $\hat{\mu}^{\text{cf}}_{\text{aipw}}$ is fixed. The only remaining randomness comes from the imputation mechanism.

For every accepted imputation vector,

$$\left|\hat{\mu}^{\text{rej}}_{\text{imp}} - \hat{\mu}^{\text{cf}}_{\text{aipw}}\right| \leq e_n.$$

Therefore,

$$\left(\hat{\mu}^{\text{rej}}_{\text{imp}} - \hat{\mu}^{\text{cf}}_{\text{aipw}}\right)^2 \leq e_n^2.$$

Taking expectation with respect to the imputation mechanism gives

$$\mathbb{E}_I\left[\left(\hat{\mu}^{\text{rej}}_{\text{imp}} - \hat{\mu}^{\text{cf}}_{\text{aipw}}\right)^2\right] \leq e_n^2.$$

For any random variable X and any fixed constant a, we have

$$\begin{aligned}\mathbb{E}\left[(X-a)^2\right] &= \mathbb{V}(X) + \{\mathbb{E}(X) - a\}^2 \\ &\geq \mathbb{V}(X).\end{aligned}$$

Applying this identity with $X = \hat{\mu}_{\text{imp}}^{\text{rej}}$ and $a = \hat{\mu}_{\text{aipw}}^{\text{cf}}$, we obtain

$$\begin{aligned}\mathbb{V}_I\left(\hat{\mu}_{\text{imp}}^{\text{rej}}\right) &\leq \mathbb{E}_I\left[\left(\hat{\mu}_{\text{imp}}^{\text{rej}} - \hat{\mu}_{\text{aipw}}^{\text{cf}}\right)^2\right] \\ &\leq e_n^2.\end{aligned}$$

This proves (4.3.2). Finally, if

$$e_n = o(n^{-1/2}),$$

then the acceptance rule implies

$$\left|\hat{\mu}_{\text{imp}}^{\text{rej}} - \hat{\mu}_{\text{aipw}}^{\text{cf}}\right| = o(n^{-1/2}).$$

Hence,

$$\hat{\mu}_{\text{imp}}^{\text{rej}} - \hat{\mu}_{\text{aipw}}^{\text{cf}} = o_p(n^{-1/2}).$$

The two estimators are therefore asymptotically equivalent at the root-n scale. The final conclusion follows from Slutsky's theorem.

□

The previous proposition also shows that the additional variability caused by rejective imputation is very small. If $e_n = 0$ and the acceptance set is nonempty, every accepted imputation vector satisfies

$$\hat{\mu}_{\text{imp}}^{\text{rej}} = \hat{\mu}_{\text{aipw}}^{\text{cf}}.$$

As a result,

$$\mathbb{V}_I\left(\hat{\mu}_{\text{imp}}^{\text{rej}}\right) = 0.$$

Different accepted completed files may still assign the value one to different nonrespondents, but they all produce exactly the same estimate of the population mean. Thus, exact matching eliminates the imputation variance of the point estimator.

When $e_n > 0$, the completed-data estimator may vary from one accepted imputation vector to another. However, its conditional imputation variance is bounded by e_n^2. If

$$e_n = o(n^{-1/2}),$$

then

$$e_n^2 = o(n^{-1}).$$

The first-order variance of the cross-fitted AIPW estimator is typically of order n^{-1}. Therefore, the additional imputation variance produced by the rejective procedure is of smaller order and is asymptotically negligible.

4.3.4 Choice of the tolerance

The tolerance must satisfy two different requirements. First, it must be large enough for the acceptance set to be nonempty: $e_n \geq d_n/N$. Second, to preserve the first-order properties of the cross-fitted AIPW estimator, it must satisfy $e_n = o(n^{-1/2})$. Therefore, the feasibility gap should satisfy

$$\frac{d_n}{N} = o_p(n^{-1/2})$$

if we want to choose a tolerance that is both feasible and asymptotically negligible.

In the simulation study, the sample is selected by simple random sampling without replacement and the tolerance is $e_n = 1/n$. As shown in the equal-weight example above, when the AIPW target lies in the feasible interval,

$$\frac{d_n}{N} \leq \frac{1}{2n}.$$

Therefore, $e_n = 1/n$ is large enough to include at least one attainable completed-data mean. At the same time,

$$\frac{1}{n} = o(n^{-1/2}),$$

so this tolerance is small enough to preserve the first-order properties of the cross-fitted AIPW estimator.

4.3.5 Extension to a categorical variable

We now extend rejective stochastic imputation to a categorical variable with more than two categories. Let $y_k \in \{1, \ldots, L\}$, where $L > 2$ is the number of possible categories. For

example, a variable describing employment status could contain the categories employed, unemployed, retired, and student. To estimate the proportion of units in each category, we represent the categorical variable using L binary indicators. For category ℓ, define

$$y_{k\ell} = 1(y_k = \ell), \qquad \ell = 1, \ldots, L.$$

Thus, $y_{k\ell} = 1$ if unit k belongs to category ℓ, and $y_{k\ell} = 0$ otherwise. Since every unit belongs to exactly one category,

$$\sum_{\ell=1}^{L} y_{k\ell} = 1.$$

The finite population proportion in category ℓ is

$$\mu_\ell = \frac{1}{N} \sum_{k \in U} y_{k\ell}, \qquad \ell = 1, \ldots, L.$$

The parameter of interest is therefore the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^\top$. The elements of this vector describe the distribution of the categorical variable in the finite population.

For sampled unit k, let $j(k)$ denote the cross-fitting fold containing that unit. We define $\hat{f}_{k\ell} = \hat{f}_\ell^{(-j(k))}(\boldsymbol{v}_k)$ as the cross-fitted estimate of $P(y_k = \ell \mid \boldsymbol{v}_k)$. For every sampled unit, the estimated class probabilities satisfy

$$\sum_{\ell=1}^{L} \hat{f}_{k\ell} = 1.$$

We also define $\hat{p}_k = \hat{p}^{(-j(k))}(\boldsymbol{v}_k)$ as the cross-fitted estimate of the response probability.

For category ℓ, the cross-fitted AIPW estimator is

$$\hat{\mu}_{\text{aipw},\ell}^{\text{cf}} = \frac{1}{N} \sum_{k \in S} w_k \left[\hat{f}_{k\ell} + \frac{r_k}{\hat{p}_k} \left(y_{k\ell} - \hat{f}_{k\ell} \right) \right].$$

Thus, we obtain one AIPW estimator for each category. Under the conditions introduced in Chapter 3, each estimator has the usual root-n properties.

For category ℓ, let

$$T_{r\ell} = \sum_{k \in S_r} w_k y_{k\ell}$$

denote the weighted total observed among the respondents. The weighted total that should be assigned to the nonrespondents in order to reproduce the AIPW estimator is

$$\tau_{\text{aipw},\ell} = N \hat{\mu}^{\text{cf}}_{\text{aipw},\ell} - T_{r\ell}.$$

Thus, $\tau_{\text{aipw},\ell}$ is the target total for category ℓ among the nonrespondents.

The L target totals are linked to one another. Since the class indicators and the estimated class probabilities both sum to one, we have

$$\sum_{\ell=1}^{L} \tau_{\text{aipw},\ell} = \sum_{k \in S_m} w_k.$$

The right-hand side is the total survey weight of the nonrespondents. Therefore, only $L - 1$ of the class-total constraints are independent. Once the totals for $L-1$ categories have been fixed, the total for the last category is determined automatically.

For each nonrespondent, we must generate one and only one imputed category. Define

$$\boldsymbol{Z}_k = (Z_{k1}, \ldots, Z_{kL})^\top,$$

where $Z_{k\ell} \in \{0, 1\}, \qquad \ell = 1, \ldots, L$, and

$$\sum_{\ell=1}^{L} Z_{k\ell} = 1. \tag{4.3.3}$$

If $Z_{k\ell} = 1$, unit k is assigned to category ℓ. The condition in (4.3.3) prevents a nonrespondent from being assigned to several categories or to no category.

For each nonrespondent, we independently generate

$$\boldsymbol{Z}_k \sim \text{Multinomial}\left(1; \hat{f}_{k1}, \ldots, \hat{f}_{kL}\right), \qquad k \in S_m.$$

This is equivalent to drawing one categorical value with probabilities $\hat{f}_{k1}, \ldots, \hat{f}_{kL}$.

For a proposed collection of multinomial draws, the completed-data estimator of the proportion in category ℓ is

$$\hat{\mu}_{I,\ell}(\boldsymbol{Z}) = \frac{1}{N}\left\{T_{r\ell} + \sum_{k \in S_m} w_k Z_{k\ell}\right\}.$$

The corresponding vector of estimates is

$$\hat{\boldsymbol{\mu}}_I(\boldsymbol{Z}) = \left(\hat{\mu}_{I,1}(\boldsymbol{Z}), \ldots, \hat{\mu}_{I,L}(\boldsymbol{Z})\right)^\top.$$

We want this vector to be close to the vector of cross-fitted AIPW estimates

$$\hat{\boldsymbol{\mu}}^{\text{cf}}_{\text{aipw}} = \left(\hat{\mu}^{\text{cf}}_{\text{aipw},1}, \ldots, \hat{\mu}^{\text{cf}}_{\text{aipw},L}\right)^\top.$$

Let $e_n \geq 0$ denote a tolerance on the scale of the population proportions. We accept a proposed collection of multinomial imputations if

$$\max_{1 \leq \ell \leq L} \left|\hat{\mu}_{I,\ell}(\boldsymbol{Z}) - \hat{\mu}^{\text{cf}}_{\text{aipw},\ell}\right| \leq e_n.$$

Thus, the same collection of imputed categories must satisfy the acceptance rule for every category at the same time. We do not accept or reject the categories separately.
Let $\mathcal{Z}_m$ denote the set of all possible category assignments for the nonrespondents:

$$\mathcal{Z}_m = \left\{\boldsymbol{z} : z_{k\ell} \in \{0,1\}, \quad \sum_{\ell=1}^{L} z_{k\ell} = 1 \text{ for every } k \in S_m\right\}.$$

For an assignment $\boldsymbol{z} \in \mathcal{Z}_m$, define

$$\hat{\mu}_{I,\ell}(\boldsymbol{z}) = \frac{1}{N}\left\{T_{r\ell} + \sum_{k \in S_m} w_k z_{k\ell}\right\}.$$

We define

$$d_{n,L} = \min_{\boldsymbol{z} \in \mathcal{Z}_m} \max_{1 \leq \ell \leq L} \left|\hat{\mu}_{I,\ell}(\boldsymbol{z}) - \hat{\mu}^{\text{cf}}_{\text{aipw},\ell}\right|.$$

The quantity $d_{n,L}$ is the smallest possible maximum distance between the vector of completed-data proportions and the vector of AIPW proportions. It is the multiclass version of the feasibility gap introduced in the binary case.

The acceptance set is nonempty if and only if

$$d_{n,L} \leq e_n.$$

If $e_n < d_{n,L}$, no assignment of the nonrespondents to the L categories can satisfy the acceptance rule. If $e_n \geq d_{n,L}$, at least one possible assignment is sufficiently close to all the AIPW targets.

The rejective stochastic imputation procedure is carried out as follows:

1. Compute the cross-fitted class probabilities $\hat{f}_{k\ell}$, the cross-fitted response probabilities $\hat{p}_k$, and the L cross-fitted AIPW estimators

$$\hat{\mu}^{\text{cf}}_{\text{aipw},1}, \ldots, \hat{\mu}^{\text{cf}}_{\text{aipw},L}.$$

2. For every nonrespondent, independently generate

$$\boldsymbol{Z}_k \sim \text{Multinomial}\left(1; \hat{f}_{k1}, \ldots, \hat{f}_{kL}\right).$$

3. For each category ℓ, compute

$$\hat{\mu}_{I,\ell}(\boldsymbol{Z}) = \frac{1}{N}\left\{T_{r\ell} + \sum_{k \in S_m} w_k Z_{k\ell}\right\}.$$

4. Accept the complete collection $\{\boldsymbol{Z}_k : k \in S_m\}$ if

$$\max_{1 \leq \ell \leq L} \left|\hat{\mu}_{I,\ell}(\boldsymbol{Z}) - \hat{\mu}^{\text{cf}}_{\text{aipw},\ell}\right| \leq e_n.$$

   Otherwise, generate a new collection of multinomial draws for all nonrespondents and repeat the previous steps.

5. For the accepted collection, set

$$y^*_{k\ell} = Z_{k\ell}, \qquad k \in S_m, \qquad \ell = 1, \ldots, L.$$

The final imputed categorical value is

$$y^*_k = \sum_{\ell=1}^{L} \ell y^*_{k\ell}.$$

The resulting rejective stochastic imputed estimator of the proportion in category ℓ is

$$\hat{\mu}^{\text{rej}}_{\text{imp},\ell} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_{k\ell} + \sum_{k\in S_m} w_k y^*_{k\ell}\right\}, \qquad \ell = 1, \ldots, L.$$

Each estimator can be computed directly from the completed categorical variable and the survey weights.

Proposition 4.6. *Suppose that the number of categories L is fixed. For every accepted collection of imputations,*

$$\max_{1\le\ell\le L}\left|\hat{\mu}^{\text{rej}}_{\text{imp},\ell} - \hat{\mu}^{\text{cf}}_{\text{aipw},\ell}\right| \le e_n. \tag{4.3.4}$$

For every category ℓ, conditionally on the observed sample and the fitted nuisance functions,

$$\mathbb{V}_I\left(\hat{\mu}^{\text{rej}}_{\text{imp},\ell}\right) \le e_n^2.$$

If

$$e_n = o(n^{-1/2}),$$

then

$$\max_{1\le\ell\le L}\left|\hat{\mu}^{\text{rej}}_{\text{imp},\ell} - \hat{\mu}^{\text{cf}}_{\text{aipw},\ell}\right| = o_p(n^{-1/2}).$$

Consequently, the vector of rejective completed-data estimators has the same first-order asymptotic properties as the vector of cross-fitted AIPW estimators.

Proof. The bound in (4.3.4) follows directly from the acceptance rule.

For a given category ℓ, every accepted value of $\hat{\mu}^{\text{rej}}_{\text{imp},\ell}$ belongs to the interval

$$\left[\hat{\mu}^{\text{cf}}_{\text{aipw},\ell} - e_n,\, \hat{\mu}^{\text{cf}}_{\text{aipw},\ell} + e_n\right].$$

This interval has length $2e_n$. Therefore,

$$\mathbb{V}_I\left(\hat{\mu}^{\text{rej}}_{\text{imp},\ell}\right) \leq e_n^2.$$

Finally, if $e_n = o(n^{-1/2})$, the difference between each completed-data estimator and its AIPW target is negligible at the root-n scale. The result follows from Slutsky's theorem.

□

The proposition shows that the same transfer result obtained for a binary variable also holds for a categorical variable. We do not need to reproduce all the AIPW proportions exactly. It is sufficient for the largest difference across the L categories to be smaller than the usual root-n estimation error.

4.4 Simulation study

We conducted two Monte Carlo experiments to study the performance of rejective stochastic imputation. The first experiment considered a binary survey variable, while the second considered a categorical variable with four categories. In both experiments, we examined the bias and variability of the rejective estimator, the accuracy of its variance estimator, and the coverage of nominal 95% confidence intervals. The two experiments share the same overall simulation design; they differ only in how the survey variable and its conditional probabilities were generated. We first describe the design common to both experiments, and then detail each experiment separately.

4.4.1 Common simulation design

A finite population of size $N = 10{,}000$ was generated in each experiment. For each population unit $k \in U$, a vector of five auxiliary variables, $\boldsymbol{v}_k = (v_{k1}, \ldots, v_{k5})^\top$, was independently

generated from the normal distribution

$$v_{kj} \sim \mathcal{N}(50, 25^2), \qquad j = 1, \ldots, 5.$$

The auxiliary variables were generated once at the beginning of each experiment and remained unchanged over all Monte Carlo replications.

At each replication, a simple random sample without replacement of size $n = 1{,}000$ was selected, corresponding to a sampling fraction of $\frac{n}{N} = 0.10$.

Nonresponse was subsequently introduced by generating a response indicator for every sampled unit. The response probability followed the logistic model

$$p(\boldsymbol{v}_k) = \frac{\exp\left(\alpha_2 v_{k2} + \alpha_3 v_{k3} + \alpha_4 v_{k4}\right)}{1 + \exp\left(\alpha_2 v_{k2} + \alpha_3 v_{k3} + \alpha_4 v_{k4}\right)},$$

with

$$\alpha_2 = -0.03768859, \qquad \alpha_3 = 0.02005180, \qquad \alpha_4 = 0.02017648.$$

These coefficients were common to both experiments and yielded an expected nonresponse rate of approximately 50%. Since the response probability depended exclusively on the fully observed auxiliary variables, the resulting nonresponse mechanism was MAR in both experiments. When the response probabilities were estimated, the estimated values were truncated from below, $\hat{p}_k \leftarrow \max(\hat{p}_k, 0.01)$, so as to avoid extreme inverse-probability weights.

The response probabilities and the conditional outcome probabilities were estimated using the cross-fitting procedure described in Chapter 3. These estimates were subsequently used to compute the cross-fitted AIPW estimator of the quantity of interest. For the cross-fitted AIPW estimator, the sample was randomly partitioned at each replication into $J = 5$ folds of (almost) equal size, without stratification on the response indicator.

For every nonrespondent, an imputed value (or category, in the four-class experiment) was generated from the estimated conditional distribution. The complete set of imputations was retained whenever the imputed estimator fell within a tolerance e_n of the cross-fitted AIPW estimator, the exact form of this criterion being specific to each experiment. Otherwise, a new set of imputations was generated.

4.4.2 Simulation design for the binary variable

The variable of interest was binary, with $y_k \in \{0, 1\}$. Its conditional probability of success was determined by the auxiliary variables through

$$f(\boldsymbol{v}_k) = \frac{v_{k1}^2 + v_{k2}^2}{v_{k1}^2 + v_{k2}^2 + |v_{k4} v_{k5}|}.$$

The response variable was then generated independently for each unit according to

$$y_k^{(b)} \sim \text{Bernoulli}(f(\boldsymbol{v}_k)), \qquad k \in U,$$

at Monte Carlo replication (b). The corresponding finite population mean was

$$\mu^{(b)} = \frac{1}{N} \sum_{k \in U} y_k^{(b)}.$$

The average conditional probability of success was approximately 69.1%.

The cross-fitted AIPW estimator of the population mean is denoted $\hat{\mu}_{\text{aipw}}^{\text{cf}}$. The complete set of imputations was retained whenever

$$\left| \hat{\mu}_I - \hat{\mu}_{\text{aipw}}^{\text{cf}} \right| \leq e_n,$$

with

$$e_n = \frac{1}{n} = 10^{-3}.$$

The rejective imputed estimator of the population mean was finally computed as

$$\hat{\mu}_{\text{imp}}^{\text{rej}} = \frac{1}{N} \left(\sum_{k \in S_r} w_k y_k + \sum_{k \in S_m} w_k y_k^* \right),$$

where S_r and S_m denote the sets of respondents and nonrespondents, respectively.

The entire simulation experiment was repeated $B = 10{,}000$ times.

4.4.3 Simulation design for the four-category variable

The survey variable had four possible categories: $y_k \in \{1, 2, 3, 4\}$. For category ℓ, define $y_{k\ell} = 1(y_k = \ell), \qquad \ell = 1, \ldots, 4$. Thus, the vector

$$\boldsymbol{y}_k = (y_{k1}, y_{k2}, y_{k3}, y_{k4})^\top$$

contains exactly one element equal to one and three elements equal to zero.

For each unit, the conditional probabilities of the four categories were defined as functions of the auxiliary variables. Specifically,

$$f_1(\boldsymbol{v}_k) = \frac{|4v_{k1} + 5v_{k2} + 6v_{k3}|}{A(\boldsymbol{v}_k)},$$

$$f_2(\boldsymbol{v}_k) = \frac{v_{k2}^2}{A(\boldsymbol{v}_k)},$$

$$f_3(\boldsymbol{v}_k) = \frac{|v_{k3}v_{k4}|}{A(\boldsymbol{v}_k)},$$

and

$$f_4(\boldsymbol{v}_k) = \frac{\exp(v_{k5}/8)}{A(\boldsymbol{v}_k)},$$

where

$$A(\boldsymbol{v}_k) = |4v_{k1} + 5v_{k2} + 6v_{k3}| + v_{k2}^2 + |v_{k3}v_{k4}| + \exp(v_{k5}/8).$$

Thus, the four probabilities are nonnegative and sum to one. At Monte Carlo replication b, the class-indicator vector was generated from

$$\boldsymbol{y}_k^{(b)} \sim \text{Multinomial}\left(1; f_1(\boldsymbol{v}_k), f_2(\boldsymbol{v}_k), f_3(\boldsymbol{v}_k), f_4(\boldsymbol{v}_k)\right), \qquad k \in U.$$

The finite population proportion in category ℓ was then

$$\mu_\ell^{(b)} = \frac{1}{N}\sum_{k\in U} y_{k\ell}^{(b)}, \qquad \ell = 1, \ldots, 4.$$

The average conditional probabilities of the four categories, $\frac{1}{N}\sum_{k\in U} f_\ell(\boldsymbol{v}_k)$, were approximately 10.3%, 33.7%, 30.8%, and 25.2% for categories 1, 2, 3, and 4, respectively, so that the expected finite population proportions were approximately equal to these values.

For each category ℓ, the cross-fitted AIPW estimator is denoted $\hat{\mu}_{\text{aipw},\ell}^{\text{cf}}$. For every nonrespondent, a candidate category was generated from the cross-fitted multinomial probabilities. The complete collection of imputed categories was accepted when

$$\max_{1\le\ell\le 4} \left|\hat{\mu}_{I,\ell} - \hat{\mu}_{\text{aipw},\ell}^{\text{cf}}\right| \le e_n,$$

where

$$e_n = 10^{-3}.$$

If the condition was not satisfied, a new collection of multinomial imputations was generated.

For an accepted collection of imputations, the rejective estimator of the proportion in category ℓ was

$$\hat{\mu}_{\text{imp},\ell}^{\text{rej}} = \frac{1}{N}\left\{\sum_{k\in S_r} w_k y_{k\ell} + \sum_{k\in S_m} w_k y_{k\ell}^*\right\}, \qquad \ell = 1, \ldots, 4.$$

It is important to use the class indicators $y_{k\ell}$ in this expression. Taking the weighted mean of the numerical labels $1, 2, 3, 4$ would estimate the average class label rather than the population proportion in each category.

The Monte Carlo experiment was repeated $B = 5{,}000$ times.

4.4.4 Machine learning methods

For the outcome model, we considered random forest, K-nearest neighbours, and gradient boosting. For the response propensity, we considered random forest, gradient boosting, and classification trees. Not every possible pair of methods was retained in the final table.

Random forest was fitted using 500 trees, a minimum terminal-node size of 5, and `mtry` $= 2$ for both the outcome and response models. The K-nearest-neighbour outcome model used $K = 10$. For the outcome model, gradient boosting was fitted using 1,000 trees, an interaction depth of 1, and a shrinkage parameter equal to 0.001. For the response model, gradient boosting was fitted using 1,500 trees, an interaction depth of 1, and a shrinkage parameter equal to 0.01. The classification-tree response model used a maximum depth of 4, a minimum bucket size of 10, and a complexity parameter equal to 0.01.

For each point estimator, we computed its Monte Carlo percent relative bias and its relative efficiency with respect to the estimator that would have been computed under full response. Under the convention used in the tables, a relative efficiency equal to 100 indicates the same Monte Carlo variance as the full-response estimator, whereas a value greater than 100 indicates a loss of efficiency. For example, a value equal to 300 means that the Monte Carlo variance is three times as large as that of the full-response estimator.

For each variance estimator, we computed its Monte Carlo percent relative bias. We also constructed nominal 95% confidence intervals and recorded their empirical coverage rate. A satisfactory variance estimator should have a relative bias close to zero and a coverage rate close to 95%.

4.4.5 Simulation results

Table 4.1 presents the results for the binary experiment. The Monte Carlo relative bias of the rejective estimator is very small for all nine combinations of machine learning methods. It ranges approximately from -0.8% to 0.5%. Thus, the rejective estimator is

approximately unbiased for the population mean.

The relative-efficiency values in the binary experiment range from 240 to 277. Under the convention used in the table, this means that the variance of the rejective estimator is approximately 2.4 to 2.8 times the variance of the full-response estimator. This loss of efficiency reflects the effect of nonresponse and of the estimation of the nuisance functions. The additional loss of efficiency due to rejective imputation is negligible.

The variance estimator performs well in the binary experiment. Its relative bias ranges from approximately -6.5% to 3.4%. The largest underestimation occurs when random forest is used for both nuisance functions, while combinations using gradient boosting for the response propensity tend to produce small positive biases. Despite these differences, the coverage rates range from 93.7% to 95.4% and are generally close to the nominal 95% level.

Table 4.2 presents the results for the four-category experiment. Once again, the relative bias of the point estimator is small. All reported biases lie between approximately -1.7% and 1.2%. The rejective estimator is approximately unbiased for the population proportions in all four categories.

The main difference between the categories concerns efficiency. The relative-efficiency values range from 218 to 413. The largest values are observed for the combination of K-nearest neighbours for the outcome model and classification trees for the response model. This combination gives relative-efficiency values of 413, 400, 348, and 283 for categories 1, 2, 3, and 4, respectively. It is therefore consistently less efficient than the other combinations considered.

Category 1, which is the least common category, generally has a relatively large variance ratio. However, prevalence alone does not explain all the results. Category 2 also has large relative-efficiency values even though it is the most common category. The choice of the outcome and response learners therefore also plays an important role. The tables alone do not allow us to determine precisely why a particular combination is less efficient.

The variance estimators remain well calibrated in the four-category experiment. Their relative biases range from approximately -1.3% to 8.7%. Most of the biases are

positive, indicating a small tendency to overestimate the variance. Nevertheless, all empirical coverage rates lie between 94.2% and 96.0%, which is close to the nominal 95% level.

Overall, the simulation results support the main objective of rejective stochastic imputation. The completed files contain only admissible binary or categorical values, while the resulting estimators have very small Monte Carlo biases. The variance estimators also provide confidence intervals with coverage rates close to their nominal level.

Table 4.1: Performance of the rejective stochastic imputed estimator for the binary variable

ML methods (f, p)	RB (%)	RE	VRB (%)	CR (%)
RF, RF	−0.5	258	−6.5	93.7
RF, Boosting	0.5	277	3.1	95.1
RF, Trees	−0.8	267	−1.2	95.2
Boosting, RF	−0.5	240	−2.4	95.2
Boosting, Boosting	0.4	264	3.0	94.9
Boosting, Trees	−0.8	257	0.0	95.2
KNN, RF	−0.3	253	−4.8	94.4
KNN, Boosting	0.4	273	3.4	95.0
KNN, Trees	−0.6	265	−2.4	95.4

Note: RB denotes the Monte Carlo percent relative bias of the point estimator; RE denotes its relative efficiency with respect to the Horvitz–Thompson estimator; VRB denotes the Monte Carlo percent relative bias of the variance estimator; and CR denotes the empirical coverage rate of the nominal 95% confidence interval.

Table 4.2: Performance of the rejective stochastic imputed estimator for the four categories

Class	ML methods (f, p)	RB (%)	RE	VRB (%)	CR (%)
1	RF, RF	−0.3	359	6.2	94.9
	RF, Boosting	−0.6	302	3.8	94.8
	Boosting, Boosting	−1.7	307	5.3	94.8
	KNN, Trees	0.2	413	−0.1	94.4
	KNN, Boosting	−0.4	316	1.3	94.4
2	RF, RF	0.6	371	4.8	95.8
	RF, Boosting	0.5	307	2.7	95.3
	Boosting, Boosting	1.1	320	8.7	95.5
	KNN, Trees	−1.2	400	−1.3	94.2
	KNN, Boosting	0.6	314	2.3	95.2
3	RF, RF	−0.4	268	8.3	95.9
	RF, Boosting	−0.6	225	0.7	95.1
	Boosting, Boosting	−0.9	229	6.9	95.7
	KNN, Trees	0.2	348	1.0	96.0
	KNN, Boosting	−1.1	236	2.6	95.0
4	RF, RF	−0.1	249	6.3	95.0
	RF, Boosting	0.3	218	4.6	95.0
	Boosting, Boosting	0.4	218	6.9	95.6
	KNN, Trees	1.2	283	1.0	94.9
	KNN, Boosting	0.6	227	4.8	95.0

Note: RB denotes the Monte Carlo percent relative bias of the point estimator; RE denotes its relative efficiency with respect to the Horvitz–Thompson estimator; VRB denotes the Monte Carlo percent relative bias of the variance estimator; and CR denotes the empirical coverage rate of the nominal 95% confidence interval.

5

Concluding remarks and future research

In this thesis, we studied the use of machine learning methods for treating item nonresponse in surveys. The first two substantive chapters reviewed existing methods and results on imputation and inference in the presence of missing data, including the double/debiased machine learning framework introduced by Chernozhukov et al. (2018). The main objective of this framework is to use flexible machine learning methods while preserving valid statistical inference under suitable conditions. In particular, we are interested not only in obtaining accurate predictions for the missing values, but also in constructing estimators that are root-n consistent, asymptotically normal, and accompanied by valid variance estimators and confidence intervals.

The usual imputed estimator is particularly attractive in survey practice. Once a completed data file has been produced, the estimator of a population mean can be written as the weighted mean of the observed and imputed values. This form is simple for data users because they can compute point estimates without refitting the model used to produce the imputations. However, when the imputation function is estimated using a flexible nonparametric or machine learning method, valid inference becomes more difficult. The

usual imputed estimator is not Neyman orthogonal with respect to the outcome regression function. In simple terms, this means that an estimation error in the machine learning model can have a first-order effect on the final estimator. As a result, consistency of the prediction model alone does not guarantee root-n consistency of the imputed estimator. Establishing a Central Limit Theorem and estimating the variance of the imputed estimator may then require a separate theoretical analysis for each machine learning method.

To address this problem, we considered the cross-fitted Augmented Inverse Probability Weighting (AIPW) estimator. This estimator combines an estimate of the conditional mean of the survey variable with an estimate of the response probability. Its score is Neyman orthogonal, and cross-fitting prevents the same observation from being used both to fit the nuisance functions and to evaluate its contribution to the estimator. Under suitable regularity conditions, including positivity, consistency, and a product-rate condition, the leading effect of estimating the two nuisance functions is governed by the product of their estimation errors. Therefore, neither nuisance function needs to be estimated at the parametric root-n rate. The resulting cross-fitted AIPW estimator is root-n consistent and asymptotically normal, and its variance can be estimated consistently. These properties are presented in detail in Dagdoug and Haziza (2026a).

Although the cross-fitted AIPW estimator provides a basis for valid inference, it does not directly produce a completed data file having the usual form required by data users. In addition, for a binary or categorical variable, the completed file should contain values that belong to the original support of the variable. For example, the imputed values for a binary variable should belong to $\{0, 1\}$. More generally, if a categorical variable takes values in $\{1, \ldots, L\}$, the imputed values should also belong to these L categories. Fractional machine learning predictions cannot be released as if they were observed categorical values.

The main contribution of this thesis was therefore to use rejective stochastic imputation within the double/debiased machine learning framework to construct random categorical imputations while keeping the estimator computed from the completed file close to the cross-fitted AIPW estimator. For a binary variable, the completed value can be written as

$$\widetilde{y}_k = r_k y_k + (1 - r_k) y_k^*, \qquad y_k^* \in \{0, 1\}.$$

The population proportion can then be estimated by applying the usual survey weights directly to the completed variable:

$$\widehat{\mu}_I = \frac{1}{N} \sum_{k \in S} w_k \widetilde{y}_k.$$

For a variable with several categories, the same completed file can be used to calculate the usual weighted count or proportion for each category.

The cross-fitted AIPW estimator determines the weighted total that should be assigned to the nonrespondents. This total becomes the target of the imputation procedure. In the binary case, the objective is then to generate zeros and ones whose weighted total is equal, or sufficiently close, to this target. This formulation separates the inference problem from the imputation problem. The cross-fitted AIPW estimator provides the statistical benchmark, while the imputation procedure transfers this benchmark to a completed categorical data file.

The main procedure studied in this thesis was rejective stochastic imputation. Candidate imputed values are first generated from Bernoulli or multinomial distributions based on cross-fitted outcome probabilities. The entire set of imputed values is accepted only when the estimator computed from the completed file is sufficiently close to the cross-fitted AIPW target. Otherwise, a new set of imputations is generated. The proposal probabilities could also be adjusted, when feasible, so that the expected weighted totals of the proposed imputations agree with the targets. This adjustment would not change the acceptance rule and could reduce the number of rejected proposals.

An important advantage of the rejective procedure is that the additional imputation variability can be controlled directly. If exact matching is possible and imposed, all accepted completed files produce exactly the same point estimate as the AIPW target. In that case, the imputation variance of the point estimator is equal to zero, even though different accepted files may assign the value one to different nonrespondents. If a small tolerance e_n is used instead, the conditional imputation variance of each estimated proportion is bounded by e_n^2. The tolerance must be large enough for at least one possible set of categorical imputations to satisfy the acceptance rule. When $e_n = o(n^{-1/2})$, this additional variance is $o(n^{-1})$, and is therefore negligible relative to the usual first-order sampling and nonresponse variance. Thus, the procedure preserves the binary or categorical support without introducing a first-order variance penalty.

The simulation results support these theoretical arguments. In both the binary and four-category simulations, the rejective imputed estimators had small Monte Carlo biases across the different combinations of machine learning methods. The variance estimators were reasonably accurate, and the coverage rates of the confidence intervals remained close to the nominal 95% level. These results indicate that, in the settings considered, the rejective procedure can produce a usable completed categorical data file while providing reliable estimates and confidence intervals.

The rejective method is simple to understand and implement, but it may become computationally demanding when the tolerance is very small or when several balancing constraints must be satisfied simultaneously. For example, if proportions must be preserved for several domains or several categories, the probability that independent proposals satisfy all the constraints may become small. It is therefore useful to consider alternative procedures that produce dependent imputed values and impose the balancing conditions more directly.

One important possibility is the cube method proposed by Deville and Tillé (2004). The cube method was originally developed to select a sample with prescribed first-order inclusion probabilities while balancing the Horvitz–Thompson estimators of auxiliary totals. In the binary imputation setting, the nonrespondents can be viewed as an artificial finite population, and the binary imputed value can be viewed as a sample-selection indicator. A value equal to one means that one is imputed, while a value equal to zero means that zero is imputed. To target the cross-fitted AIPW estimate, the imputation probabilities would first need to be chosen so that the expected weighted total for the nonrespondents agrees with the AIPW target, when this is feasible. Suitable balancing variables could then be used to keep the realized weighted total close to this target.

The cube method generates the binary values jointly rather than through repeated independent proposals. Its flight phase moves the vector of probabilities toward a vector of zeros and ones while preserving the balancing equations. The random updates preserve the original probabilities in expectation. If some components remain fractional, a landing phase is used to complete the binary vector, possibly with a remaining balancing discrepancy. This approach is closely related to the literature on balanced random imputation developed by Chauvet, Deville, and Haziza (2011) and Chauvet and Do Paco (2018). These papers show how balancing can substantially reduce, and in some settings eliminate, the additional variance caused by random imputation.

Developing a complete cube-based version of the procedure studied in this thesis would be an interesting direction for future work. Such a study should examine how to choose the imputation probabilities, how to deal with the landing phase, and how closely the resulting completed-data estimator matches the cross-fitted AIPW target under unequal survey weights and several simultaneous constraints. Any remaining balancing error would need to be small enough to preserve the first-order properties of the AIPW estimator. For variables with more than two categories, the procedure would also need to ensure that exactly one category is assigned to each nonrespondent. It would also be useful to compare the computational cost of the cube method with that of rejective imputation.

Another promising direction is the recent genetic algorithm proposed by Hasler, Eustache,

and Tillé (2026). This algorithm searches for a probability distribution over binary samples that preserves prescribed marginal probabilities while having a small expected balancing cost. In the imputation setting, the balancing cost can be defined as the squared difference between the weighted total of the imputed values and the AIPW target. When the imputation probabilities are chosen to match the target in expectation, this expected cost equals the conditional variance of the imputed total. The algorithm does not guarantee that the globally optimal design is found, and its computational requirements may be important. Nevertheless, it may produce imputation designs that are more closely balanced than those obtained with the traditional cube method. Its adaptation to item nonresponse therefore deserves further theoretical and numerical study.

Several other extensions could also be considered. Future work could study variables with many categories, rare categories, domain means, and situations with very low response probabilities.

References

Abadie, A. and G. W. Imbens (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. In: *Econometrica* 74.1, pp. 235–267.

Beaumont, J.-F. and C. Bocci (2009). Variance Estimation When Donor Imputation Is Used to Fill in Missing Values. In: *Canadian Journal of Statistics* 37.3, pp. 400–416.

Beaumont, J.-F. and D. Haziza (2016). A Note on the Concept of Invariance in Two-Phase Sampling Designs. In: *Survey Methodology* 42.2, pp. 319–323.

Biau, G. (2012). Analysis of a Random Forests Model. In: *Journal of Machine Learning Research* 13.38, pp. 1063–1095.

Breiman, L. (2001). Random Forests. In: *Machine Learning* 45.1, pp. 5–32.

Bühlmann, P. and B. Yu (2003). Boosting with the L_2 Loss: Regression and Classification. In: *Journal of the American Statistical Association* 98.462, pp. 324–339.

Cao, W., A. A. Tsiatis, and M. Davidian (2009). Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. In: *Biometrika* 96.3, pp. 723–734.

Chauvet, G., J.-C. Deville, and D. Haziza (2011). On Balanced Random Imputation in Surveys. In: *Biometrika* 98.2, pp. 459–471.

Chauvet, G. and W. Do Paco (2018). Exact Balanced Random Imputation for Sample Survey Data. In: *Computational Statistics & Data Analysis* 128, pp. 1–16.

Chen, J. and J. Shao (2000). Nearest Neighbor Imputation for Survey Data. In: *Journal of Official Statistics* 16.2, pp. 113–131.

Chernozhukov, V. et al. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. In: *The Econometrics Journal* 21.1, pp. C1–C68.

Dagdoug, M. and D. Haziza (2026a). A Debiased Machine Learning Single-Imputation Framework for Item Nonresponse in Surveys. arXiv:2611.01485 [stat.ME], version 1 (1 November 2026).

— (2026b). Machine Learning Methods for Finite Population Parameter Estimation in Survey Sampling. arXiv:2604.01160 [stat.ME], version 3 (21 July 2026).

Deville, J.-C. and Y. Tillé (2004). Efficient Balanced Sampling: The Cube Method. In: *Biometrika* 91.4, pp. 893–912.

Fay, R. E. (1991). A Design-Based Perspective on Missing Data Variance. In: *Proceedings of the 1991 Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, pp. 429–440.

Hasler, C., E. Eustache, and Y. Tillé (2026). In Search of the Most Balanced Sampling Design. arXiv:2607.26544 [stat.ME].

Horvitz, D. G. and D. J. Thompson (1952). A Generalization of Sampling Without Replacement from a Finite Universe. In: *Journal of the American Statistical Association* 47.260, pp. 663–685.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. In: *Statistical Science* 22.4, pp. 523–539.

Kim, J. K. and D. Haziza (2014). Doubly Robust Inference with Missing Data in Survey Sampling. In: *Statistica Sinica* 24.1, pp. 375–394.

Narain, R. (1951). On Sampling Without Replacement with Varying Probabilities. In: *Journal of the Indian Society of Agricultural Statistics* 3, pp. 169–175.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. In: *Journal of the American Statistical Association* 89.427, pp. 846–866.

Särndal, C.-E. (1992). Methods for Estimating the Precision of Survey Estimates when Imputation has been Used. In: *Survey Methodology* 18.2, pp. 241–252.

Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of Random Forests. In: *The Annals of Statistics* 43.4, pp. 1716–1741.

Shao, J. and P. Steel (1999). Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions. In: *Journal of the American Statistical Association* 94.445, pp. 254–265.

Stone, C. J. (1982). Optimal Global Rates of Convergence for Nonparametric Regression. In: *The Annals of Statistics* 10.4, pp. 1040–1053.

Yang, S. and J. K. Kim (2019). Nearest Neighbor Imputation for General Parameter Estimation in Survey Sampling. In: *The Econometrics of Complex Survey Data: Theory and Applications*. Ed. by K. P. Huynh, D. T. Jacho-Chávez, and G. Tripathi. Vol. 39. Advances in Econometrics. Emerald Publishing Limited, pp. 209–234.